CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS
ULBRA
Recredenciado pela Portaria Ministerial nº 3.607, de 17/10/05, D.O.U. nº 202, de 20/10/2005
ASSOCIAÇÃO EDUCACIONAL LUTERANA DO BRASIL

William Christhie Caproni de Oliveira

SENTIMENTALL: FERRAMENTA PARA ANÁLISE DE SENTIMENTOS EM PORTUGUÊS

Palmas - TO

2015

William Christhie Caproni de Oliveira

SENTIMENTALL: FERRAMENTA PARA ANÁLISE DE SENTIMENTOS EM PORTUGUÊS

Trabalho de Conclusão de Curso (TCC) elaborado e apresentado como requisito parcial para obtenção do título de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientadora: Prof. M.Sc. Parcilene Fernandes de Brito.

Palmas - TO

2015

William Christhie Caproni de Oliveira

SENTIMENTALL: FERRAMENTA PARA ANÁLISE DE SENTIMENTOS EM PORTUGUÊS

Trabalho de Conclusão de Curso (TCC) elaborado e apresentado como requisito parcial para obtenção do título de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientadora: Prof. M.Sc. Parcilene Fernandes de Brito.

Aprovada em:

BANCA EXAMINADORA

_____

Prof. M.Sc. Parcilene Fernandes de Brito.
Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.Sc. Jackson Gomes de Souza
Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M. Sc. Fabiano Fagundes
Centro Universitário Luterano de Palmas – CEULP

Palmas - TO

2015

Dedico este trabalho a todas as pessoas que, devido aos caminhos tortuosos e às dificuldades próprias da vida de cada um, não tiveram a oportunidade de concluir uma graduação. Dedico também aos meus amados filhos Isaac e Pedro, na esperança que meu esforço e dedicação para a realização deste trabalho os inspire a lançar voos mais altos em busca de conhecimento.

AGRADECIMENTOS

Agradeço primeiramente à Deus, pois tudo o que tenho e sou é por meio de Sua graça em minha vida. Em Colossenses 3.17 o apóstolo Paulo bem nos instrui “Tudo o que fizerem, seja em palavra seja em ação, façam-no em nome do Senhor Jesus, dando por meio dele graças a Deus Pai”. Deste modo, a Ele seja rendida toda honra e toda a glória agora e para sempre. Em verdade, tantos foram os momentos difíceis que enfrentei nestes longos anos de estudo que, se não fosse Sua mão amiga a me conduzir, e a Sua voz suave a me motivar, certamente não estaria aqui hoje celebrando esta vitória.

Agradeço também aos meus pais, aos meus avós e a todos meus familiares, pois de uma forma ou de outra, foram também fundamentais para que eu pudesse estar onde estou hoje. Em especial agradeço à minha mãe, pelas dificuldades financeiras que enfrentou durante minha infância, e aos meus avós maternos Geraldino e Maria de Lourdes, por todo o suporte dado durante o período em que morei com eles, para que eu concluísse o ensino médio. Agradeço também à minha madrinha Maria Aparecida Ribeiro de Godoy ✞ (dindinha) e à minha avó paterna Ernesta ✞, pois embora já não estejam mais em nosso meio, nunca deixaram de estar presentes em minhas lembranças com suas palavras de motivação e seus sábios conselhos.

Não poderia deixar de agradecer de maneira especial à minha amiga Thatiane de Oliveira Rosa, pois sempre acreditou em meu potencial, e mesmo no momento em que passei pelas piores dificuldades, esteve presente a meu lado, me motivando a voltar e concluir a graduação. Também neste sentido destaco o apoio e motivação dado pela Cristina Filipakis no momento que soube que eu estava pensando em voltar. Deixar para depois? Jamais!

Lembro saudoso e agradeço aos meus amigos já falecidos Carlos Eduardo Vilas Boas (Du) ✞, que me incentivou a entrar na universidade, e a meu amigo de estudos Gleisson Martins (Robinho Montanha) ✞. Que Deus os tenha em Seus braços. Porém, os amigos são tantos que daria um livro somente com as histórias e agradecimentos. Letícia, Ana Paula, Emiliano, Elizabeth, Cristiane, Douglas, Denise, Ranyelson, Sérgio, Andrew e tantos outros, não pensem que foram esquecidos,

sintam-se todos cobertos por minha gratidão. Embora este espaço seja pequeno para citá-los, seu lugar é cativo em meu coração.

Igualmente agradeço à minha orientadora Parcilene, pois não foi simplesmente uma orientadora. De fato, a despeito de ter aceitado me orientar neste trabalho mesmo sem ter horários disponíveis, não foram poucas as madrugadas, feriados e finais de semana em que esteve sempre pronta a ajudar. Orientadores não fazem isso, mas sim amigos! Também pude contar com a prontidão de todos os professores em momentos de dúvidas e tribulações, e por isto serei sempre grato.

O apoio oferecido pelo Tribunal de Justiça, através da concessão de horário especial, foi de fundamental importância. Sem ele certamente este trabalho não poderia atingir o resultado obtido dentro do prazo estipulado. Neste sentido gostaria de agradecer especialmente a meus colegas de trabalho, que se desdobraram para cobrir minha ausência durante o período de estudos. Esta vitória também é de vocês!

Gostaria de registrar aqui meu agradecimento aos pesquisadores William Daniel Colen de Moura Silva (USP), Paula Carvalho e equipe (REACTION), Sandra Aluísio e equipe (USP) e Miriam Lúcia Domingues (ICS – UFPA). Sem a pronta cooperação e cessão de suas pesquisas e materiais não seria possível o desenvolvimento deste trabalho. Não são poucos os desafios em se desenvolver pesquisa científica, principalmente no Brasil, e é graças a dedicação e doação de pesquisadores como estes que ela continua a evoluir.

Enfim, agradeço a todos que direta ou indiretamente contribuíram para o desenvolvimento desta minha caminhada na graduação, que iniciou-se há muito tempo atrás (2005), mas que após muitas e inacreditáveis reviravoltas da vida, encontrou um desfecho feliz. Que esta seja somente a primeira de muitas aventuras que ainda hão de vir... Muito Obrigado!

RESUMO

CHRISTHIE, William. **SENTIMENTALL: FERRAMENTA PARA ANÁLISE DE SENTIMENTOS EM PORTUGUÊS**. 2015. 139 f. Trabalho de Conclusão de Curso (Graduação) – Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas, Palmas/TO, 2015.

Este trabalho objetiva o desenvolvimento de um protótipo de ferramenta capaz de realizar a Análise de Sentimentos em textos escritos em Português, no nível de aspectos, utilizando a abordagem léxica e o Processamento de Linguagem Natural. São apresentados os conceitos da Análise de Sentimentos nos níveis de documento, sentença e aspecto. As etapas para a análise no nível de aspectos utilizando *lexicons* de sentimentos são descritas, bem como os conceitos envolvendo o Processamento de Linguagem Natural e suas etapas. Também são apresentadas as técnicas de *Web Crawling* e *Web Scraping*, utilizadas para a extração de conteúdos da internet. Neste sentido, para a aplicação das técnicas e conceitos apresentados, um protótipo de ferramenta foi desenvolvido utilizando a linguagem de programação Java, a API do corretor ortográfico “CogrOO”, e os *lexicons* de sentimentos “SentiLex” e “LIWC”. No que tange ao teste da ferramenta, foi utilizado o contexto do turismo, especificamente os comentários sobre atrações, hotéis e restaurantes de destinos turísticos brasileiros. Um corpus de quase um milhão e meio de comentários, escritos em português, foi extraído do site especializado em turismo “Trip Advisor”. Para a extração foi utilizada a ferramenta Import.IO, para transformação dos dados a ferramenta “Kettle – Spoon”, e para armazenamento o SGBD “SQL Server”. Com a utilização da ferramenta desenvolvida, foram efetuados sobre este corpus: o processo de normalização, o Processamento de Linguagem Natural, e a Análise de Sentimentos no nível de aspectos. Deste modo, um conjunto de informações resultantes de cada etapa da análise são apresentados e discutidos.

PALAVRAS-CHAVE: Análise de Sentimentos, Processamento de Linguagem Natural, Turismo.

ABSTRACT

This work presents the development of a prototype tool that performs Sentiment Analysis in texts written in Portuguese, at the aspect level, using the lexical approach and Natural Language Processing. The concepts of Sentiment Analysis are presented at the document, sentence and aspect levels. The stages for analyzing the aspect level using sentiment lexicons are described, as well as the concepts involving the Natural Language Processing and its stages. The techniques of Web Crawling and Web Scraping, used for the internet content extraction, are also introduced. In this sense, for the application of techniques and concepts presented, a prototype tool was developed using the Java programming language, the API of the spellchecker “CoGrOO”, and the sentiment lexicons in Portuguese "SentiLex" and "LIWC". Regarding the test of the tool, the context of national tourism was used, specifically the reviews on attractions, hotels and restaurants of Brazilian touristic destinations. A corpus of nearly a million and a half of comments, written in Portuguese, was extracted from the site specialized in tourism "Trip Advisor". The Import.IO tool was used for extraction, the "Kettle - Spoon" tool for data transformation, and the DBMS "SQL Server" for storage. With the developed tool, this corpus underwent: the Normalization Process, the Natural Language Processing, and the Sentiment Analysis at the Aspect Level. Thus, a set of resulting information from each step of the analysis are presented and discussed.

Keywords: Sentiment Analysis, Natural Language Processing, Tourism.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ANEXOS

LISTA DE APÊNDICES

LISTA DE ABREVIATURAS

ADVP - Adverbial Phrase
ANP - Adjective Noun Pairs
API - Application Programming Interface
EOS - End-of-speech
ETL - Extract, Transform and Load
HMM - Hidden Markov Models
IDE - Integrated Development Environment
ISK - Institute of Linguistics and Communication Studies
MWE – Multi Word Expression
NGD - Normalized Google Distance
NLP - Expectation-Maximization
NLP - Natural Language Processing
NP - Noun Phrases
NWD - Normalized Web Distance
PIB - Produto Interno Bruto
PoS - Part-of-speech
PP - Prepositional Phrases
SDU - University of Southern Denmark
SGBD - Sistema de Gerenciamento de Banco de Dados
VISL - Visual Interactive Syntax Learning
VP - Verb Phrases

SUMÁRIO

1 INTRODUÇÃO

O constante crescimento da internet, aliado ao aumento de funcionalidades com interatividade social vêm provocando um crescimento exponencial no volume de dados disponível nas redes sociais, nos blogs e nos sites especializados de modo geral. Concomitantemente, a alta competitividade e dinamismo do mundo atual exigem constante atualização em relação às informações de mercado. Assim, são necessários sistemas computacionais que sejam capazes de processar tal volume de informações em tempo hábil, e de modo a facilitar a inferência de novas informações e conhecimentos. A extração automática de informações oriundas de fontes não estruturadas tem proporcionado novas oportunidades de negócio e de análises.

Neste sentido, com o intuito de automatizar o processo de extração, classificação e sumarização de sentenças que contenham opiniões, surgiram os sistemas de Análise de Sentimentos, também conhecidos como Mineração de Opiniões. “A análise de sentimentos essencialmente tenta inferir o sentimento das pessoas baseando-se em suas expressões de linguagem” (LIU, 2010, p. 6, tradução nossa).

As opiniões, de modo geral, refletem o sentimento do usuário a respeito de um produto ou serviço, podendo ser classificadas em positivo ou negativo. A contabilização dos sentimentos de um grande volume de opiniões, ou seja, a sua sumarização, pode capturar a essência dos sentimentos coletivos, e assim viabilizar ações de correção e melhoria nos produtos e serviços.

A análise das opiniões pode ser efetuada no nível de documento, de sentenças, ou de aspectos, e pode utilizar abordagens de aprendizado de máquinas, léxicas, estatísticas e semânticas.

Uma importante ferramenta na abordagem léxica são os *lexicons*, ou dicionários que mapeiam léxicos em suas polaridades. “O léxicon de sentimentos é o recurso mais crucial para a maioria dos algoritmos de análise de sentimentos” (FELDMAN, 2013, p. 86, tradução nossa).

Para possibilitar que os textos escritos em linguagem natural sejam analisados, são utilizadas as técnicas de Processamento de Linguagem Natural ou em inglês, *Natural Language Processing* (NLP). O NLP “se ocupa do estudo da linguagem voltado para a construção de ferramentas computacionais específicas relacionadas a tarefas básicas direcionadas ao processamento da informação baseada na linguagem humana” (DOMINGUES, 2011, p. 27).

Alguns trabalhos buscam estratégias para realizar a análise de sentimentos valendo-se de técnicas de tradução de máquina e assim poder aproveitar da ampla gama de recursos linguísticos na língua inglesa. Em geral, nestes algoritmos traduz-se o texto para o inglês e então se efetua a análise de sentimentos. Como ponto positivo esta técnica tem todo o arcabouço tecnológico existente para a língua inglesa, tais como grande gama de dicionários léxicos de sentimentos e algoritmos eficientes de NLP. Entretanto, como ponto negativo estão os problemas no processo de tradução automática, que não consegue uma boa precisão, assim como não consegue capturar todas as particularidades das línguas originárias dos textos. Neste sentido não foi utilizada a estratégia de tradução de máquina neste trabalho, mas sim ferramentas desenvolvidas especificamente para o Português. Uma boa descrição desta abordagem e de trabalhos relacionados pode ser encontrada em Liu (2012, p. 41).

O presente trabalho é parte integrante de um projeto de pesquisa multidisciplinar desenvolvido em parceria entre os grupos de pesquisa Engenharia Inteligente de Dados, do CEULP, e Economia Comportamental e Análise do Comportamento do Consumidor, da PUC-GO. O projeto objetiva a utilização de técnicas computacionais de Análise de Sentimentos e de técnicas psicológicas de Análise Comportamental aplicadas no estudo do contexto do Turismo Nacional. Um dos ambientes estudados neste projeto é o site especializado em turismo “Trip Advisor”[1], que alega ter a maior comunidade de viajantes do mundo. Em 2013 existiam mais de 170 milhões de avaliações, dentre as quais, dois milhões e oitocentas mil avaliações somente sobre os 1479 destinos turísticos brasileiros cadastrados. São mais de 280 milhões de usuários registrando suas opiniões sobre pontos turísticos de 45 países em todo o mundo.

[1] http://tripadvisor.com.br

Neste cenário, o presente trabalho busca responder ao seguinte problema de pesquisa: Como implementar a técnica de análise de sentimentos no contexto das avaliações públicas de sites sobre destinos turísticos brasileiros? A hipótese aqui abordada é de que é possível realizar a análise de sentimentos ao nível de aspecto, nos comentários públicos escritos em língua portuguesa, utilizando-se os recursos linguísticos e computacionais existentes.

Portanto, o objetivo geral deste trabalho é utilizar técnicas computacionais para criar um protótipo que efetue a análise de sentimentos nas avaliações públicas de destinos turísticos brasileiros em língua portuguesa apresentados em web sites. Neste sentido objetiva-se especificamente:

- Extrair avaliações públicas sobre pontos turísticos brasileiros, escritas em língua portuguesa e disponíveis em web sites;
- Realizar o Processamento de Linguagem Natural nas avaliações extraídas;
- Efetuar a análise de sentimentos nas avaliações processadas.

Este trabalho está estruturado da seguinte forma. O capítulo 2 apresenta no referencial teórico um estudo sobre os assuntos abordados no trabalho: Análise de Sentimentos (2.1), Processamento de Linguagem Natural (2.2), *Web Crawling* e *Web Scraping* (2.3), e por fim o processo de avaliação (2.4). No capítulo 3 é apresentada a metodologia e os materiais utilizados no desenvolvimento do trabalho. O capítulo 4 apresenta os resultados obtidos, alguns dados são apresentados e discutidos. Por fim as considerações finais são apresentadas no capítulo 5, e as referências bibliográficas utilizadas, no capitulo 6.

2 REFERENCIAL TEÓRICO

A Análise de Sentimentos, ou Mineração de Opiniões, é uma área de pesquisa que objetiva tratar computacionalmente as opiniões expressas em textos escritos em linguagem natural. O processo envolve a utilização de teorias e técnicas computacionais capazes de lidar com estas informações, expressas de modo não estruturado, dentre as quais processamento de linguagem natural, aprendizado de máquina e outras.

Este capítulo provê inicialmente uma visão geral dos conceitos e técnicas utilizadas na análise de sentimentos, alguns dos quais serão utilizados neste trabalho. Segue-se na seção 2.2 com a apresentação dos conceitos de Processamento de Linguagem Natural. Explicam-se os conceitos envolvidos na extração de conteúdos disponíveis em páginas da internet na seção 2.3 e, por fim, apresenta-se a técnica utilizada para avaliação do processamento realizado na seção 2.4.

2.1. Análise de Sentimentos ou Mineração de Opiniões

Como saber a opinião pública de um modo geral sobre determinado serviço ou produto? Considerando-se a grande quantidade de informação que está disponível na internet, pode-se questionar se elas não poderiam ser utilizadas para responder a esta pergunta. De fato, as técnicas da área de análise de sentimentos objetivam utilizar a sumarização desta ampla gama de informações, de modo a inferir estas opiniões.

Conforme Liu (2010), a informação textual amplamente disponível na internet em linguagem natural pode ser categorizada em dois tipos: opiniões e fatos. Ainda segundo o autor, fatos relatam acontecimentos passados ou verdades, enquanto que as opiniões referem-se a pontos de vista subjetivos, expressos sobre as mais diversas coisas, fatos e pessoas. Isto pode ser observado no comentário a seguir, cujos trechos foram numerados para facilitar a explicação:

Exemplo 1: Juliana, em 03/08/2014: “(1) Fui à Baía do Sancho este final de semana. (2) Realmente o visual é de perder o fôlego. (3) O acesso é um pouco complicado, (4) mas melhorou bastante. (5) Leve água e proteção para o sol.”.

Neste comentário existem fatos, como (1) e (5) e opiniões como (2), (3) e (4). As sentenças opinativas podem ser computacionalmente coletadas e sumarizadas, de modo a demonstrar ou verificar a opinião das pessoas sobre uma determinada entidade.

Liu (2010, p. 3) formalizou um conjunto de conceitos envolvidos no problema da análise de sentimentos. Tais conceitos são descritos a seguir:

- **Objeto**: A entidade sobre a qual um determinado comentário opinativo está sendo efetuado. É expresso por um conjunto finito de características ou aspectos do objeto $F = \{f_1, f_2, f_3, \dots, f_n\}$, o qual contém o próprio objeto como um aspecto especial. Neste sentido, considerando o Exemplo 1, tem-se um comentário expresso sobre o Objeto: “Baía do Sancho”, o qual pode ser descrito pelo conjunto de seus aspectos F= {Visual, Acesso, Baía do Sancho}. O último atributo deve-se ao fato de algumas opiniões não serem expressas sobre uma característica específica, mas sobre o objeto como um todo;

- **Aspectos implícitos e explícitos**: Dada uma sentença s, diz-se que o aspecto $f_i \in F$ de um objeto o é explícito caso ele ou um de seus sinônimos apareçam em s. Caso nem f_i nem algum de seus sinônimos apareçam, mas f_i possa ser inferido, ele é chamado de implícito. Assim, possui um conjunto finito de sinônimos $W_i = \{w_1, w_2, w_3, \dots w_m\}$ e um conjunto finito de indicadores $I_i = \{i_1, i_2, i_3, \dots i_q\}$. Deste modo, por exemplo, pode-se dizer que a característica “Valor” possui um conjunto de sinônimos W= {custo, preço,...} e um conjunto de indicadores I= {caro, barato,...}. Com isso é possível mapear ocorrências de sinônimos e aspectos implícitos nas frases opinativas;

- **Titular da opinião**: É o autor da opinião propriamente dita;

- **Orientação da opinião**: É a polaridade da opinião, geralmente expressa em termos de positiva, negativa, ou neutra;

- **Documento opinativo**: É qualquer documento que contenha opiniões de um conjunto de titulares, sobre um conjunto de objetos, expressas sobre um subconjunto de aspectos destes objetos;

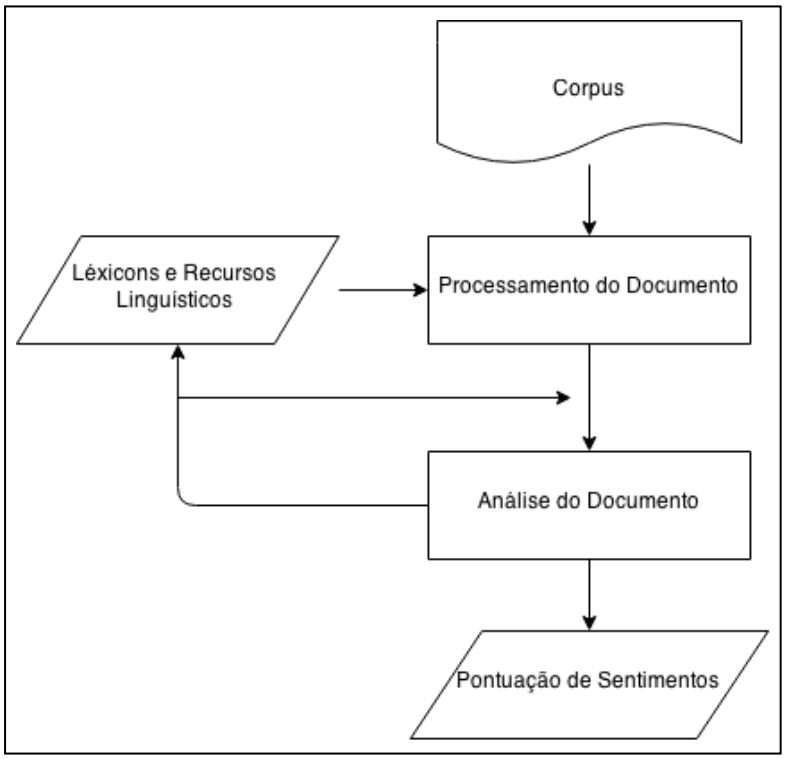
- **Opinião direta**: É expressa na forma de uma quíntupla $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, ou seja, é expressa em um aspecto f_{jk} de um objeto o_j, apresentando uma polaridade h_i, expressa por um titular oo_{ijkl} em um momento específico no tempo t_l. Deste modo, o Exemplo 1 poderia ser representado pelo conjunto de registros a

seguir, contendo os cinco atributos supracitados para cada opinião expressa pelo autor: {(Baía do Sancho, visual, Juliana, Positivo, 03/08/2014), (Baía do Sancho, acesso, Juliana, Negativo, 03/08/2014)};

- **Opinião comparativa**: Expressa uma relação de igualdades ou de diferenças entre um ou mais objetos distintos. Geralmente são expressas usando-se frases superlativas ou comparativas. Um comentário deste tipo poderia ser, por exemplo, “A Baía do Sancho é a melhor!” ou então, “Prefiro Baía do Sancho a Praia do Forte...”.

Definida a terminologia, será explicado o processo de análise de sentimentos, demonstrado na figura a seguir.

Figura 1 – Processo de Análise de Sentimentos.



Fonte: Feldman (2013).

Na Figura 1, Feldman (2013) apresenta uma visão geral da arquitetura de um sistema genérico de AS. A entrada do sistema é um corpus, ou conjunto de documentos a serem analisados. Sobre estes documentos, o módulo de processamento do documento utiliza um conjunto de recursos linguísticos para efetuar tarefas tais como tokenização, lematização, ou marcações PoS. Estas etapas serão mais bem descritas na seção 2.2, a qual descreverá o Processamento de Linguagem Natural. Após o processamento do documento, o módulo “Análise do

Documento” utiliza os dados processados, juntamente com *lexicons* de sentimentos, para etiquetar os documentos com as polaridades das opiniões detectadas. Esta análise pode ser efetuada no documento como um todo, para cada sentença ou para cada aspecto observado. Os níveis de análise serão abordados na seção a seguir. Para apresentar os resultados globais da análise ao usuário, o sistema possui um módulo de pontuação de sentimentos o qual é responsável por contabilizar as polaridades anotadas na etapa anterior. As etapas do processo de Análise de Sentimentos estarão mais bem descritas na seção 2.1.2. Como pode ser observado na Figura 1, o módulo de Análise do Documento pode fornecer insumos para geração de *lexicons*. Embora não seja uma etapa da AS propriamente dita, na seção 2.1.3 serão explicados os conceitos envolvidos na geração de dicionários léxicos de sentimentos, uma importante ferramenta auxiliar do processo de análise de sentimentos.

2.1.1. Nível de Análise

Analisar opiniões sobre determinado objeto requer a análise do sentimento expresso nos documentos. Entretanto, “um documento com opinião positiva sobre um determinado objeto não significa que o autor tem opinião positiva sobre todos os aspectos ou características do objeto” (LIU, 2010, p. 16, tradução nossa). O tipo de análise no nível de documento é conhecido por “*document-level sentiment classification*” (LIU, 2010, p. 2). Sendo assim, um exemplo de análise no nível de documento poderia considerar o seguinte comentário:

Exemplo 2: “(1) Conheci hoje a praia do forte. (2) O visual era simplesmente maravilhoso. (3) Água transparente, linda! (4) Mas não gostei da temperatura da água, um gelo...”.

Como visto no exemplo, o documento possui várias opiniões, sendo algumas positivas e outras não. Uma análise de sentimentos no nível de documento verificaria qual sentimento é mais predominante no texto como um todo e o classificaria como tal. Assim, temos as opiniões 2 e 3 com sentimentos positivos e a opinião 4 como negativa, o que classificaria o documento como positivo.

Pode-se então, buscar um maior detalhamento no nível de análise, verificando se cada frase contida no documento possui ou não opiniões. Este processo também é chamado “*subjectivity classification*” (LIU, 2010, p. 2). Ainda segundo o autor, caso esta seja uma frase opinativa, deve ser classificada quanto a sua polaridade, tarefa conhecida como “*sentence-level sentiment classification*”.

Segundo o Exemplo 2, cada frase deve ser classificada quanto a sua polaridade. A análise excluiria a frase 1 ou a consideraria como neutra, por não se tratar de frase opinativa, e consideraria a polaridade de cada frase, sem sumarizar a polaridade do documento.

Muitas vezes, aplicações do mundo real precisam de análises ainda mais detalhadas, sendo que no nível dos aspectos expressos nas opiniões, a análise é chamada "*feature-based sentiment analysis*" (LIU, 2010, p. 2). Segundo o autor, para que sejam possíveis melhorias nos produtos e serviços, faz-se necessário observar o que os consumidores estão gostando ou não, e isto só é possível através deste tipo de análise. Ainda segundo o Exemplo 2, uma análise no nível de aspecto poderia ser:

Tabela 1 – Exemplo de análise no nível de aspecto

Trecho	Objeto	Aspecto	Polaridade
2	Praia do Forte	Visual	positivo
3	Praia do Forte	Água	positivo
3	Praia do Forte	Água	positivo
4	Praia do Forte	Água	negativo

Fonte: Próprio autor

A Tabela 1 apresenta um exemplo de análise de sentimentos efetuada no nível de aspecto para o Exemplo 2. Observa-se a repetição do aspecto "Água", que ocorre devido ao trecho 3 apresentar duas opiniões positivas (transparente e linda) e ao trecho 4 apresentar uma opinião negativa (temperatura). A seguir serão apresentadas as etapas para que a análise no nível de aspecto possa ser realizada.

2.1.2. Etapas de análise

Para que seja possível a análise no nível dos aspectos dos objetos será necessária a identificação de, pelo menos, o objeto, o aspecto e a polaridade. Becker e Tumitan (2013) definem três etapas para a mineração de opiniões, a saber: Identificação, Classificação de Polaridade e Sumarização.

Na etapa de Identificação, as frases opinativas devem ser selecionadas para extração. Basicamente a tarefa consiste em determinar se uma frase é subjetiva com opiniões, ou objetiva, o que pode ser definido como um problema de classificação de textos. A classificação automática de textos pode ser vista como o problema de determinar automaticamente a classe a qual os documentos textuais pertencem.

Wang e Liu (2011) utilizam técnicas de Aprendizado de Máquina para a classificação de frases opinativas. Como a informação disponível na internet é vasta, mas a disponibilidade de dados etiquetados é pequena, eles utilizam um método semissupervisionado para a tarefa. Através do método *Expectation-Maximization* (EM) com uma restrição de distribuição de classe, eles treinam um classificador *Naive Bayes* e assim efetuam a tarefa de classificação.

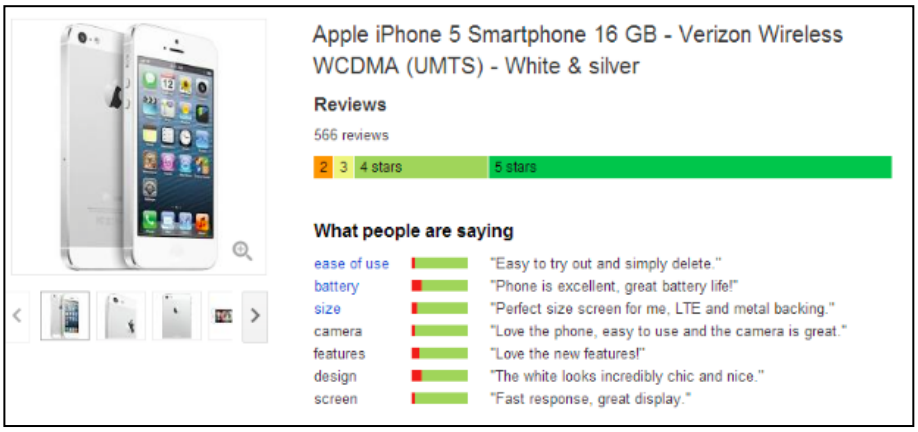
Como as opiniões são subjetivas, diferentes pontos de vista podem fornecer diferentes opiniões sobre o mesmo aspecto de um mesmo objeto. Neste sentido, a etapa de sumarização busca formas eficientes de se capturar a essência das opiniões.

Nesta etapa é efetuada a contabilização das opiniões em métricas qualitativas e quantitativas, de modo a representar o sentimento geral sobre um determinado objeto e seus aspectos. Liu (2012) explica que para identificar a opinião média ou prevalecente, a opinião expressa por um pequeno grupo de pessoas não é suficiente, sendo necessário analisar uma grande quantidade de opiniões. Segundo o autor, medidas quantitativas devem fornecer o percentual de pessoas que opinaram positivamente ou negativamente, sobre quais aspectos de quais objetos, além dos aspectos mais comentados positiva ou negativamente.

Para que estas medidas possam ser mais bem visualizadas, tendo em vista o grande volume de informações, devem ser utilizados gráficos para sumarizar a informação. Por exemplo, a informação pode ser apresentada inicialmente com poucos detalhes, em um alto grau de sumarização, e ir aumentando o nível de detalhes, conforme a necessidade do usuário do sistema que utilize a análise de sentimentos, até chegar ao nível dos comentários que geraram a análise.

Bons exemplos de utilização dos sumários de opiniões são apresentados em Becker e Tumitan (2013), apresentados a seguir.

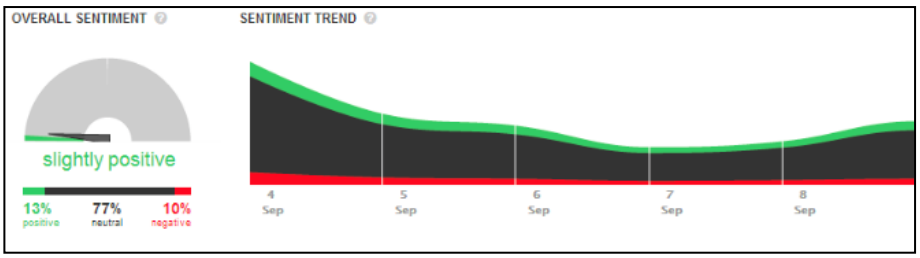
Figura 2 – Sumarização de AS em aspectos de produtos



Fonte: Google Shopping (2013 *apud* BECKER e TUMITAN, 2013).

A Figura 2 apresenta uma sumarização para a análise de sentimentos efetuada no sistema Google Shopping[2], o qual identifica e sumariza os sentimentos de consumidores em diversas lojas de comércio eletrônico. Os aspectos de cada produto são identificados e pontuados, além de ser efetuada uma pontuação geral baseada em classificação por estrelas.

Figura 3 – Sumarização de AS em redes sociais



Fonte: UberVU (2013 *apud* BECKER e TUMITAN, 2013).

Na Figura 3, os autores demonstram a sumarização efetuada pela empresa UberVU, que acompanha o sentimento e os comentários do público geral em redes sociais sobre marcas e empresas. O relatório apresenta a variação dos sentimentos em relação ao tempo e pode ser útil para analisar o impacto de eventos específicos, como lançamentos ou vendas.

[2] http://www.google.com/shopping

Figura 4 – Sumarização de AS em preços de ações

Fonte: Social Market Analytics (2013 *apud* BECKER e TUMITAN, 2013).

Na Figura 4 os autores apresentam a sumarização efetuada pela empresa *Social Market Analytics*, que efetua a análise de sentimentos nos comentários do Twitter sobre os preços de ações de empresas específicas. Ela criou suas próprias métricas para representar o sentimento dos compradores antes da abertura e no fechamento de cada pregão. Elas baseiam-se na série de tempo normalizada e ponderada ao longo de um período retrospectivo (S-Score) e na média suavizada desta (S-Mean).

Para que seja possível a sumarização das opiniões como demonstrado, é necessária uma etapa anterior que seja responsável por classificar cada opinião em relação a sua polaridade. Entretanto, classificar frases opinativas entre positivo ou negativo, mesmo entre humanos, nem sempre é objeto de consenso e, portanto, não é trivial. Isto poderia ser visualizado no Exemplo 1, modificando-se o trecho (5) para: "... (5) Leve água e proteção para o sol, (6) pois lá embaixo não tem nada.". Algumas pessoas poderiam classificar o trecho (6) como neutro por julgarem ser um fato, enquanto outras poderiam classificá-lo como negativo. As pesquisas atuais ignoram frases como estas ou as assumem com a polaridade neutra, devido ao fator subjetivo de polaridade não clara.

Para a classificação de polaridades podem ser utilizadas abordagens de quatro grupos distintos: léxicas; aprendizado de máquina; estatísticas; e semânticas, sendo que as técnicas destas diferentes abordagens podem ser combinadas para melhoria de resultados.

A abordagem léxica vale-se de dicionários (ou *léxicon*) de sentimentos, os quais possuem tuplas de léxicos e suas respectivas polaridades. Podem ser utilizados *lexicons* existentes de sentido geral, ou podem ser criados dicionários específicos para o domínio em questão. Assim, de acordo com Silva, Lima e Barros (2012), a classificação é feita com base nos adjetivos encontrados no texto, também

chamados de Palavras Opinativas, e na respectiva polaridade no léxicon utilizado. Feldman (2013), bem como Becker e Tumitan (2013) afirmam que se deve ainda observar a existência de palavras adversativas e/ou de negação, além de ironia e sarcasmo para a definição final da polaridade. Entretanto, segundo os autores, ironia e sarcasmo são extremamente complexos de serem detectados, além de serem mais comuns em alguns domínios tais como política e esporte, e por isso nem sempre são considerados. O aprofundamento sobre este assunto está fora do escopo deste trabalho, porém mais informações podem ser encontradas em Carvalho et al. (2009 *apud* BECKER e TUMITAN, 2013), Liu (2012) e Tsur et al. (2010 *apud* FELDMAN, 2013).

Apesar de alguns trabalhos considerarem somente as palavras opinativas, substantivos podem ser neutros e adjetivos podem ser dependentes do contexto no qual estão inseridos, conforme observado em Borth et al. (2013). Tendo isto em vista, torna-se muito difícil, se não impossível, manter dicionários de sentimentos universais (QIU et al., 2009).

Para melhor entendimento do processo, é apresentado o exemplo a seguir:

Exemplo 3: “(1) A água do chuveiro estava gelada! (2) Mas chegando no restaurante a pizza quentinha valeu o sacrifício... (3) Ela estava deliciosa!”.

Existem palavras opinativas que são fortemente polarizadas, como é o caso do adjetivo "deliciosa" em (3), o qual está qualificando o aspecto pizza do objeto restaurante. Mas existem palavras opinativas que são dependentes do contexto, como é o caso de gelada (1) e quentinha (2). Para este tipo de situações utilizam-se dicionários contendo entradas do tipo (água_de_chuveiro, fria, negativo) e (pizza, quente, positivo). Assim, ao efetuar uma busca pelo léxicon de sentimentos, podem-se mapear as palavras opinativas encontradas no texto em polaridades sensíveis ao contexto. Neste sentido, por exemplo, Silva construiu um léxicon de sentimentos específico para o domínio de comentários em inglês sobre celulares. Borth et al. (2013) confeccionaram uma ontologia de sentimentos visuais, composta por mais de três mil pares entre adjetivos e substantivos, para a detecção de sentimentos em conteúdos visuais como imagens e vídeos. Na seção a seguir serão apresentadas as técnicas envolvidas na geração de *lexicons,* elemento chave nas etapas identificação e classificação de polaridades ao utilizar-se a abordagem léxica.

2.1.3. Geração de lexicons

De acordo com Feldman (2013), a aquisição dos dicionários léxicos pode ocorrer por abordagem manual, por utilização de dicionários ou por utilização de corpus específicos do domínio. A abordagem manual é pouco utilizada devido à necessidade de grande esforço para a construção dos dicionários. Já na utilização de dicionários, um pequeno quantitativo de palavras iniciais alimenta o processo que procura por sinônimos e antônimos em algum dicionário on-line para aumentar este conjunto inicial (LIU, 2012, p. 91). Na abordagem baseada em corpus, utiliza-se um conjunto inicial de termos para a localização de outras palavras opinativas em um grande conjunto de documentos, valendo-se de técnicas estatísticas ou linguísticas.

Ding e Liu (2008 *apud* LIU, 2012, p. 96) demonstram que, em diferentes contextos, algumas palavras podem ter diferentes orientações, mesmo que no mesmo domínio. Eles propõem um modelo de pares (característica, palavra opinativa) para o dicionário. Igualmente Borth et al. (2013), que utilizam uma base léxica composta de pares "*Adjective Noun Pairs*" (ANP), ou pares entre substantivos e adjetivos. Com a utilização dos ANPs, obtêm-se dicionários com léxicos contextuais de sentimentos fortes, além de tornar tais conceitos mais detectáveis (BORTH et al., 2013).

No trabalho de Silva, Lima e Barros (2012), ao desenvolverem uma base léxica contendo pares (características, palavras opinativas) para tratar a semântica das palavras opinativas, os autores dividem o processo em duas etapas centrais. Na primeira são extraídos os pares de características e palavras opinativas em um corpus representativo não etiquetado. Na segunda etapa, é utilizada a medida *Normalized Google Distance* (NGD) para o cálculo da polaridade dos pares detectados na etapa anterior. Esta medida e as etapas para a criação dos léxicos de sentimentos serão detalhadas nas subseções a seguir.

Extração de aspectos e palavras opinativas

Utilizando algumas palavras opinativas iniciais, Qiu et al. (2009) demonstraram ser possível a geração de um dicionário de palavras opinativas específicas de domínio. Eles também demonstraram que estas palavras opinativas quase sempre estão associadas aos aspectos dos objetos. Em um processo iterativo, o algoritmo *Double Propagation,* criado por Qiu et al. (2009), descobre novos aspectos e palavras opinativas a cada ciclo, valendo-se das descobertas das iterações anteriores, até que não sejam encontrados novos aspectos ou novas

entradas para o dicionário. Ele está dividido em quatro tarefas, cada qual com um conjunto específico de regras para explorar os três tipos de relações existentes entre palavras opinativas e aspectos:

1. Extrair palavras opinativas utilizando outras palavras opinativas;
2. Extrair aspectos utilizando palavras opinativas;
3. Extrair palavras opinativas utilizando aspectos;
4. Extrair aspectos utilizando aspectos.

O algoritmo adota uma estratégia baseada em regras, construídas a partir dos três tipos de relações entre palavras opinativas e aspectos. Por exemplo, uma das regras apresentadas em Qiu et al. (2009) para extração de aspectos é selecionar substantivos que possuem palavra opinativa como adjunto modificador. Os autores apresentam um conjunto de oito regras, duas para cada tarefa.

Para identificar os aspectos implícitos, o processo deve mapear os indicadores de aspectos, dos quais, conforme Borth et al. (2013), os mais comuns são advérbios e adjetivos. Ainda segundo os autores, após o processo de formação dos pares, uma etapa extra de analise deve ser empregada para evitar pares que sejam iguais a entidades com sentido alterado, como por exemplo, “cachorro quente”.

Após a extração dos pares, estes devem ser passados para a próxima fase, ou seja, o cálculo da polaridade dos sentimentos relacionados, a qual será explicada a seguir.

Cálculo da polaridade

Para o cálculo da polaridade dos pares de aspectos e palavras opinativas, deve-se medir a distância de informação entre o conjunto de palavras formado por cada um dos pares e dois conjuntos de informações, um contendo palavras fortemente positivas e outro contendo palavras fortemente negativas. “Distância da informação mede a informação entre um par de objetos” (BENNETT et al., 1998 *apud* COHEN; VITÁNYI, 2013, p. 1, tradução nossa) ou entre pares de conjuntos finitos não vazios (LI et al., 2008 *apud* COHEN; VITÁNYI, 2013).

Para aferir a similaridade entre palavras e/ou frases e assim obter conhecimento sobre a similaridade de objetos, no trabalho de Cilibrasi e Vitányi (2007 *apud* VITÁNYI; CILIBRASI, 2010) criou-se a medida NGD. Posteriormente a chamaram de *Normalized Web Distance* (NWD), já que para o cálculo pode ser

utilizado qualquer buscador que: apresente o total de páginas retornadas nas consultas; e que possua uma grande quantidade de páginas indexadas.

A técnica vale-se da vastidão de informações disponíveis na internet, “efetivamente o maior banco de dados semântico eletrônico do mundo” (VITÁNYI; CILIBRASI, 2010, p. 294, tradução nossa), para calcular uma aproximação computacional normalizada da distância da informação entre dois objetos, e assim obter uma medida universal de similaridade. Segundo Cohen e Vitányi (2013, p.6, tradução nossa), “a contagem relativa de páginas retornadas pelo Google é tão grande e diversificada que se aproxima do uso real das palavras e frases”.

Considerando dois termos x e y, a similaridade entre eles varia de 0 (os termos são idênticos) até 1 (totalmente díspares), o que pode ser calculado pela fórmula:

$$NGD_{(x,y)} = \frac{max\{\log_2 f(x), \log_2 f(y)\} - \log_2 f(x,y)}{\log_2 N - min\{\log_2 f(x), \log_2 f(y)\}} \quad (1)$$

Onde:

$f(x)$ representa o número de páginas com ocorrências de x;

$f(x,y)$ representa o número de páginas com ocorrências de ambos x e y;

e N representa o número total de páginas indexadas pelo mecanismo de busca.

Sua utilização pode ser mais bem entendida no seguinte exemplo:

Exemplo 4: Conjunto de pares de aspectos e palavras opinativas P = {<visual, perder_o_fôlego>, <acesso, complicado>, <agua transparente>, ...}; conjunto de palavras conhecidamente positivas B = {bom, ótimo, maravilhoso, excelente, gostoso, delicioso, ...}; conjunto de palavras conhecidamente negativas R = {ruim, péssimo, horrível, ...}.

A polaridade de cada par P_i do conjunto P pode ser definida em função das distâncias de informação entre P_i e os conjuntos B e R. Deste modo, para realizar o cálculo das distâncias para o conjunto <acesso, complicado> serão executadas as seguintes pesquisas em um buscador fictício que contém um total de páginas indexadas de aproximadamente 100.910.000.000:

acesso complicado é bom – 283.000 resultados;

“acesso complicado” – 3.810.000 resultados;

bom – 31.300.000 resultados;

acesso complicado é ruim – 2.320.000 resultados;

ruim – 280.000.000 resultados.

N = 100.910.000.000

Substituindo na fórmula temos:

$$NGD_{(P_i,B)} = \frac{max\{\log_2 3.810.000\,, \log_2 31.300.000\} - \log_2 283.000}{\log_2 100910000000 - min\{\log_2 3.810.000\,, \log_2 31.300.000\}}$$

$$NGD_{(P_i,B)} = \frac{7}{14{,}69291863}$$

$$NGD_{(P_i,B)} = 0{,}462074076$$

Assim, obtém-se a distância entre o par “acesso complicado” e o conjunto de palavras positivas. Da mesma forma, calcula-se a distância do par para o conjunto de palavras negativas:

$$NGD_{(P_i,R)} = 0{,}244555232$$

Como no exemplo fictício <acesso, complicado> possui uma distância 0,462 do conjunto B e 0,244 do conjunto R, pode-se inferir uma polaridade negativa, haja vista a menor distância de informação para o conjunto de palavras negativas.

Nesta seção foram apresentados os principais conceitos da Análise de Sentimento, alguns dos quais serão utilizados neste trabalho. Na seção a seguir será apresentada uma visão geral dos conceitos da área de NLP, para o tratamento de textos não estruturados, escritos em linguagem natural.

2.2. Processamento de Linguagem Natural (NLP)

Segundo Domingues (2011), a linguística computacional surgiu da interseção entre a linguística e a inteligência artificial, investigando a linguagem e as línguas naturais através do emprego de recursos computacionais. Menuzzi (2005 *apud* DOMINGUES, 2011, p. 15) classifica NLP como sendo uma subárea da linguística computacional que estuda a linguagem com foco na construção de ferramentas específicas para o processamento de informação textual e falada.

“Linguagens são feitas de palavras que se combinam via morfossintaxe para codificar significado na forma de frases e sentenças” (BALDWIN e KIM, 2010, tradução nossa). De acordo com Liddy (1998), NLP é um conjunto de técnicas computacionais para a análise e representação de textos, sendo que a análise pode ser feita em um ou mais tipos.

Conforme Luger (2004 *apud* DOMINGUES, 2011, p. 39), a Análise Fonética estuda o relacionamento entre as palavras e os sons, enquanto que a Análise Morfológica estuda os morfemas que constituem as palavras, tais como os prefixos, sufixos e radicais. Segundo Liddy (1998), a análise Léxica concentra-se ao nível de palavra, incluindo o significado léxico. A análise Sintática aborda “as regras para se combinar palavras, frases e sentenças e o uso destas regras para analisar e gerar sentenças corretas” (LUGER, 2004 *apud* DOMINGUES, 2011, p. 39). Já a Análise Semântica, trata do “significado das palavras e de como esses significados se combinam em sentenças para expressar uma determinada ideia” (DOMINGUES, 2011, p. 39). Embora didaticamente seja viável apresentar os diversos tipos de análise da língua de forma independente, na prática o que se observa é uma grande interdependência entre os mesmos.

As ferramentas de processamento de linguagens naturais inicialmente eram elaboradas utilizando-se a abordagem baseada em regras, na qual os profissionais da computação e da linguística codificavam o conhecimento diretamente no programa. Hoje, com o aumento da participação da computação estatística, as técnicas evoluíram e utilizam modelos estatísticos extraídos de grandes corpus. Um corpus é uma grande coleção eletrônica de textos que é amostrada de modo a ser representativo em uma linguagem ou variação desta (XIAO, 2010, p. 147). Existem ainda, sistemas que combinam estas duas abordagens, por regras e por modelos estatísticos, extraindo o melhor de cada uma.

Nos sistemas baseados em regras são obtidos melhores resultados (SILVA, 2013), porém são difíceis de codificar e dar manutenção (GÜNGÖR, 2010), pois especialistas humanos são os responsáveis por codificar as regras computacionais sobre como processar a informação diretamente nos programas. Silva (2013, p. 6) cita como exemplos: programas autômatos de estado finito para tokenização; gramática de restrições utilizadas no marcador Palavras (aplicativo web no site do projeto VISL[3] que “é o estado da arte em etiquetagem *Part-of-speech* (PoS) para o português” (BICK, 2000 *apud* SILVA, 2013, p. 9)); e redes de transição aumentada no corretor ReGra (corretor ortográfico e gramatical utilizado no Microsoft Office e

[3] http://visl.sdu.dk

desenvolvido em parceria entre a USP e Itautec/Philco (NUNES e OLIVEIRA, 2000 *apud* SILVA, 2013, p. 16)).

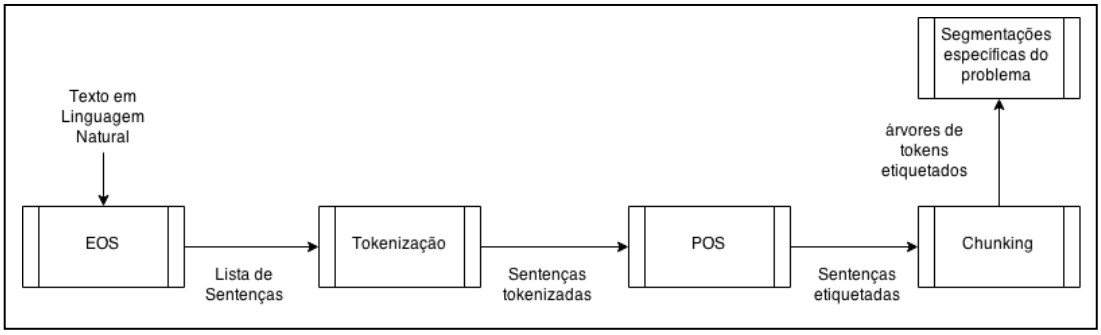
Nas estratégias baseadas em processamento estatístico, técnicas de aprendizado de máquina, tais como *Hidden Markov Models (HMM), N-Gram* e *Maximum Entropy*, são utilizadas para a extração de modelos estatísticos de um grande corpus (SILVA, 2013).

Devido a grande complexidade existente nas tarefas relacionadas ao NLP, o processamento é dividido em algumas etapas. Geralmente as ferramentas apresentam, para cada etapa, módulos que podem ser utilizados independentemente, de modo que seu correto entendimento faz-se importante.

2.2.1. Etapas no NLP

O NLP pode ser altamente complexo e por isso costuma ser dividido em sistemas menores (SILVA, 2013). Nadkarni, Ohno-Machado e Chapman (2011), bem como Russell (2013) o dividem em cinco tarefas de baixo nível. Estas tarefas são executadas sem levar em consideração o motivo final para o qual a saída será gerada. Geralmente cada etapa constitui um módulo nos sistemas de NLP, os quais podem ser utilizados em conjunto com os demais ou isoladamente. O processo está representado na figura a seguir:

Figura 5 – Processo de NLP



Fonte: Próprio autor

Conforme demonstrado na Figura 5, o processo está dividido em etapas bem definidas, iniciando-se pelo recebimento do texto escrito em linguagem natural e seguindo um fluxo no qual a saída de uma etapa constitui a entrada da etapa seguinte. Estas etapas serão descritas a seguir:

- **Detecção do fim da sentença (EOS):** Esta tarefa é responsável por quebrar um texto em uma coleção de unidades lógicas ou sentenças que serão passadas para as demais etapas de análises (RUSSELL, 2013, p. 191). Por

exemplo, poderiam ser utilizados os sinais de pontuação em um texto para gerar um vetor de sentenças como saída desta etapa.

- **Tokenização:** é a tarefa de quebrar uma sequência de caracteres em *Tokens*. *Tokens* são "instâncias de uma sequência de caracteres em um documento particular que estão agrupadas como uma unidade semântica útil para o processamento" (MANNING, RAGHAVAN e SCHÜTZE, 2009, p. 22, tradução nossa). Nesta tarefa cada frase do vetor de frases da etapa anterior é analisada separadamente, e transformado em um vetor de léxicos. Por exemplo, a frase "Conheci Búzios hoje.", daria origem ao vetor [conheci, Búzios, hoje].

- **Marcação Morfossintática (PoS):** A etiquetagem morfossintática consiste em classificar e etiquetar cada palavra do vetor de *tokens* com as categorias gramaticais corretas (DOMINGUES, 2011, p. 15). Por exemplo, a marcação efetuada pelo sistema PALAVRAS[4] para a frase "Estou amando minhas férias aqui em Angra dos Reis" foi "{[[estar], <vK> <vi> V PR 1S IND VFIN], [[amar] <vt> V GER], [[meu] <poss 1S> DET F P], [[féria] <per> <mon> <temp> N F P], [[aqui] <aloc> ADV], [[em] PRP], [[Angra=dos=Reis] PROP M/F S/P]}". Cada palavra foi retornada a sua forma canônica e adicionada com etiquetas morfossintáticas. A etiquetagem morfossintática será mais bem apresentada posteriormente.

- **Segmentação textual (*chunking*):** A segmentação textual, ou *chunking* em inglês, trata de dividir uma sentença em um ou mais conjuntos distintos de palavras interrelacionadas sintaticamente. Através deste processo, também conhecido por análise superficial ou "*Shallow Parsing*", adiciona-se informação mais geral ou abstrata às marcações PoS efetuadas na etapa anterior, agrupando *tokens* que expressam um conceito lógico. Através deste processo podem-se obter dicas sobre a estrutura das frases em um nível mais abstrato do que ao nível das palavras. Por exemplo, os conjuntos podem ser marcados como frases verbais (VP), nominais (NP), preposicionais (PP), ou adverbiais (ADVP).

- **Segmentações específicas do problema:** Nesta tarefa, a saída da etapa anterior é analisada do ponto de vista de cada problema específico, e novas marcações são efetuadas (RUSSELL, 2013, p. 195; NADKARNI, OHNO-MACHADO e CHAPMAN, 2011, p. 545). A análise de sentimentos em si, na abordagem léxica,

[4] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/complex.php

pode ser vista como enquadrada nesta etapa, pois utiliza as marcações efetuadas até a etapa de *chunking* para descobrir os aspectos e palavras opinativas no texto.

Geralmente essas etapas são executadas em sequência, utilizando a saída de uma etapa como entrada para a próxima. O processo tem início na quebra do texto em sentenças e termina com a segmentação específica do problema, embora alguns problemas não necessitem de todas as etapas do processamento, como por exemplo, os sistemas de Recuperação da Informação, os quais, a depender do nível de implementação das funcionalidades, podem utilizar somente a lematização e a tokenização.

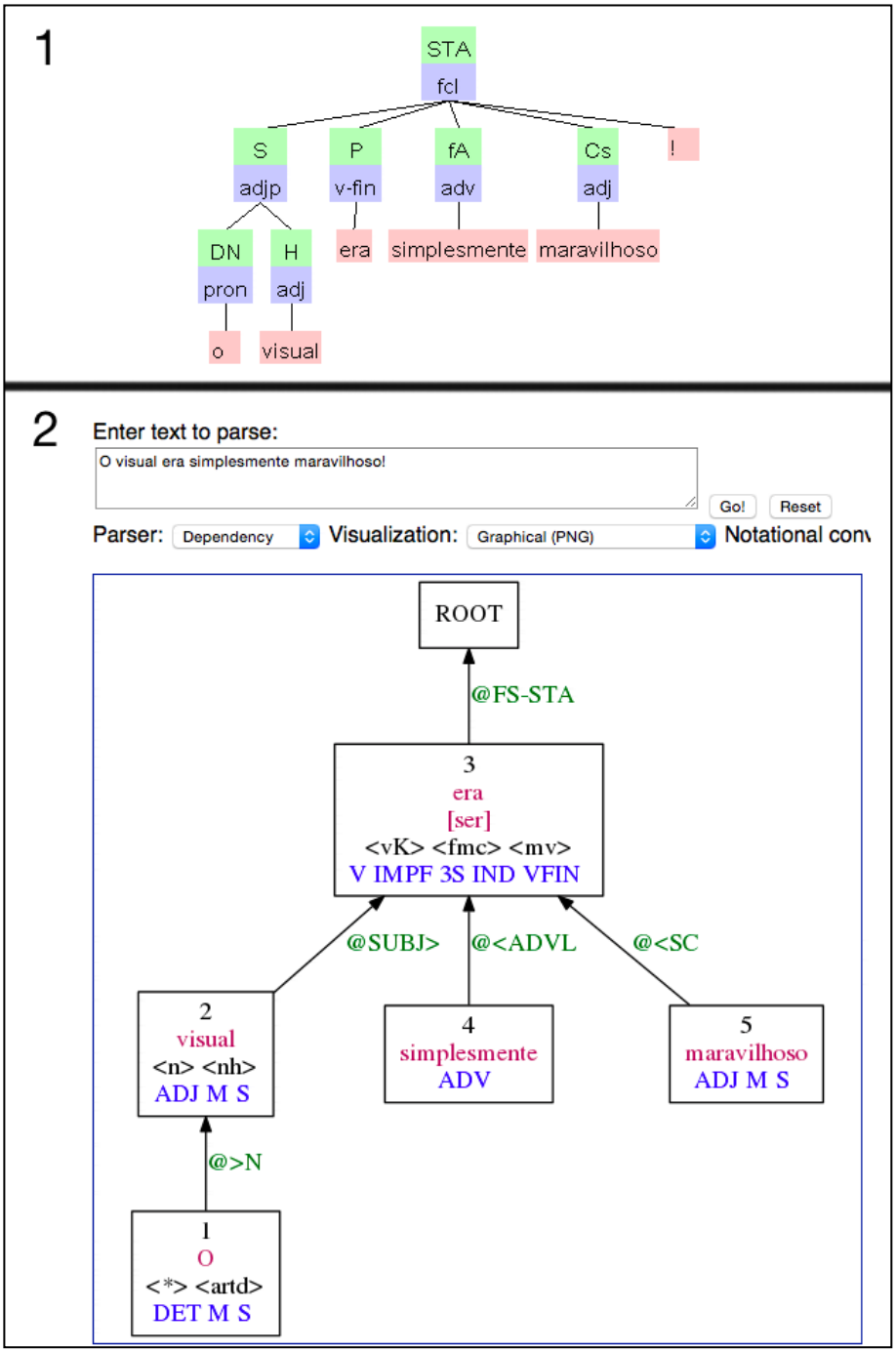
Na etapa de etiquetagem PoS, uma das mais importantes no Processamento de Linguagem Natural, para marcar corretamente cada *token* presente no vetor recebido como entrada, inicia-se pela análise léxica. Segundo Manning, Raghavan e Schütze (2009, p. 32), a análise léxica possui duas tarefas comuns que são a Lematização e o *Stemming*. Elas objetivam reduzir as palavras às suas formas inflexionadas, o que, para idiomas como o português, pode representar uma economia considerável de espaço, devido a sua riqueza morfológica (HIPPISLEY, 2010).

A lematização reduz as palavras à sua forma canônica, ou seja, sua forma mais neutra ou de dicionário (MANNING, RAGHAVAN e SCHÜTZE, 2009, p. 32). Geralmente os verbos ficam no infinitivo e os adjetivos e substantivos na forma masculina singular. O *stemming,* por outro lado, reduz a palavra ao seu radical, cortando o final das palavras através de heurísticas e eliminando afixos derivacionais (MANNING, RAGHAVAN e SCHÜTZE, 2009, p. 32; HIPPISLEY, 2010, p. 32). Por exemplo, enquanto o lema de “viajando” é “viajar”, o *stem* poderia ser “viaj”.

A etiquetagem morfossintática geralmente apresenta como saída os *tokens* em sua forma lematizada, acrescidos de etiquetas com a análise sintática. “A sintaxe é uma parte da gramática que estuda a ordem dos constituintes em uma oração” (TORRES 2012, p. 25). Ela visa analisar uma sequência de palavras a fim de fornecer uma “descrição estrutural de acordo com uma gramática formal” (Ljunglöf e Wirén, 2010, p. 59, tradução nossa). Esta estrutura sintática representa os constituintes da sequência de palavras, classificados nas diferentes classes gramaticais, e suas relações fundamentais de domínio ou precedência (Torres, 2012).

Pustet (2003 *apud* GÜNGÖR, 2010, p. 205) afirma que a maioria dos linguistas considera como principais morfossintáticos o substantivo, o verbo e o adjetivo. Embora alguns modelos linguísticos utilizam outras marcações (GÜNGÖR, 2010), tais como o advérbio, adjuntos e artigos. Segundo o autor, o tamanho do conjunto de marcações costuma ser grande e conter uma vasta gama de *tags*. A figura a seguir representa a saída de um exemplo de análise.

Figura 6 – Exemplos de Processamento de Linguagem Natural



Fonte: Próprio autor

A Figura 6 apresenta dois exemplos de processamentos de linguagem natural efetuadas em português pela ferramenta PALAVRAS[5] para a frase “O visual era simplesmente maravilhoso!”. Na figura 5-1 pode ser observada a árvore sintática, que é “uma representação de todos os passos na derivação da sentença, a partir do nó raiz” (LJUNGLÖF e WIRÉN, 2010, p. 62, tradução nossa). Nela podem ser observados os *tokens* e as etiquetas adicionadas pela marcação morfossintática e pela segmentação textual. Na figura 5-2 observa-se a análise de dependências, que teve origem nas teorias de Lucien Tesnière que se baseia nas relações de dependência entre as palavras da oração e tem o verbo como elemento central.

Observa-se ainda na Figura 6 a presença das etiquetas morfossintáticas. Estas etiquetas dependem do conjunto de etiquetas adotado pelo sistema que está realizando a análise. Por exemplo, “pron” refere-se a pronome pessoal (o), “adj” refere-se a adjetivo (visual), “S” e “@SUBJ” referem-se ao sujeito (o visual). O conjunto completo de *tags* adotado pela ferramenta utilizada para o teste pode ser consultado em [6].

2.2.2. Expressões Multipalavras (MWE)

Um fator de grande importância para as análises textuais é a correta identificação de Expressões Multipalavra, ou em inglês *Multiword Expression* (MWE). MWE podem ser descritas como “uma sequência de palavras que atua como uma única unidade, em algum nível da análise linguística” (CALZOLARI et al., 2002, tradução nossa). Deste modo podemos entender MWE como sendo uma expressão que, embora seja formada por um conjunto de palavras, pode ser vista como sendo uma única palavra. Seu sentido mais amplo é diferente da consideração de cada palavra em isolado. Neste sentido, tem-se MWE como “um conjunto de palavras que possuem um significado próprio e cuja frequência em certo texto tem caráter idiossincrático” (SAG et al., 2011 *apud* TORRES, 2012). Ou seja, a semântica das palavras que compõem a expressão é diferente da obtida ao observar-se cada palavra em isolado. Ainda segundo os autores, as MWE podem ser expressões idiomáticas, compostos nominais, nomes próprios, construções verbais, verbos de suporte ou frases institucionalizadas.

---

[5] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/complex.php

[6] http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html

Neste trabalho serão consideradas as MWE do tipo compostos nominais. Em Boos, Prestes e Villavicencio (2014), são utilizados os padrões morfológicos apresentados a seguir para a detecção de expressões.

Tabela 2 – Padrões morfológicos MWE

Padrão	**Exemplo**
NN	Nações Unidas
NA	Governo Federal
NNN	Supremo Tribunal Federal
NNA	Fundo Monetário Internacional
NAA	Produto Interno Bruto
NPN	casa de praia

Fonte: Boos, Prestes e Villavicencio (2014, tradução nossa)

A Tabela 2 apresenta as composições morfológicas demonstradas pelos autores, segundo a qual N representa substantivos, A representa adjetivos e P preposições. Entretanto a exclusiva utilização dos padrões morfológicos da expressão não é capaz de representar toda a informação existente entre a junção de suas partes componentes.

Neste sentido, em Church e Hanks (1990) é apresentada a medida de associação entre palavras PMI, baseada no conceito de informação mútua da área da Teoria da Informação. A medida pode ser calculada em função da probabilidade de ocorrência de uma palavra y em conjunto com uma palavra x, dada a ocorrência desta palavra x.

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)\,P(y)}.$$

Em Cruys (2011) é apresentada uma generalização da medida PMI, capaz de lidar com mais de duas variáveis aleatórias, observada pela fórmula a seguir:

$$SI_{(x_1,x_2,\ldots,x_n)} = \log_2 \frac{P(x_1, x_2, \ldots, x_n)}{\prod_{i=1}^{n} P(x_i)} \quad (2)$$

Onde:

$P(x_1, x_2, \ldots, x_n)$ representa a probabilidade de ocorrência da expressão em sua totalidade;

$\prod_{i=1}^{n} P(x_i)$ representa a multiplicação das probabilidades de ocorrência de cada um dos *tokens* componentes da expressão.

Para o melhor entendimento da equação considere o cálculo do PMI para a expressão “café da manhã”, considerando que a probabilidade de ocorrência do termo em sua totalidade é de 6,53%, e que a probabilidade de seus componentes é de: café 8,04%, da 38,8% e manhã 7,11%.

Substituindo na fórmula temos:

$$SI_{(café\ da\ manhã)} = \log_2 \frac{P(café\ da\ manhã)}{P(café) * P(da) * P(manhã)}$$

$$SI_{(café\ da\ manhã)} = \log_2 \frac{0{,}0653}{0{,}0804 * 0{,}388 * 0{,}0711}$$

$$SI_{(café\ da\ manhã)} = 4{,}87$$

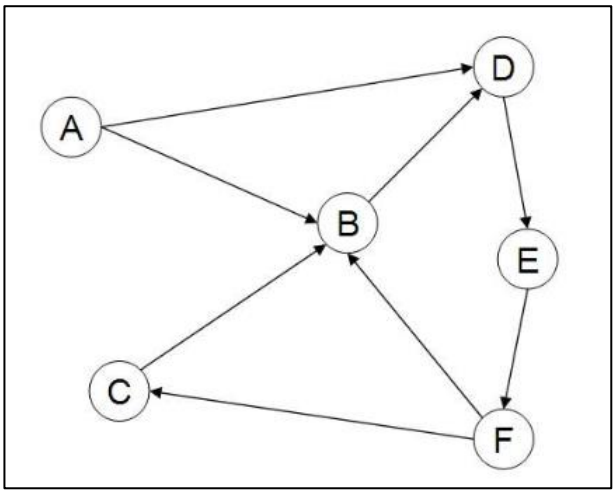
Sem o correto tratamento das Expressões Multipalavras existentes no texto a análise incorrerá na detecção incorreta de aspectos e polaridades. E.g., a avaliação “O atendente estava de cara feia” poderia gerar o aspecto “cara” com a palavra opinativa negativa “feia”, enquanto que na verdade “cara feia” é uma expressão negativa do aspecto “atendente”. Ou então, a avaliação “o cachorro quente estava delicioso” poderia gerar o aspecto “cachorro” com a palavra opinativa “quente”.

2.3. *Web Crawling* e *Web Scraping*

Conforme Manning, Raghavan e Schütze (2009, p. 443, tradução nossa), “*web crawling* é o processo pelo qual coletamos páginas da web”. Ainda conforme os autores, o processo inicia-se com um conjunto de URLs iniciais. Ao visitar cada um destes endereços, o sistema deve coletar seu conteúdo; extrair novas URLs presentes em *links* para outras páginas, e remover o endereço já processado da lista de endereços. O conteúdo coletado, por exemplo, pode ser disponibilizado para um indexador, em mecanismos de busca; ou para um *scraper,* no caso de sistemas de extração da informação.

Conforme Manning, Raghavan e Schütze (2009, p. 443), o processo pode ser visto como o percorrer de um web grafo. De fato, a web pode ser vista como um grafo direcionado, segundo o qual os vértices são as páginas, e as arestas representam os *links* que as interligam. Um exemplo de grafo web direcionado é apresentado a seguir.

Figura 7 – Exemplo de grafo web



Fonte: Manning, Raghavan e Schütze (2009, p. 426)

Na Figura 7 é apresentado um grafo web com seis páginas (de A até F) e oito links interligando-as. A página B possui grau de entrada três e grau de saída um. A página A, por outro lado, tem grau dois de saída, e nenhuma outra página aponta para ela.

Neste sentido, considerando que um *web crawler* receba como endereço inicial para o processo a página F, ele seria capaz de chegar à página C utilizando apenas um nível de profundidade, e à página E com três níveis (F-B, B-D, D-E). Entretanto, o *crawler* não seria capaz de chegar à página A, uma vez que não existem *links* apontando para a mesma. Se a página inicial do *crawler* fosse configurada para a página A, por outro lado, seria possível alcançar todas as demais páginas, embora fosse necessário percorrer um número maior de níveis de profundidade (4 níveis para chegar à página C).

O objetivo de um *crawler* eficiente é coletar o máximo de páginas úteis no menor tempo possível (MANNING, RAGHAVAN e SCHÜTZE, 2009, p. 443). Portanto, conforme os autores, os web *crawlers* devem ser robustos e corretos, i.e., devem ser resilientes a falhas e respeitar políticas implícitas e explícitas que regulam seu funcionamento.

Políticas implícitas referem-se ao volume de solicitações simultâneas realizadas a um mesmo servidor. Se este volume for muito grande, o número de acessos pode prejudicar o funcionamento do site, podendo chegar a indisponibilizar o serviço.

Políticas explícitas referem-se às porções de determinados sites que não devem ser visitados por web *crawlers*. Isto pode ser configurado pelos gestores dos portais através do padrão conhecido como *Robots Exclusion Protocol.*

Além destes dois atributos, os autores apresentam outro conjunto de qualidades que os sistemas deveriam conter, a saber:

- **Distributividade**: Deve poder ser executado em um ambiente distribuído, executando em paralelo em muitos equipamentos distintos.
- **Escalabilidade**: A arquitetura do sistema deve permitir o incremento da taxa de extração através da melhoria do Hardware, aumento de equipamentos ou largura de banda disponível para a tarefa.
- **Eficiência**: O sistema deve fazer uso eficiente dos recursos utilizados, e.g. rede, processador, memória.
- **Qualidade**: O sistema deve ser capaz de extrair as páginas mais úteis antes de páginas de menor qualidade em termos de conteúdo.
- **Novidade**: Capacidade de operar em modo contínuo, detectando possíveis alterações no conteúdo, que deve ser mantido sempre atualizado.
- **Extensibilidade**: Permitir a extensão de suas funcionalidades para que possa assimilar novos padrões de conteúdo, protocolos, e outros.

Como supracitado, uma das principais funções do *crawler* é a extração do conteúdo, seja para processos de indexação, seja para *web scraping*. Entretanto, as páginas web não são formadas simplesmente pelo conteúdo da informação a ser transmitido pela mesma. Adicionalmente existe uma grande gama de dados de *layout*, propaganda, paginação e navegação. Para um usuário humano, a abstração destas informações para interpretar o conteúdo propriamente dito é um processo natural. Contudo, para um sistema computacional extrair as informações presentes em uma página, é necessário um processo capaz de distinguir entre o conteúdo útil para o contexto e o restante.

O processo para extrair dados das páginas web é conhecido como *web scraping*. De acordo com Russell (2013, p. 183, tradução nossa), “O problema não é tão simples como apenas retirar as marcações HTML e processar o texto que permanecer”. Como citado anteriormente, existem diversas informações que também devem ser removidas da página, além das marcações HTML. Também deve ser considerado que, muitas vezes, as marcações HTML possuem informações úteis ao problema em questão.

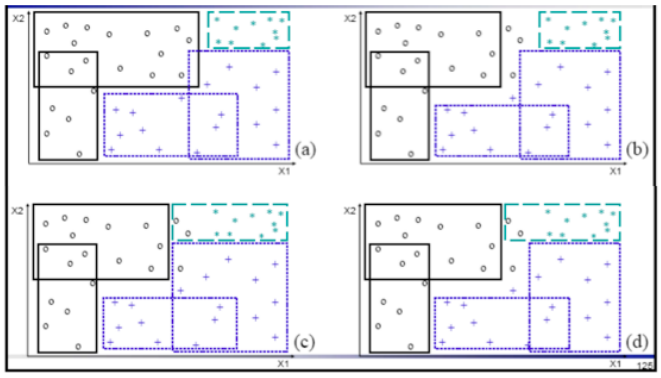
Existem diversas abordagens utilizadas para a extração da informação desejada. Um bom trabalho é apresentado em Kohlschütter, Fankhauser e Nejdl (2010), no qual os autores apresentam um conjunto superficial de características textuais que podem ser utilizadas para classificar os elementos individuais de texto em uma página. Eles comparam sua técnica com uma série de técnicas complexas conhecidas como estado da arte neste tipo de problema.

Geralmente são utilizadas heurísticas e técnicas de aprendizado supervisionado de máquina para a seleção de quais informações devem ser extraídas de determinados *templates* ou padrões de página. Alguns exemplos de ferramentas que podem ser utilizadas para estas tarefas são a API "boilerpipe library"[7] (KOHLSCHÜTTER, FANKHAUSER E NEJDL, 2010), uma biblioteca baseada em Java, e o *web crawler* "Import.IO"[8] (WHITE, PAINTER e FOGG, 2015), uma ferramenta web disponível para a extração de dados de páginas web.

2.4. Avaliação

O processo de avaliação objetiva mensurar o quão bem um sistema executa suas atividades. Em se tratando de problemas de classificação, isso significa observar a precisão e a completude do sistema. Precisão ou consistência refere-se à quantidade de exemplos corretamente classificados, enquanto que completude refere-se à quantidade de exemplos de uma classe abrangidos pelo classificador, o que pode ser observado na figura a seguir:

Figura 8 – Completude e consistência



Fonte: Pozo (2006).

[7] https://code.google.com/p/boilerpipe/

[8] https://import.io

Demonstram-se na Figura 8 quatro classificações diferentes executadas por um algoritmo genérico. Em 6-a, observa-se uma classificação completa e consistente, pois todos os objetos das três classes foram corretamente classificados. Já em 6-b, observa-se uma classificação incompleta, mas consistente, uma vez que nem todos os exemplos foram abrangidos pela classificação (alguns “0” e “+” ficaram de fora), entretanto, todos os que foram classificados o foram de forma correta. Em 6-c, tem-se uma classificação completa, mas inconsistente, uma vez que alguns exemplos foram erroneamente classificados. Já em 6-d, uma classificação incompleta e inconsistente.

Para a correta mensuração destes conceitos, utilizam-se as métricas adotadas pelos sistemas de recuperação da informação, *Precision* e *Recall*. Conforme apresentado em Manning, Raghavan e Schütze (2009 p. 155), estas medidas podem ser definidas por:

$$precision = \frac{tp}{tp + fp} \tag{3}$$

$$recall = \frac{tp}{tp + fn} \tag{4}$$

Onde tp refere-se aos exemplos corretamente classificados, fp aos exemplos erroneamente classificados e fn aos exemplos erroneamente não classificados.

Dependendo da aplicação, é conveniente que se obtenha uma métrica capaz de mensurar estas duas medidas em um único quantificador. Para este propósito utiliza-se a medida “F-Measure”, que conforme Manning, Raghavan e Schütze (2009 p. 156), pode ser definida por:

$$F = \frac{1}{\propto \frac{1}{P} + (1 - \propto)\frac{1}{R}} = \frac{(\beta^2 + 1)\,PR}{\beta^2 P + R}, sendo\ \beta^2 = \frac{1 - \propto}{\propto} \tag{5}$$

Onde P representa o valor de *Precision*, R representa o valor de *Recall*, e $\propto$ um fator que determina o peso para precisão e completude, o qual pode variar de 0 a 1. Com uma distribuição balanceada entre as duas medias, chega-se a medida conhecida como F1, demonstrada por:

$$F_1 = \frac{2\,PR}{P + R} \tag{6}$$

Outra forma efetiva de mensuração do modelo de classificação é a Matriz de Confusão. Trata-se de uma matriz na qual são confrontadas as classificações

realizadas pelo classificador e as classificações corretas. Um exemplo de Matriz Confusão está apresentado na tabela a seguir.

Tabela 3 – Exemplo de Matriz Confusão

		PREDITO	
		Positivo	**Negativo**
REALIZADO	**Positivo**	100	5
	Negativo	10	50

Fonte: Próprio autor

Conforme apresentado na Tabela 3, a Matriz Confusão confronta os valores realizados pelo classificador com as classes reais do conjunto de registros de treinamento. Em sua diagonal principal (destacada) está o quantitativo de exemplos corretamente classificados, no exemplo apresentado, 100 para a classe positivo e 50 para a classe negativo. As demais células apresentam erros de classificação, sendo 10 falsos negativos e 5 falsos positivos para o exemplo apresentado.

Neste capítulo foram apresentadas as teorias que permeiam a análise de sentimentos efetuada em textos escritos em linguagem natural. Foram descritas as principais nomenclaturas e conceitos sobre os três níveis de análise. Sobre a análise no nível de aspectos, foram apresentadas as etapas existentes para a abordagem léxica, na qual os dicionários de sentimentos representam um componente central. Neste sentido apresentou-se algumas técnicas envolvidas no processo de geração automática de lexicons de sentimentos específicos de domínio. Também foram apresentados os conceitos envolvidos no Processamento de Linguagem Natual, outro ponto de fundamental importância para a Análise de Sentimentos na abordagem léxica. Como os maiores volumes de informação pública estão disponíveis na internet, foram apresentadas técnicas de *web mining* para a extração de dados existentes em web sites. A seguir apresenta-se a metodologia utilizada durante a realização deste trabalho.

3 MATERIAIS E MÉTODOS

Este capítulo apresentará os materiais e métodos adotados neste trabalho, o qual pode ser classificado como uma pesquisa aplicada, pois se utiliza de referenciais teóricos para a resolução de um problema do mundo real. No contexto específico deste trabalho, o desenvolvimento de um protótipo analisador de sentimentos em avaliações sobre destinos turísticos escritos em português.

A natureza deste trabalho pode ser classificada como quanti-qualitativa, uma vez que se objetiva a mensuração das opiniões públicas coletivas expressas nos comentários.

3.1. População e Amostra

Para este estudo considerou-se como população a totalidade dos comentários existentes no site de turismo “Trip Advisor”[9]. Desta população foram extraídos como amostragem um milhão quatrocentos e quinze mil comentários públicos. Os comentários estão redigidos na língua portuguesa e apresentam avaliações sobre pontos turísticos, hotéis e restaurantes de destinos turísticos brasileiros.

3.2. Materiais

Na primeira etapa do trabalho foi efetuado o estudo para a elaboração do referencial teórico. Deste modo, foram utilizadas diversas fontes bibliográficas tais como livros, artigos científicos, teses, dissertações e monografias.

Para a elaboração dos exemplos de Processamento de Linguagem Natural apresentados no texto, foi utilizada a ferramenta online PALAVRAS[10]. Ela foi criada pelo projeto de pesquisa e desenvolvimento “*Visual Interactive Syntax Learning*” (VISL) do Instituto de Linguística e Comunicação (ISK) da Universidade do Sul da Dinamarca (SDU). Esta ferramenta é considerada o estado da arte em etiquetagem PoS para o português, atingindo uma taxa de acerto de 99% para a etiquetagem morfológica e de 96% a 97% para a etiquetagem sintática. Entretanto, a ferramenta

[9] http://www.tripadvisor.com

[10] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/complex.php

não possui API disponível para programação, disponibilizando somente uma ferramenta online para a utilização das análises textuais.

Para extrair os dados existentes em páginas HTML disponíveis na internet foi utilizado o *web Crawler* “import.io” (WHITE, PAINTER e FOGG, 2015). A ferramenta oferece várias formas para a extração de informações *online*, sendo utilizada para este trabalho a opção “crawler”. Ao efetuar um cadastro gratuito no site dos desenvolvedores[11], é liberado o acesso ao instalador da ferramenta. As principais vantagens em utilizar-se esta tecnologia foram a facilidade de treinamento da ferramenta e a fácil extração dos dados disponíveis, reduzindo a necessidade de manipulação de marcações HTML.

Para limpeza, transformação e carga dos dados foi utilizada a ferramenta “Kettle – Spoon” (PENTAHO, 2004). Trata-se de uma ferramenta de código aberto, licenciada pela *Apache Licence 2.0*. Com esta ferramenta é possível configurar facilmente, através de componentes visuais e do recurso arrastar e soltar, trabalhos de extração, transformação e carga de dados. Ela suporta uma ampla gama de conectores e extratores, permitindo assim a interoperabilidade com os mais diversos tipos de arquivos e bancos de dados.

Como forma de armazenamento dos dados foi utilizado o Sistema de Gerenciamento de Banco de Dados (SGBD) “SQL Server 2014” (MICROSOFT, 2015).

No desenvolvimento de código Java foi utilizada a *Integrated Development Environment* (IDE) “IntelliJ” (JETBRAINS, 2000). A ferramenta foi escolhida por ser gratuita, de fácil aprendizado e possuir recursos poderosos que facilitam a edição do código fonte.

Para o processamento de linguagem natural em português, foi utilizada a *Application Programming Interface* (API) Java do verificador gramatical CoGrOO (SILVA, 2013). Ele é um projeto *open source* desenvolvido para fornecer ao “Apache OpenOffice” a capacidade de detectar erros comuns encontrados na escrita de textos em Português do Brasil, e utilizado por milhares de usuários (SILVA, 2013).

No processo de análise de sentimentos, para a determinação das polaridades das palavras opinativas, foi utilizado o dicionário SentiLex-PT (SILVA; CARVALHO;

[11] http://import.io

SARMENTO, 2012), um léxicon de sentimentos com 7.014 lemas e 82.347 formas flexionadas em português de Portugal, bem como suas polaridades. Também foi utilizado o dicionário LIWC em português do Brasil (BALAGE FILHO; PARDO; ALUÍSIO, 2013), um dos *lexicons* que compõem o núcleo do Software *Linguistic Inquiry and Word Count* (LIWC), um software comercial de análise textual. Segundo os autores, o dicionário agrupa 127.149 palavras em uma ou mais categorias que o software utiliza para análises psicolinguísticas em textos. Neste trabalho utilizaram-se as categorias "*posemo"* (emoções positivas)(12.878 palavras), e "*negemo"* (emoções negativas) (15.115 palavras).

3.3. Procedimentos

O desenvolvimento deste trabalho foi dividido em três etapas: aquisição dos comentários, pré-processamento e análise de sentimentos.

Na etapa de Aquisição dos comentários foi efetuado um processo que pode ser comparado com o processo *Extract, Transform and Load* (ETL). A Extração dos dados ou *web mining* utilizou um web *crawler* para extrair as cidades (denominadas como destinos); as atrações, hotéis e restaurantes (denominados objetos); e as avaliações de cada objeto. Ao final da extração de cada entidade, um processo específico de Transformação e Carga foi elaborado para inserir os registros em uma base de dados.

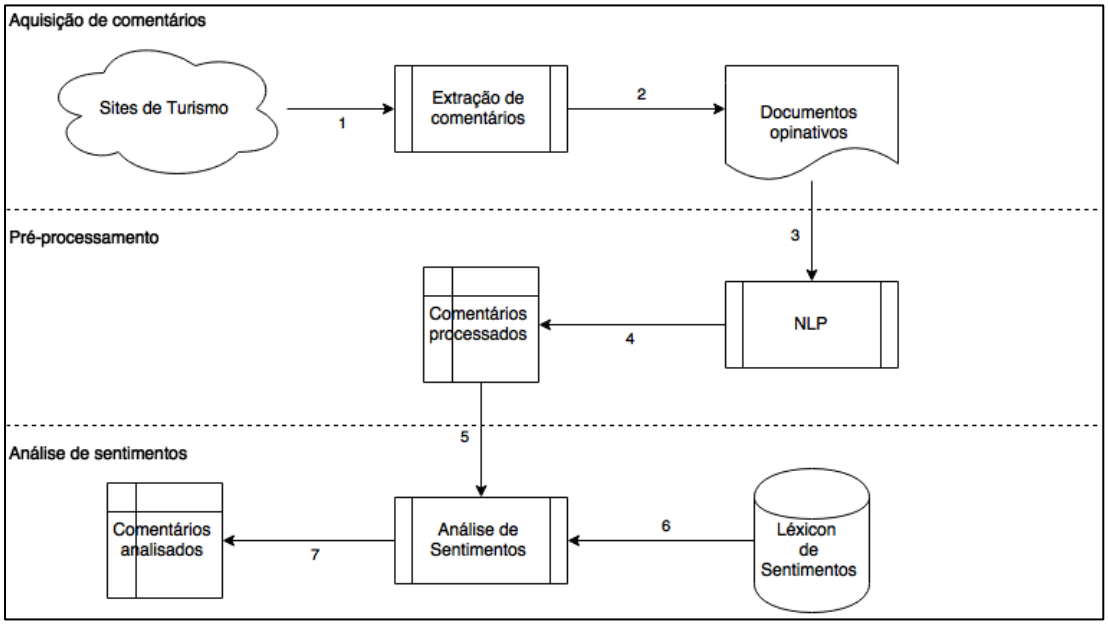
Foram extraídas todas as 2574 cidades brasileiras disponíveis no site "Trip Advisor", e para cada uma foi extraído o número de avaliações disponíveis (totalizando 3.127.461 avaliações), sendo selecionadas para extração as cem cidades mais comentadas. Foram extraídos todos os comentários em português de 20% dos objetos mais comentados em cada tipo (hotéis, atrações e restaurantes).

Na etapa de pré-processamento foi desenvolvida uma ferramenta para a realização do NLP nos comentários extraídos. Utilizou-se a linguagem de programação Java e a API CogrOO para a realização da tarefa. Devido ao grande volume de informações, técnicas computacionais de paralelismo e o devido controle do consumo de memória tiveram que receber especial atenção para que a tarefa pudesse ser realizada em tempo satisfatório.

Na etapa de análise de sentimentos, os métodos e classes responsáveis pela tarefa foram adicionados à ferramenta. Indexações e otimizações foram efetuadas na base de dados para que a análise fosse realizada em tempo satisfatório.

A metodologia está ilustrada na figura a seguir.

Figura 9 – Metodologia



Fonte: Próprio autor

Na Figura 9 observa-se que a metodologia está dividida em três etapas. Na aquisição de comentários, o sistema de *web crawling,* em (1), extrai da internet alguns comentários públicos sobre os pontos turísticos brasileiros. Estes comentários passam a compor em (2) o corpus de comentários extraídos. Em (3), já na etapa de pré-processamento, estes comentários passam pelo processamento de linguagem natural, cujo resultado é indexado e passa a compor em (4) a base de comentários processados. Na etapa de análise de sentimentos, o processo homônimo recebe em (5) os comentários processados e em (6) o léxicon de sentimentos, para determinar as polaridades dos comentários e os aspectos comentados. A análise realizada é então salva em (7).

Apresentados os materiais utilizados bem como a metodologia adotada neste trabalho, o capítulo a seguir descreverá os resultados obtidos.

4 RESULTADOS E DISCUSSÃO

Neste capítulo apresentam-se os resultados obtidos no desenvolvimento da solução proposta. Deste modo, a divisão deste capítulo segue as três etapas de divisão do desenvolvimento do trabalho, cada qual apresentando seus resultados específicos e discussões. Na seção a seguir apresenta-se a aquisição dos comentários, na 4.2, o pré-processamento textual realizado sobre os mesmos, e na seção 4.3 a etapa de Análise de Sentimentos.

4.1. Aquisição dos comentários

A primeira etapa para a realização do trabalho foi a aquisição dos dados a serem analisados. Para a extração dos comentários existentes no site “Trip Advisor”, utilizou-se um processo análogo ao processo de *Extract, Transform and Load* (ETL) no *Data Mining*. Deste modo, inicialmente serão apresentadas as extrações realizadas para a realização desta etapa, seguidas pelas tarefas de transformação e carga dos dados.

4.1.1. Extrações Realizadas

Inicialmente foi configurado um *crawler* na ferramenta “Import.IO” (APÊNDICE A – CRIANDO UM NOVO CRAWLER) para a extração das quantidades de avaliações de Atrações, Hotéis e Restaurantes de todas as cidades brasileiras disponíveis no site “Trip Advisor”[12], o que permitiu a extração do quantitativo de comentários de 2574 destinos turísticos. De posse dos dados extraídos, foram efetuadas tarefas de limpeza e importação. Foram importados os cem destinos com maior número de comentários (Atrações + Hotéis + Restaurantes).

Um *crawler* foi treinado para extrair os objetos e suas quantidades de avaliações para os destinos selecionados. Após a devida importação para a base de objetos, outros três *crawlers* foram treinados, um para cada tipo de objeto, o que permitiu a extração dos comentários propriamente ditos. Em cada tipo (Atrações,

[12] http://www.tripadvisor.com.br/

Hotéis e Restaurantes) selecionou-se para extração dos comentários os vinte por cento de objetos mais comentados em cada destino.

Foi criada uma consulta SQL na base de objetos extraídos, responsável por escrever as URLs dos parâmetros "*Where to start*", "*Where to crawl*" e "*Where to extract data from*" do Import.IO. O texto da consulta, contendo os endereços, foi copiado para o respectivo campo do Import.IO para que a extração ocorresse somente nos objetos selecionados. Um exemplo de resultado retornado pelo SQL Server é demonstrado a seguir para um restaurante de São Paulo.

"*Where to start*": http://www.tripadvisor.com.br/Restaurant_Review-g303631-d1009752-Reviews-or290-Skye_Bar_Restaurante-Sao_Paulo_State_of_Sao_Paulo.html

"*Where to crawl*": http://www.tripadvisor.com.br/Restaurant_Review-g303631-d1009752-Reviews-{any}$

"*Where to extract data*": http://www.tripadvisor.com.br/ShowUserReviews-g303631-d1009752-{any}$

Como pode ser observado, a consulta extrai os códigos de destino (letra "g" seguida de números) e de objeto (letra "d" seguida de números) da URL de cada objeto e constrói as URLs obedecendo à sintaxe da ferramenta import.io. Ressalta-se que a marcação "{any}" (qualquer caractere) seguida da marcação "$" (final da URL) indica que o padrão montado refere-se ao início de uma URL, o que permite ao extrator corresponder endereços como por exemplo:

http://www.tripadvisor.com.br/ShowUserReviews-g303631-d1009752-r259169907-Skye_Bar_Restaurante-Sao_Paulo_State_of_Sao_Paulo.html

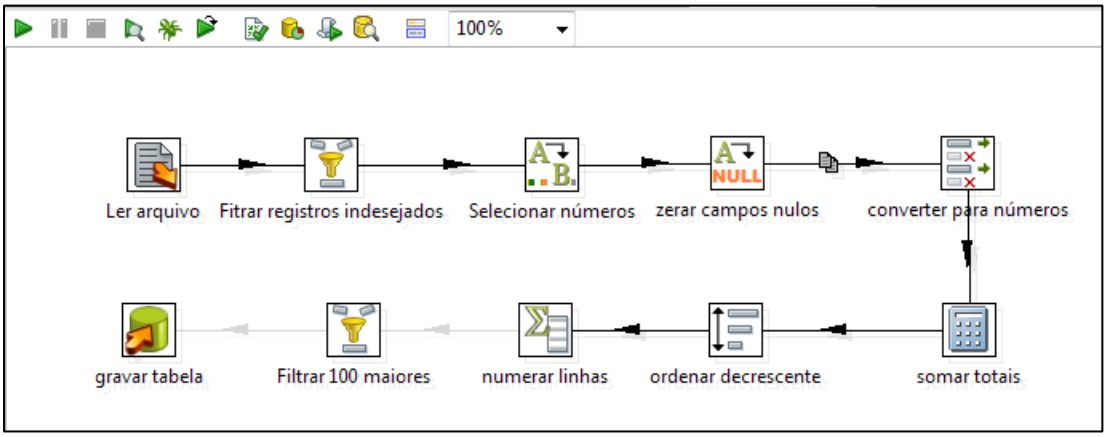
Outro ponto a ser observado é a presença do parâmetro "or290" utilizado na URL do campo "*Where to start*". Este parâmetro é utilizado pelo site "Trip Advisor" para indicar a paginação (no caso, página 30) nos comentários do objeto. A instrução SQL calcula uma página central, até cinquenta níveis abaixo da primeira página, para que a extração atinja o maior número de comentários.

4.1.2. Transformação e Carga:

Para o processo de Transformação e Carga foi utilizada a ferramenta "Kettle – Spoon". Os extratores configurados na ferramenta "Import.IO" geram como saída arquivos csv. Para inseri-los corretamente na base de dados foi necessária a configuração de uma tarefa para: leitura; limpeza e processamento; e inserção destes dados no SGBD SQL Server.

Inicialmente foi configurada uma tarefa para selecionar os cem destinos mais comentados, conforme apresentado na figura a seguir:

Figura 10 – Transformação e Carga de Destinos

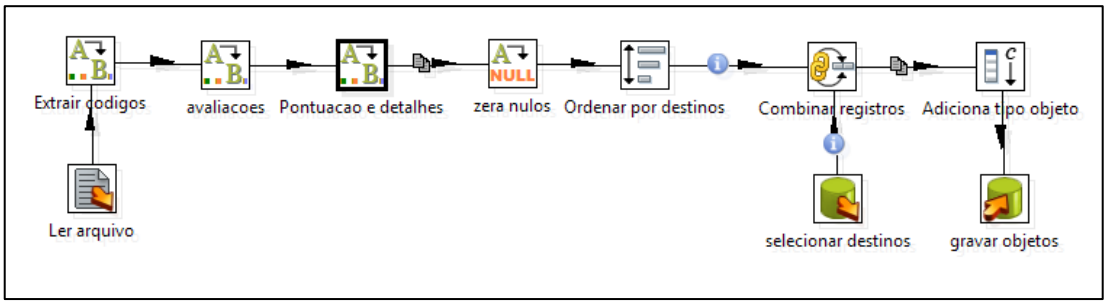


Fonte: Próprio autor

Conforme apresentado na Figura 10, a tarefa inicia-se lendo o arquivo CSV. Em seguida os registros indesejados são filtrados. Dentre os registros filtrados estão, por exemplo, destinos de outros países que foram extraídos pelo *crawler*. O próximo passo executa uma seleção da parte numérica do texto retornado nas colunas de totais, uma vez que o extrator selecionou também o texto que acompanhava os números e o sinal “.” de milhar (e.g., “1.230 avaliações”). Estes campos são convertidos de textos para números e são somados, criando então uma nova coluna com o total. Os registros são ordenados em ordem decrescente por esta nova coluna. Para permitir o filtro dos cem destinos mais comentados, é inserida uma nova coluna com o número da linha e o próximo passo filtra os registros que possuem esta coluna menor ou igual a cem. Após estes processamentos os registros são inseridos na tabela no último passo. A lista completa com os cem destinos e suas frequências pode ser conferida no APÊNDICE B.

Após a importação dos destinos, foram importados os objetos (Atrações, Hotéis e Restaurantes). Da mesma forma foi criada uma tarefa para cada tipo. Para exemplificar é demonstrado na figura a seguir o processo de importação das Atrações:

Figura 11 – Transformação e Carga de Atrações



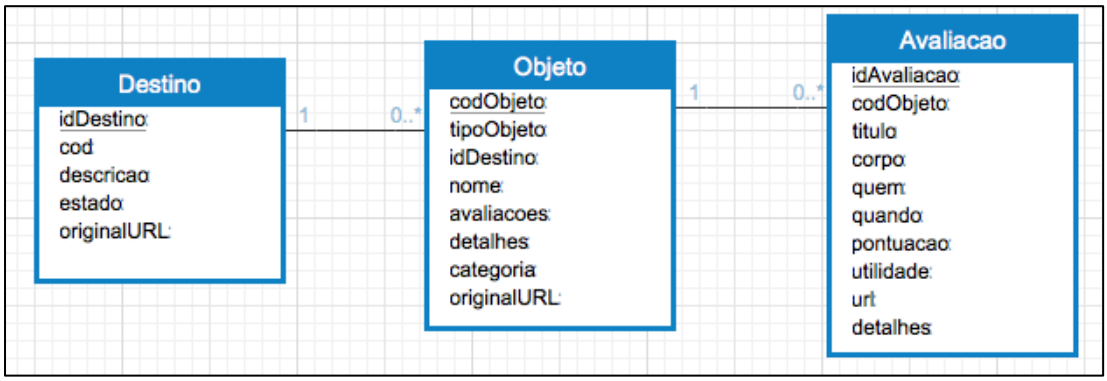
Fonte: Próprio autor

A Figura 11 apresenta a tarefa criada para Transformar e Carregar os objetos na base de dados, cujos passos são descritos a seguir:

- Ler arquivo: o arquivo com os registros é lido e carregado no sistema;
- Extrair códigos: são utilizadas expressões regulares para extrair os códigos de objeto e de destino das URLs presentes;
- Avaliações: através de expressões regulares seleciona-se somente a parte numérica das avaliações (e.g. “1.450 avaliações”);
- Pontuação e detalhes: também através de expressões regulares seleciona-se a parte numérica das pontuações (e.g. “4 de 5 estrelas”). Os objetos do tipo hotéis e restaurantes possuem uma coluna com a pontuação detalhada (e.g. “custo-benefício: 4 de 5 estrelas”), que também são tratadas nesta etapa;
- Zerar nulos: os campos nulos (sem avaliações ou sem pontuação) são substituídos pelo valor 0 (zero);
- Ordenar por destinos: os registros são ordenados por ordem alfabética de descrição dos destinos;
- Selecionar destinos: seleciona os registros de destinos inseridos na base de dados;
- Combinar registros: os dados são combinados com os *ids* previamente inseridos na base de dados;
- Adicionar tipo objeto: é criada uma coluna com a letra referente ao tipo de objeto (“A” para atrações, “H” para hotéis e “R” para restaurantes) e são inseridos na base de dados.
- Gravar objetos: os registros são inseridos na tabela objetos.

Para armazenar os dados extraídos foi elaborado o modelo conceitual apresentado a seguir:

Figura 12 – Modelo conceitual dados extraídos



Fonte: Próprio autor

A Figura 12 apresenta o modelo conceitual da estrutura elaborada para armazenar os dados extraídos, com as principais entidades e atributos. A entidade “Destino” representa as cidades extraídas do site. A coluna “cod” foi extraída da URL dos destinos. Cada cidade ou destino turístico pode conter um ou mais “Objetos”, que podem ser atrações, hotéis ou restaurantes. O atributo “avaliações” contém a quantidade de avaliações indicadas para o objeto em sua página no momento da extração, e engloba comentários de qualquer idioma.

Um objeto pode possuir muitas avaliações, que é a entidade que mantém os comentários dos usuários. O atributo “pontuação" consiste na pontuação em estrelas conferida pelos usuários do site ao objeto avaliado, enquanto que o atributo detalhes contém as pontuações detalhadas. A propriedade “utilidade” refere-se à quantidade de usuários que marcaram este comentário como útil no site.

4.1.3. Discussões

De acordo com os quantitativos apresentados pelo site no momento da extração e com os registros extraídos, elaborou-se uma tabela com algumas estatísticas, apresentadas a seguir.

Tabela 4 – Estatísticas sobre os dados extraídos em 02/01/2015

Descrição	**Quantitativo**
Total de destinos brasileiros com avaliações	2.253
Total de destinos brasileiros sem avaliações	321
Total de destinos brasileiros	2.574
Total de avaliações	3.127.461
Total de avaliações de hotéis	824.667
Total de avaliações de atrações	1.013.830
Total de avaliações de restaurantes	1.288.964
Média de avaliações de destinos	1.205
Média de avaliações de hotéis	423
Média de avaliações de atrações	872
Média de avaliações de restaurantes	633
Mediana de avaliações de destinos	23
Mediana de avaliações de hotéis	18
Mediana de avaliações de atrações	25
Mediana de avaliações de restaurantes	20

Fonte: Próprio autor

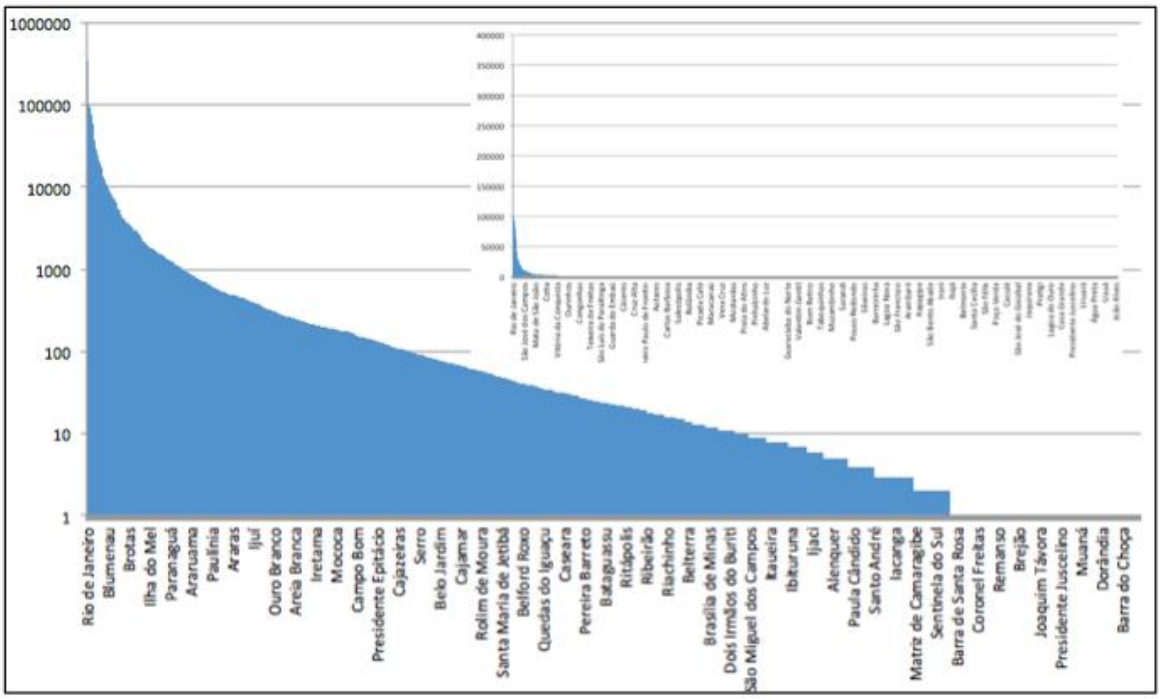
A Tabela 4 apresenta alguns números referentes à extração de destinos turísticos realizada em 02/01/2015. Como pode ser observado, existem muitos destinos turísticos que não têm nenhuma avaliação. Ao analisar-se a distribuição de frequência dos comentários, observou-se uma distribuição do tipo *Power Law*.

Conforme Newman (2005), as distribuições de frequência podem ser distribuições normais, também chamadas Gaussianas, na qual uma faixa de valores apresenta maior concentração de ocorrências. Neste tipo de distribuição as frequências dos demais elementos da amostra tendem a ser igualmente distribuídos entre uma faixa de valores acima e abaixo deste ponto central. Como exemplo deste tipo de distribuição pode-se citar as medidas de estatura de uma população, que tendem a concentrar-se em uma estatura média.

Entretanto, ainda segundo Newman (2005), o tipo de distribuição que comumente é observado ao analisarem-se as frequências de palavras de uma língua é a *Power Law*, segundo a qual um pequeno conjunto dos valores apresenta grande concentração de frequências e uma grande quantidade dos valores ocorre raramente. Outros exemplos de *Power Law* podem ser observados no princípio de

Pareto, segundo o qual 80% da riqueza do mundo está concentrado na mão de apenas 20% da população, na lei de Zipf, e no modelo de incêndios florestais, dentre outros. A distribuição de frequência dos comentários extraídos é apresentada na figura a seguir:

Figura 13 – Distribuição de frequência das avaliações de destinos turísticos



Fonte: Próprio autor

A Figura 13 apresenta a distribuição de frequência dos comentários das cidades brasileiras no site “Trip Advisor”. Como pode ser observado, os comentários estão presentes com uma elevada frequência em um pequeno percentual dos destinos, apresentando uma grande maioria de destinos com um número pequeno de comentários. A figura apresenta a imagem (canto superior direito) da frequência em escala normal e a imagem (ampliada) da distribuição de frequência em escala logarítmica.

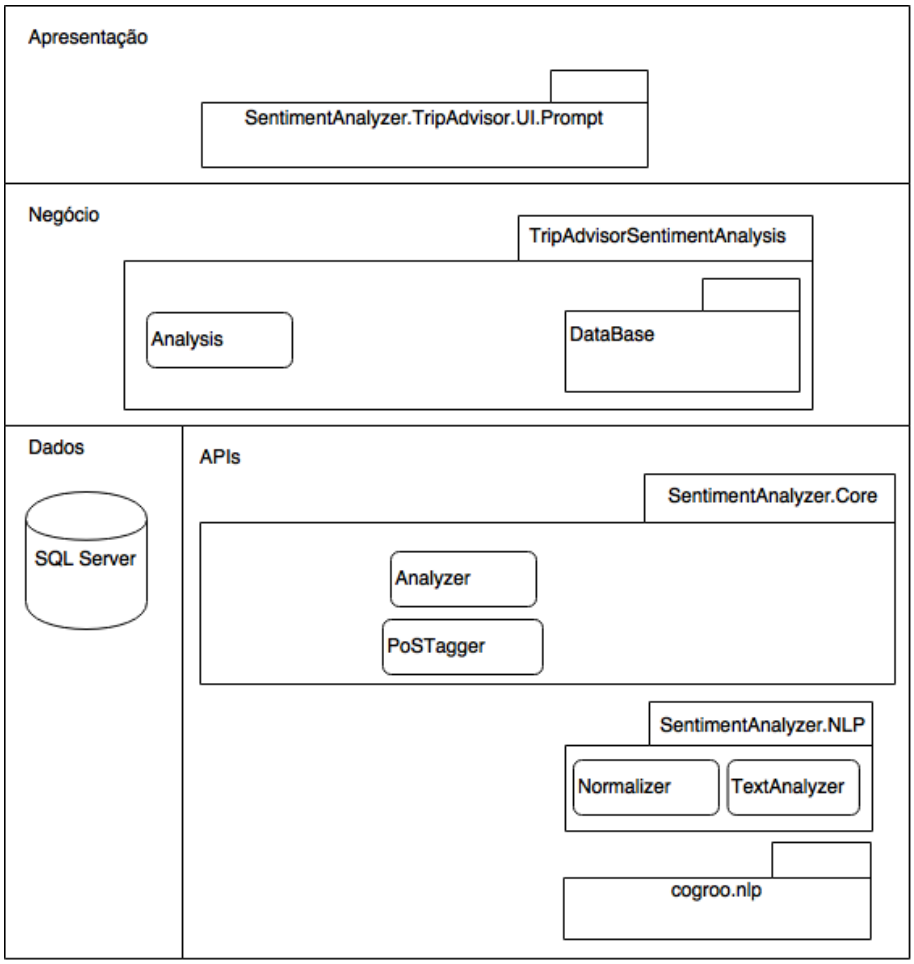
Deste modo, as cem cidades extraídas apresentam um somatório de comentários de 2.655.943, incluindo comentários de todos os idiomas. Por outro lado, somando-se os próximos cem registros em ordem decrescente de comentários, obtêm-se somente 99.372 comentários. Neste sentido, levando-se em consideração o tempo disponível para a realização do trabalho e os recursos computacionais disponíveis, foram selecionadas as cem cidades mais comentadas, o que representa 84,9 por cento do volume de comentários em destinos brasileiros. Entre os meses de

Março e Abril foram efetuadas as extrações dos objetos dos destinos selecionados e de seus comentários.

4.2. Pré-processamento

A segunda etapa para a realização do trabalho foi o pré-processamento dos dados adquiridos na etapa anterior. Para esta etapa foi realizada a análise morfossintática das avaliações, e o resultado foi indexado junto com os termos de cada sentença/avaliação. Para a realização desta tarefa utilizou-se a IDE IntelliJ, na qual foi criada uma estrutura de projeto contendo quatro módulos, representados na imagem a seguir:

Figura 14 – Camadas e principais classes Pré-processamento



Fonte: Próprio autor

A Figura 14 apresenta uma abstração da divisão em camadas existente no projeto desenvolvido para o pré-processamento dos comentários. A primeira camada de abstração representa a interface com o usuário, no caso, uma aplicação do tipo

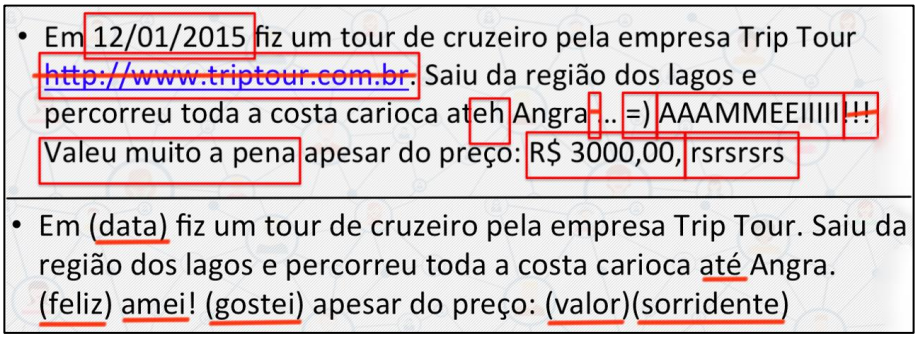
prompt de comando. Esta camada faz chamadas à camada inferior, responsável por manter os códigos específicos para o problema a ser resolvido por este trabalho.

A camada de negócio possui a classe “Analysis”, responsável por carregar os dados existentes na base de dados, enviar para a camada responsável pela análise morfossintática, e salvar o resultado novamente na base de dados. O pacote “DataBase” encapsula as classes e métodos responsáveis pelo acesso, consultas e inserções à base de dados SQL Server, representada na camada inferior de abstração.

Objetivando o reaproveitamento de código, foi criada a camada APIs, que contém os pacotes que são independentes ao contexto deste trabalho, e assim, podem ser reaproveitados em trabalhos futuros. O pacote “SentimentAnalyzer.Core” possui a classe “Analyzer”, que é responsável por realizar as diversas tarefas da análise de sentimentos. No método “PreProcess”, a classe recebe um vetor com os comentários e chama a classe responsável pela marcação morfossintática em seus elementos. Como o CogrOO não recebe uma estrutura de dados como entrada, mas sim um parâmetro do tipo texto, para cada comentário a classe adiciona o título como a primeira frase do texto e executa uma chamada à classe “Normalizer” do pacote “SentimentAnalyzer.NLP”.

Esta classe é responsável por efetuar a normalização do texto, removendo alguns elementos desnecessários à análise e corrigindo outros. Um exemplo de normalização é apresentado na imagem a seguir.

Figura 15 – Exemplo de normalização textual



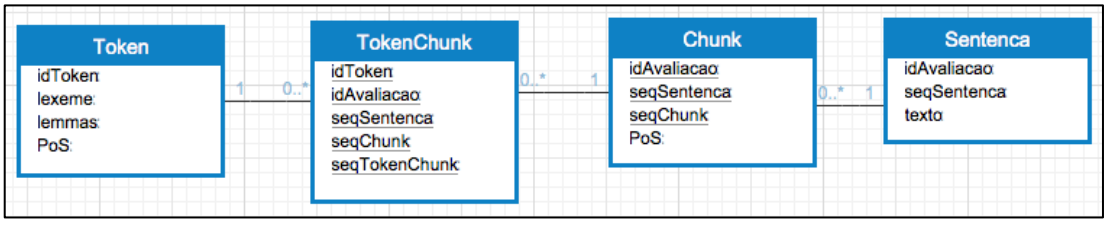
Fonte: Próprio autor

A Figura 15 apresenta um exemplo de normalização textual aplicada pela ferramenta. Como pode ser observado, os itens a serem normalizados estão destacados no primeiro texto, sendo que os elementos a serem removidos aparecem taxados. No segundo texto, os elementos substitutos aparecem sublinhados. Alguns

itens que foram removidos são URLs, e-mails, datas, horas, valores, excesso de espaços em branco, pontuações repetidas, letras repetidas, dentre outros. Palavras todas capitalizadas foram convertidas para minúsculo, entretanto palavras apenas com a primeira letra maiúscula foram mantidas para que o CogrOO as reconhecesse com a etiqueta “prop”, que identifica os nomes próprios. Algumas expressões foram substituídas por outras, como é o caso das risadas (“rsrsrs”, “kkkk”, dentre outros) e dos *emoticons*, no qual foi adotado o conjunto apresentado em Araújo, Gonçalves e Benevenuto (2013). Para a realização destas tarefas uma série de expressões regulares foi aplicada com o intuito de padronizar o texto. A lista completa das expressões regulares e substituições aplicadas pode ser encontrada no APÊNDICE C.

Após a normalização, o texto é enviado para a classe “TextAnalyzer” do pacote “SentimentAnalyzer.NLP”. Esta classe encapsula o uso dos métodos do corretor ortográfico CogrOO, representado pelo pacote “cogroo.nlp”. O analisador morfossintático do “CogOO” quebra o texto em uma lista de sentenças, e cada sentença em um conjunto de *chunks*, cada qual com sua lista de *tokens* e suas respectivas marcações. Para facilitar o processo de análise de sentimentos, e a contagem das frequências dos termos, o resultado da análise realizada no vetor de *reviews* é indexado no banco de dados. Neste sentido foi criada a estrutura apresentada na figura a seguir:

Figura 16 – Modelo conceitual



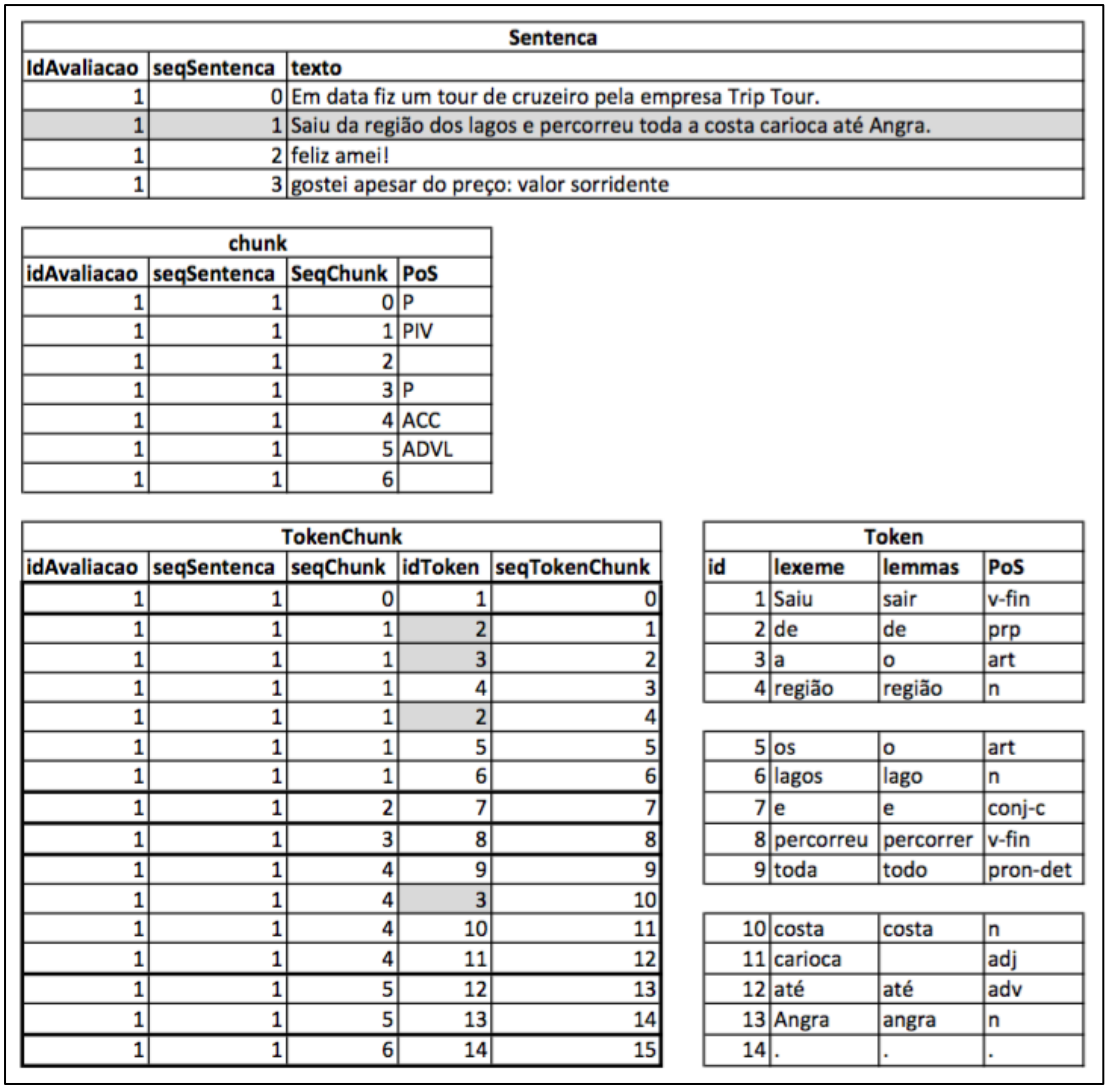
Fonte: Próprio autor

A Figura 16 demonstra a estrutura criada para armazenar o resultado do processamento de linguagem natural realizado nos comentários. Cada avaliação foi quebrada em sentenças, nas quais o atributo texto contém o resultado da normalização. Cada sentença possui vários *chunks*, que representam as partes sintáticas da sentença, cuja etiqueta foi armazenada no atributo PoS. Cada *chunk* contém muitos *tokens*, no qual a coluna “lexeme” representa o texto original da palavra, “lemmas” são as possíveis formas de dicionário da palavra, e “PoS” a

etiqueta morfológica. A entidade “TokenChunk” mantém o relacionamento entre os *chunks* e os *tokens*, sendo o campo “seqTokenChunk” responsável por manter a posição de cada *token*.

Um exemplo de pré-processamento é apresentado a seguir.

Figura 17 – Exemplo de pré-processamento



Sentenca

idAvaliacao	seqSentenca	texto
1	0	Em data fiz um tour de cruzeiro pela empresa Trip Tour.
1	1	Saiu da região dos lagos e percorreu toda a costa carioca até Angra.
1	2	feliz amei!
1	3	gostei apesar do preço: valor sorridente

chunk

idAvaliacao	seqSentenca	SeqChunk	PoS
1	1	0	P
1	1	1	PIV
1	1	2	
1	1	3	P
1	1	4	ACC
1	1	5	ADVL
1	1	6	

TokenChunk

idAvaliacao	seqSentenca	seqChunk	idToken	seqTokenChunk
1	1	0	1	0
1	1	1	2	1
1	1	1	3	2
1	1	1	4	3
1	1	1	2	4
1	1	1	5	5
1	1	1	6	6
1	1	2	7	7
1	1	3	8	8
1	1	4	9	9
1	1	4	3	10
1	1	4	10	11
1	1	4	11	12
1	1	5	12	13
1	1	5	13	14
1	1	6	14	15

Token

id	lexeme	lemmas	PoS
1	Saiu	sair	v-fin
2	de	de	prp
3	a	o	art
4	região	região	n
5	os	o	art
6	lagos	lago	n
7	e	e	conj-c
8	percorreu	percorrer	v-fin
9	toda	todo	pron-det
10	costa	costa	n
11	carioca		adj
12	até	até	adv
13	Angra	angra	n
14	.	.	.

Fonte: Próprio autor

A Figura 17 demonstra como seria o pré-processamento do exemplo apresentado na Figura 15 – Exemplo de normalização textual. Inicialmente o texto é quebrado em sentenças, como demonstrado no preenchimento da tabela sentença. Para efeitos de simplificação, será detalhada somente a segunda sentença, destacada na imagem.

A tabela “Chunk” contém as partes sintáticas da sentença (“P” para predicador, “PIV” para objeto preposicional, “ACC” para objeto direto e “ADVL” para adjunto adverbial. A lista completa de etiquetas pode ser conferida no ANEXO A).

O CogrOO não obteve uma boa classificação sintática no exemplo, como pode ser observado na Figura 17. Por exemplo o *chunk* “da região de os lagos” deveria ser classificado como adjunto adverbial “ADVL”, mas foi classificado como objeto preposicional “PIV”. Entretanto, as etiquetas sintáticas são utilizadas neste trabalho somente para determinar se um *token* específico encontra-se no sujeito da sentença ou no predicado. Verifica-se portanto, se o *chunk* que contém o *token* a ser verificado contém a etiqueta “SUBJ”. Caso afirmativo trata-se do sujeito, caso contrário considera-se como predicado. Esta informação, em conjunto com a etiqueta morfológica, é necessária para a busca por palavras e polaridades no dicionário “SentiLex”.

Os *tokens* estão inseridos na tabela “token” e seus identificadores estão vinculados a cada parte sintática (ou “chunk”) através da tabela “tokenChunk”. Como os *tokens* “de” e “a” se repetem na sentença, observa-se em destaque que somente seus identificadores são repetidos na tabela “tokenChunk”.

Para maximizar a utilização dos recursos computacionais disponíveis, o código desenvolvido vale-se de processamento paralelo, ao implementar a interface Java “*ExecutorService”*. As avaliações são divididas em grupos de x registros, e cada grupo é designado a uma *thread* para o pré-processamento, sendo executadas y threads em paralelo por ciclo. Ao término da execução de cada *thread*, um novo grupo é selecionado e disponibilizado para processamento, em um ciclo iterativo, até não existirem mais avaliações para serem processadas. Para o trabalho atual utilizou-se x = 1000 e y = 4.

4.2.1. Discussões

Após o pré-processamento realizado pela ferramenta desenvolvida, alguns dados foram verificados e são apresentados na tabela a seguir.

Tabela 5 – Estatísticas sobre os dados pré-processados

Descrição	**Quantitativo**
Total de tokens nas avaliações de Restaurantes	41.483.959
Total de *tokens* nas avaliações de Hotéis	27.867.588
Total de *tokens* nas avaliações de Atrações	19.487.532

Total de *tokens* geral	88.839.079
Total de avaliações de restaurantes	734.218
Total de avaliações de hotéis	300.408
Total de avaliações de atrações	380.850
Total de avaliações pré-processadas	1.415.476
Total de sentenças de restaurantes	3.653.648
Total de sentenças de hotéis	1.994.034
Total de sentenças de atrações	1.594.459
Total de sentenças geral	7.242.141
Total de combinações *token* / PoS	444.234
Total de *tokens* únicos (exclusivos)	356.950
Redução de *tokens* exclusivos com normalização (expressões regulares)	36%
Redução de total de *tokens* nas avaliações com a normalização (expressões regulares)	23%
Média de *tokens* por sentença em Restaurantes	11,35
Média de *tokens* por sentença em Hotéis	13,97
Média de *tokens* por sentença em Atrações	12,22
Média de *tokens* por sentença geral	12,26

Fonte: Próprio autor

A Tabela 5 demonstra algumas sumarizações referentes aos dados pré-processados. Conforme pode ser observado, os comentários estão distribuídos entre os tipos Atrações, Restaurantes e Hotéis, e seguem a proporção de 52% para Restaurantes, 27% para Atrações e 21% para Hotéis. O fato do tipo “Restaurantes” ter mais de 50% do total de comentários justifica o maior volume de *tokens* para esta categoria.

Pode-se constatar também o quantitativo de sentenças e a sua distribuição entre os tipos de objetos, seguindo aproximadamente o mesmo percentual de distribuição dos comentários. Com isto, a média de *tokens* por sentença mantém-se bem semelhante entre os tipos de objeto.

Ainda segundo a Tabela 5, observa-se que o quantitativo de *tokens* sofreu uma redução expressiva após a inclusão do método de normalização por expressões regulares, sem, no entanto, haver prejuízo semântico. Foram 36% de

redução quando se observa o quantitativo de *tokens* únicos e 23% no quantitativo de *tokens* por sentença.

Na tabela a seguir são apresentados os *tokens* mais frequentes observados.

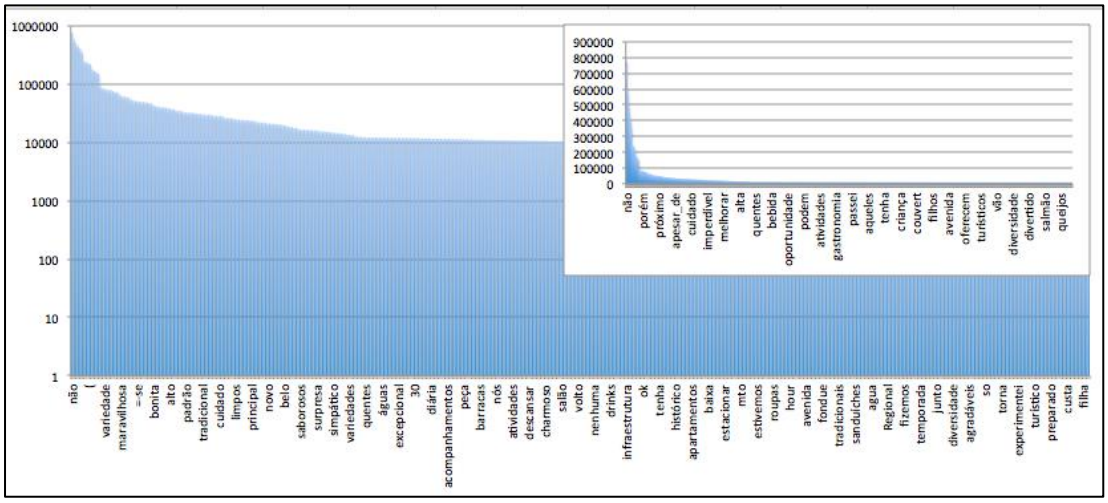
Tabela 6 – *Tokens* mais utilizados

Token	**Frequência**
não	776017
bom	586864
mas	497515
gostei	447673
excelente	400524
boa	351898
ótimo	246232
)	239117
(	234186
melhor	218832
agradável	179201
ótima	168902
"	164102
qualidade	151438
grande	90862
maravilhoso	85063
variedade	83163
nada	80963
porém	79809
melhores	79717

Fonte: Próprio autor

Conforme observado na Tabela 6, apresentam-se os 20 *tokens* mais citados nos comentários, bem como suas frequências. Para a elaboração da tabela foram retirados os *stop-words* apresentados no APÊNDICE F. Observa-se que as palavras mais frequentes do modelo coincidem com palavras bem influentes para a análise de sentimentos, ou seja, palavras de negação, adversativas, substantivos, adjetivos e delimitadores. A remoção de símbolos como os parêntesis, por exemplo, não pode ser efetuada por estarem configurados como delimitadores. A distribuição de frequência entre estes *tokens* pode ser mais bem visualizada no gráfico a seguir.

Figura 18 – Distribuição de frequência dos *tokens* mais frequentes

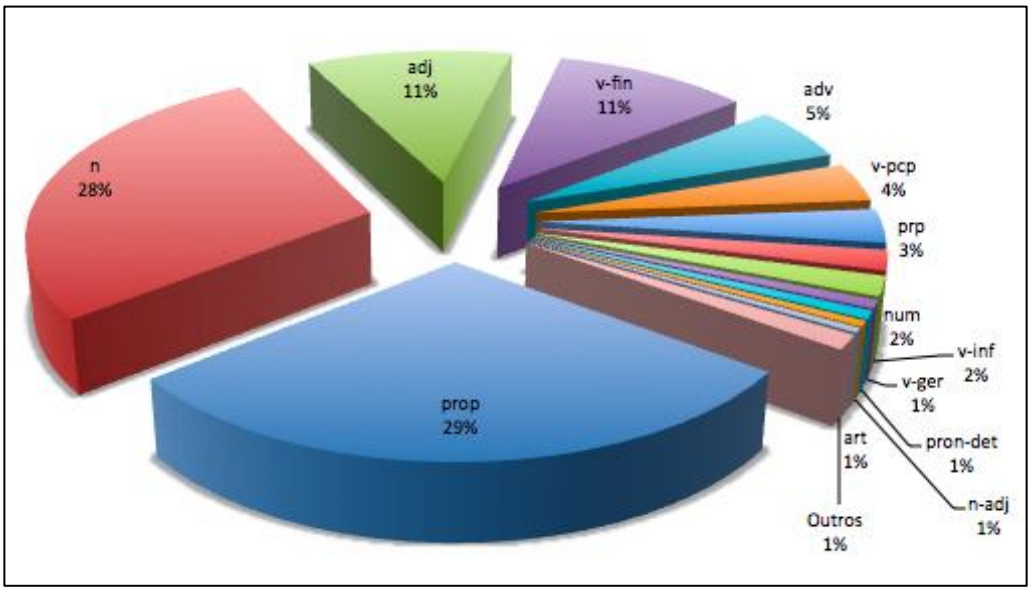


Fonte: Próprio autor

A Figura 18 apresenta a distribuição de frequência dos 500 *tokens* mais citados nos comentários em escala logarítmica (gráfico maior) e em escala normal (gráfico menor). Para a elaboração do gráfico também foram removidas as palavras presentes no APÊNDICE F – LISTA DE STOP-WORDS UTILIZADA. Como pode ser observado, a distribuição de frequência também segue o tipo Power Law, conforme apontado por Newman (2005).

A imagem a seguir demonstra como estes *tokens* estão distribuídos em relação à classificação morfológica realizada pela ferramenta.

Figura 19 – Proporção de *tokens* em relação a etiquetagem PoS

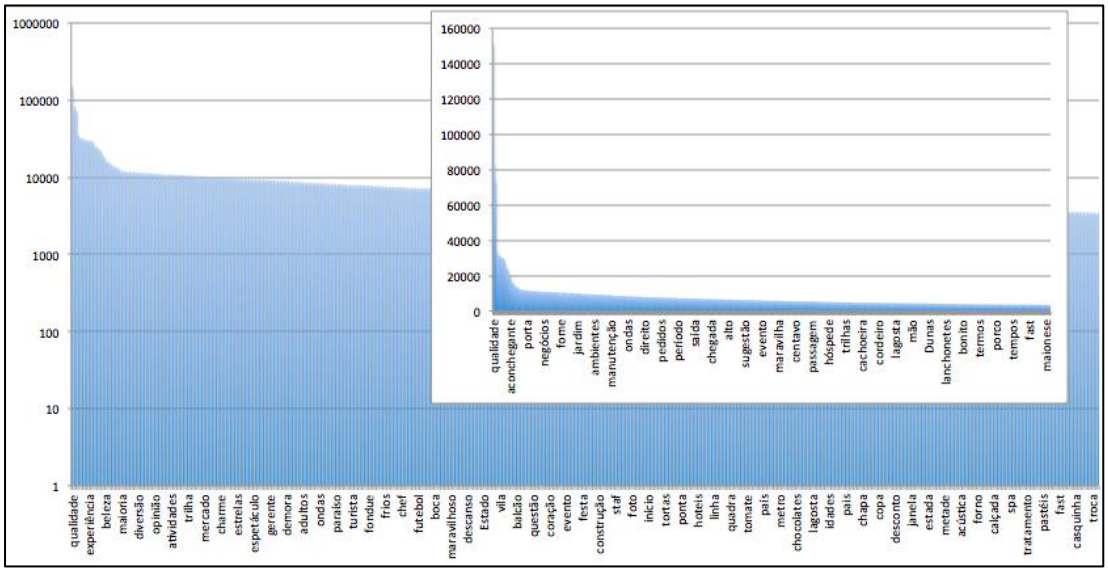


Fonte: Próprio autor

A Figura 19 apresenta como o quantitativo de *tokens* pré-processados está distribuído entre as diversas etiquetas morfológicas definidas pela API CogrOO. Observa-se que os maiores quantitativos encontram-se as etiquetas “prop”, que se referem a nomes próprios e “n”, que se referem aos substantivos. Durante os experimentos efetuados, constatou-se que o CogrOO classifica como “prop” palavras desconhecidas (muitas vezes por grafia errada ou palavras de outros idiomas) e palavras com a primeira letra capitalizada, o que justifica o quantitativo elevado de *tokens* nesta categoria. Em seguida aparecem as etiquetas “adj” (adjetivos) e “v-fin” (verbos), precedendo os demais tipos de classes morfológicas que não possuem representatividade significante.

A seguir os *tokens* da classe substantivo são apresentados em mais detalhes.

Figura 20 – Distribuição de frequência dos substantivos mais frequentes

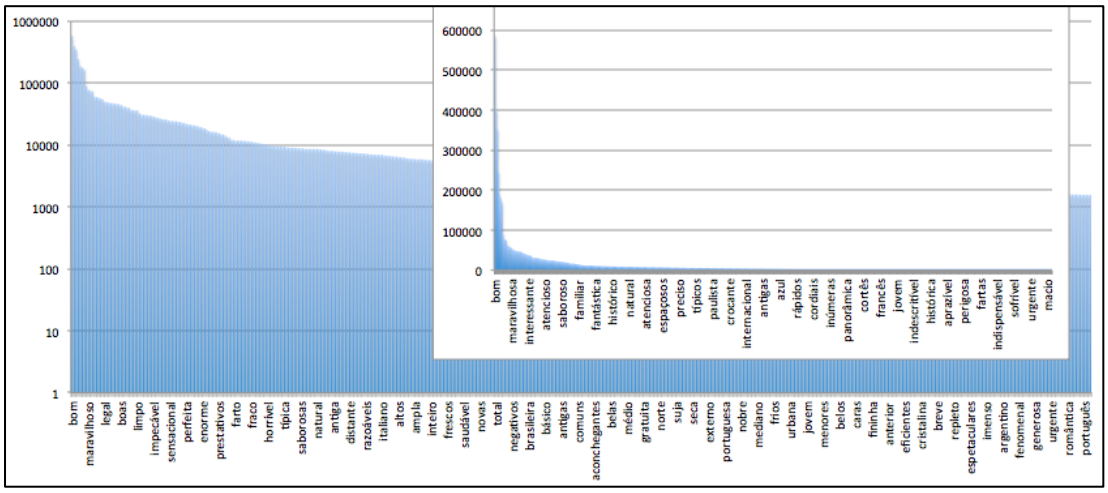


Fonte: Próprio autor

A Figura 20 apresenta os 500 mais citados *tokens* da classe substantivo (“n”), filtrados os *stop-words*, bem como suas frequências. Como pode ser observado, os *tokens* mais frequentes (e. g. Qualidade, experiência, beleza, maioria, diversão, opinião, atividades, trilha, mercado e charme) obtiveram uma classificação morfológica satisfatória, ainda mais se considerarmos a informalidade dos comentários web. A lista completa pode ser conferida no APÊNDICE D.

Na figura a seguir demonstram-se os adjetivos mais frequentes.

Figura 21 – Distribuição de frequência dos adjetivos mais frequentes



Fonte: Próprio autor

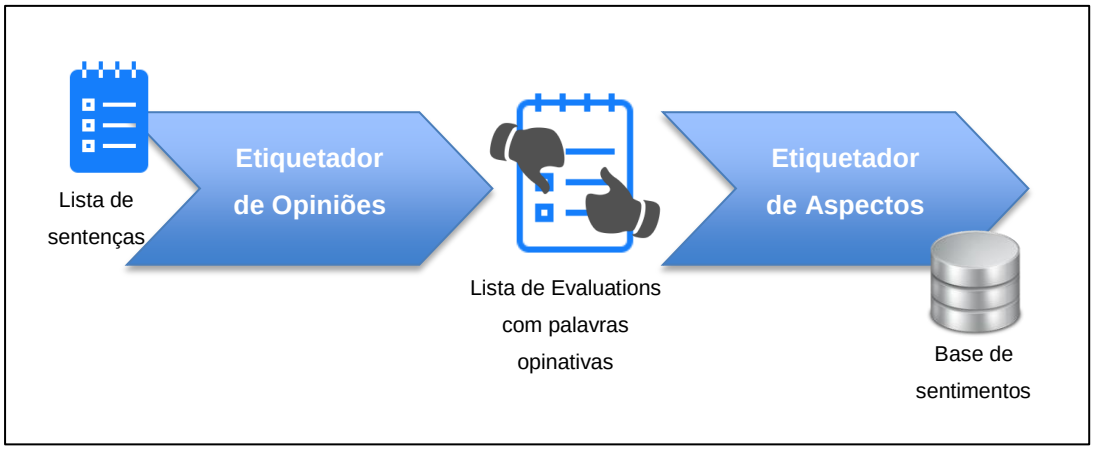
A Figura 21 demonstra os 500 *tokens* mais citados da classe “adj” (adjetivo) (e. g. bom, maravilhoso, legal, boas, limpo, impecável, sensacional, perfeita, enorme e prestativos), retirados os *stop-words*. Tal como a classe substantivo, os adjetivos mais frequentes obtiveram uma classificação satisfatória. A lista completa pode ser conferida no APÊNDICE E.

Concluída a etapa de pré-processamento, iniciou-se a etapa de Análise de Sentimentos, a qual será apresentada em detalhes na seção a seguir.

4.3. Análise de Sentimentos

A terceira e principal etapa para a realização do trabalho foi a Análise de Sentimentos. Neste sentido, uma vez que as avaliações estão divididas em sentenças e *tokens*, a tarefa atual consiste em identificar a correta polaridade das palavras opinativas e identificar os aspectos aos quais elas referem-se. Com efeito, em sua maioria as palavras opinativas são adjetivos ou substantivos adjetivados, e os aspectos são substantivos ou Expressões Multipalavras (MWE). Entretanto, palavras de negação, adversativas e delimitadores devem ser levadas em consideração a fim de determinar corretamente o contexto semântico do comentário. Neste contexto, a imagem a seguir objetiva fornecer uma visão geral de como a ferramenta aborda o problema da AS.

Figura 22 – Tarefas da Análise de Sentimentos efetuada pela ferramenta



Fonte: Próprio autor

A Figura 22 apresenta as principais tarefas executadas pela ferramenta desenvolvida na tarefa de AS. Neste sentido a análise foi dividida em duas etapas: etiquetagem de opiniões e etiquetagem de aspectos. Como demonstrado, o processo tem início com a lista de sentenças etiquetadas pela etapa de pré-processamento sendo fornecida como entrada para o Etiquetador de Opiniões. Após a devida identificação das palavras opinativas e suas polaridades, cada uma é adicionada a um objeto do tipo “Evaluation” que é adicionado a uma lista. Esta lista é fornecida como entrada para o Etiquetador de Aspectos que é responsável pelo preenchimento da lista de aspectos de cada “Evaluation” com os substantivos e Expressões Multipalavras identificados. Por fim a lista de “Evaluations” é enviada para a base de sentimentos analíticos, onde é criado um registro para cada par Palavra Opinativa, Aspecto. Um exemplo pode ser visualizado a seguir.

Figura 23 – Exemplo do processo de Análise de Sentimentos

Sentença:
bom café da manhã, com ambiente e músicas agradáveis e cuidado no atendimento ao cliente.

Evaluations		
palavra opinativa	**polaridade**	**aspectos**
bom	1	café da manhã
agradáveis	1	ambiente; músicas
cuidado	1	atendimento

Fonte: Próprio autor

A Figura 23 apresenta uma lista de objetos do tipo “Evaluations” preenchidos com a análise de uma sentença. Conforme visto na Figura 22, o Etiquetador de Opiniões inicialmente cria a lista e preenche as colunas “palavra opinativa” e

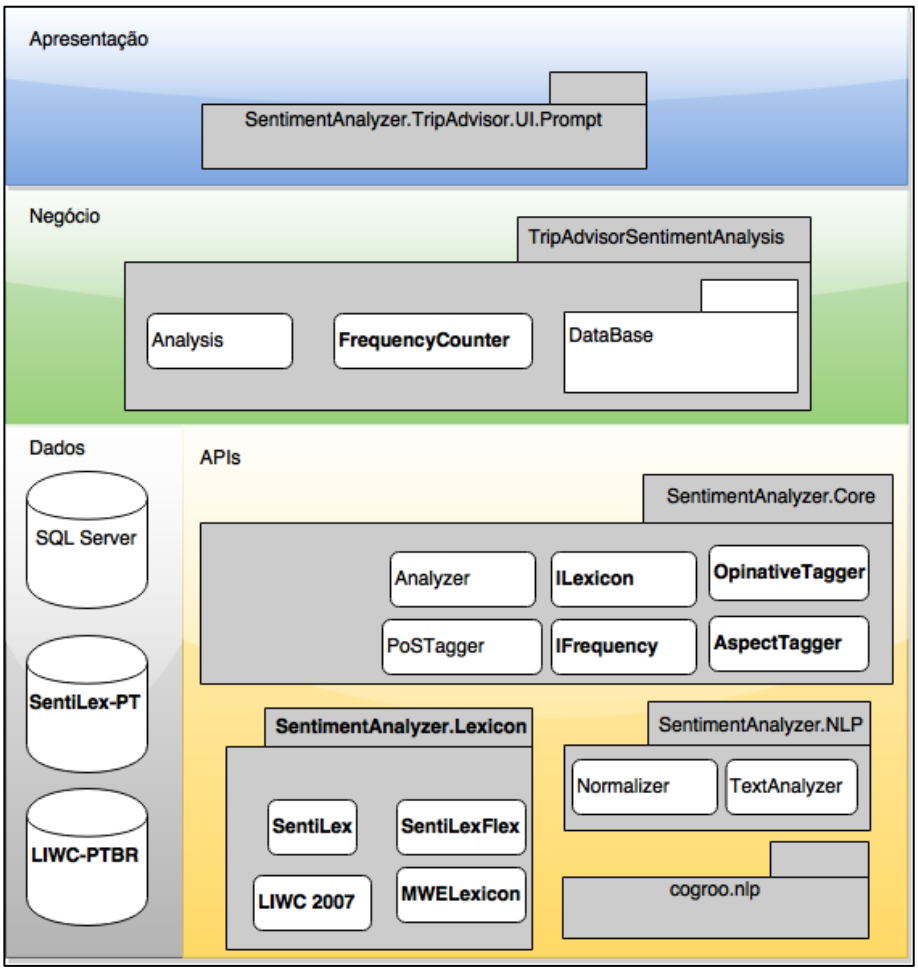
“polaridade”. Em seguida a coluna aspectos é preenchida pelo Etiquetador de Aspectos, e a lista é retornada para que possa ser salva no banco de análises.

A seção a seguir apresentará uma visão geral da arquitetura desenvolvida para a ferramenta, seguida da seção 4.3.2, na qual serão apresentadas em mais detalhes as tarefas envolvidas na primeira etapa da Figura 22 (etiquetagem de opiniões). Após a explanação do processo de identificação de Expressões Multipalavras (seção 4.3.3), as tarefas envolvidas na segunda etapa (etiquetagem de aspectos) serão então abordadas na seção 4.3.4. Por fim, em 4.3.5 apresentam-se alguns dados e discussões sobre a análise efetuada.

4.3.1. Arquitetura

A arquitetura desenvolvida na etapa de pré-processamento foi expandida para comportar as novas funcionalidades, conforme demonstrado na imagem a seguir.

Figura 24 – Camadas e principais classes da ferramenta



Fonte: Próprio autor

Observa-se na Figura 24 a estrutura em camadas, suas *namespaces*, bem como as principais classes desenvolvidas. Em negrito aparecem as maiores alterações em relação à arquitetura desenvolvida na etapa de pré-processamento.

A classe “Analyzer” da camada “APIs” recebeu um novo método “Process”, que encapsula a chamada às duas novas classes “OpinativeTagger” e “AspectTagger”. Elas são responsáveis respectivamente pelas tarefas de identificação de opiniões e identificação de aspectos, as quais serão descritas em mais detalhes nas seções subsequentes. Dentre os parâmetros de entrada, o método recebe uma lista de objetos que implementem a interface “ILexicon” e um objeto que implemente a interface “IFrequency”.

A interface “ILexicon” expõe dois métodos “Short GetPolarity(Token token, String syntacticTag);” e “Short GetPolarity(String multiWordExpression);”. Na camada “APIs”, *namespace* “SentimentAnalyzer.Lexicon” existem implementações desta interface para os *lexicons* “SentiLex-PT” e “LIWC-PTBR” (camada “Dados”). Entretanto, novos dicionários específicos de domínio podem ser adicionados ao processo de AS, bastando implementar a interface em questão.

O dicionário “SentiLex” foi desenvolvido para o contexto jurídico e o “LIWC” é um dicionário livre de contexto. Um diferencial na forma de estruturação do “SentiLex” é a busca por polaridades, que é feita considerando em conjunto a etiquetagem sintática e morfológica, conforme exemplo apresentado na tabela a seguir:

Tabela 7 – Busca por polaridades SentiLex

Sentença	**Palavra analisada**	**Etiqueta Sintática**	**Etiqueta Morfológica**	**Polaridade**
“Ódio é um sentimento de aversão”	Ódio	Sujeito	Substantivo	Neutra
“Odiei a piscina do hotel”	Odiei	Predicado	Verbo	Negativa

Fonte: Próprio autor

Na Tabela 7 é apresentada uma síntese da análise das palavras “ódio/odiei” a partir do uso do Dicionário “SentiLex”. Verifica-se que a polaridade de um termo é expressa a partir da junção de duas informações: uma de origem sintática (termo sujeito/predicado) e outra de origem morfológica (e.g. substantivo, verbo).

Embora esses dicionários obtenham uma boa cobertura no processo de AS, alguns termos não são capturados e outros o são com a polaridade errônea para contextos específicos. Como exemplo, tem-se as palavras “falam” e “acomoda”, que são classificadas como negativas, ou a palavra “recomendo”, que não existe nos dicionários. Estas palavras poderiam ser adicionadas a um dicionário específico para o domínio a ser tratado, no caso o contexto do turismo e, assim, representar uma grande melhoria nos resultados da análise.

A interface “IFrequency” define os métodos “int GetFrequency” e “int GetTotalFrequency”, objetivando retornar respectivamente a frequência de sentenças que contém a expressão e o total de sentenças. Estes métodos são utilizados pela classe “AspectTagger” para a descoberta de Expressões Multipalavra. A implementação deve ser fornecida pela camada de negócio, específica de cada problema, permitindo uma abordagem diferenciada e otimizada para cada situação específica.

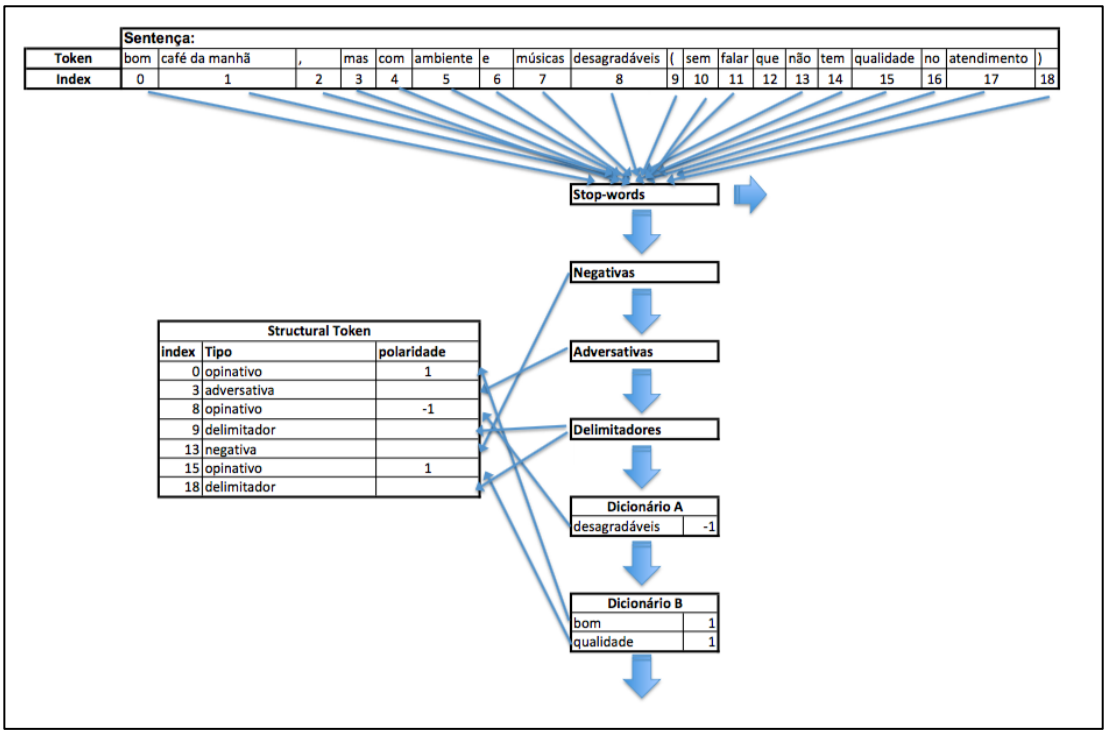
Para o problema da Análise de Sentimentos dos dados deste trabalho, implementou-se a interface “IFrequency” na classe “FrequencyCounter”. Esta classe faz chamadas às classes de acesso ao banco de dados para consultar as frequências das expressões solicitadas. Para otimizar o desempenho, ela mantém uma *hash table* através do objeto Java “HashMap” como um *cache* dos resultados das pesquisas efetuadas. Deste modo, novas consultas são pesquisadas inicialmente no *cache* local, e são pesquisadas no banco somente caso não sejam encontradas.

4.3.2. Identificação de opiniões

A tarefa de identificação das opiniões objetiva a correta identificação de palavras opinativas presentes no texto, a identificação de suas polaridades no contexto atual e a devida extração destas palavras em uma lista de objetos do tipo “Evaluation”. Assim, esta tarefa foi dividida em três passos: identificação de polaridades *a priori*, identificação de polaridades *a posteriori* e preenchimento da lista de *Evaluations*. Na identificação das polaridades *a priori*, a polaridade é definida apenas pelo valor da palavra opinativa no dicionário. Em seguida, a identificação *a posteriori* considera o contexto da palavra na sentença. Estes passos são executados pela classe “SentimentAnalyzer.Core.OpinativeTagger”, método “Execute”.

No primeiro passo (polaridades a priori), objetivando a otimização do tempo de processamento, cada token da sentença é verificado se é opinativo, adversativo, negativo ou delimitador. Os pertencentes a estas classes são mapeados em uma lista de objetos do tipo “StructuralToken”, o qual mantém um apontamento para o *token* original na sentença, e uma propriedade indicando o tipo de *token* (opinativo, adversativo, negativo ou delimitadores). Os *tokens* pertencentes à lista de *stop-words* configurada (APÊNDICE F) são filtrados, somente sendo pesquisados nas demais listas os *tokens* restantes. As listas que foram criadas para este processo de identificação podem ser conferidas no APÊNDICE G. O processo é ilustrado a seguir.

Figura 25 – Mapeamento estrutural e definição de polaridades *a priori*



Fonte: Próprio autor

A figura 25 apresenta um exemplo do processo de mapeamento estrutural, no qual o vetor de *tokens* da sentença é percorrido e cada posição é confrontada com as listas: *Stop-Word*, delimitadores, negativas, adversativas e *lexicons* (Dicionário A e B). As palavras que pertencem à lista de *Stop-Word* ou que não pertencem a nenhuma outra lista são desconsideradas. As listas determinam as palavras que são potencialmente influentes para a análise, e os dicionários determinam a polaridade *a priori* das palavras opinativas.

Os dicionários de sentimentos são passados em um parâmetro que contém uma lista de classes que implementam a interface “ILexicon”. Para cada potencial palavra opinativa, o método faz uma pesquisa sequencial entre os dicionários da lista, de modo que dicionários menores, ou mais especializados em relação ao domínio, devem ser adicionados nas primeiras posições, seguidos dos dicionários mais genéricos. O método pesquisa os demais dicionários somente caso não ache a palavra opinativa ou ela contenha a polaridade neutra.

Como observado na Figura 25, o exemplo traz os dicionários A e B. A palavra “qualidade” é pesquisada inicialmente no dicionário A, e como não é encontrada, o sistema efetua uma nova pesquisa no dicionário B, recuperando assim a polaridade positiva. A palavra “desagradável”, por outro lado, é retornada na primeira pesquisa, uma vez que se encontra no primeiro dicionário.

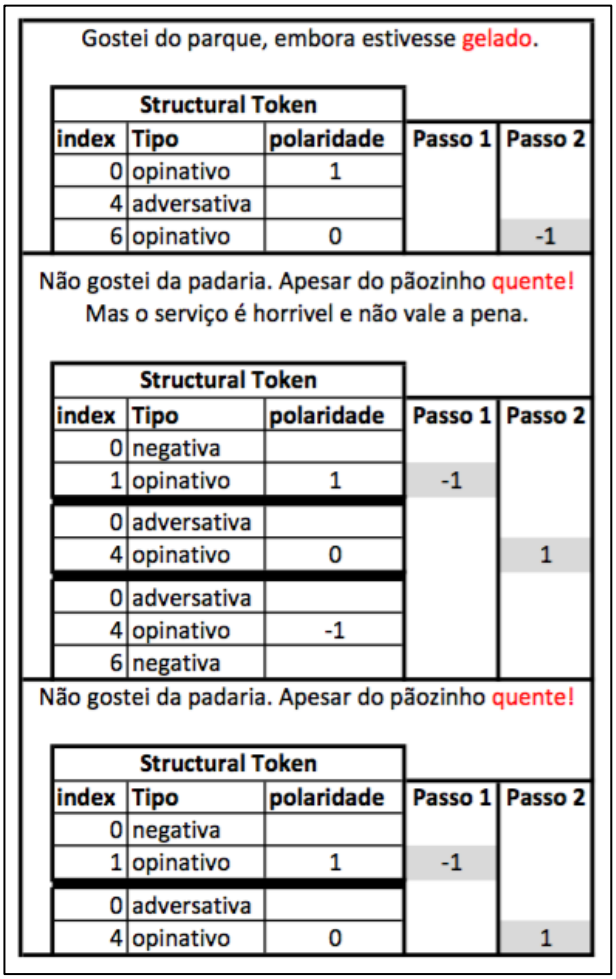
Após o processo de mapeamento estrutural dos comentários, o próximo passo é a determinação da polaridade *a posteriori* dos *tokens* opinativos. São determinados inicialmente os *Tokens* opinativos fortes e em seguida os *Tokens* opinativos fracos. *Tokens* opinativos fortes são aqueles cuja polaridade foi definida pelo *léxicon* de sentimentos, sem a necessidade de observar o contexto semântico da sentença. Neste sentido, os *tokens* opinativos fracos possuem a polaridade zero e os fortes a polaridade diferente de zero.

Para determinar a polaridade final dos *Tokens opinativos fortes*, o método verifica se o *token* estrutural anterior é de negação e se sua distância no vetor original é de até três *tokens.* Em caso afirmativo, a polaridade do *token* é invertida.

Para os *Tokens* opinativos fracos, o método tenta identificar a polaridade observando o contexto semântico do comentário. Para tanto procura, no vetor de “StruturalToken” da sentença atual, a palavra opinativa de polaridade diferente de zero mais próxima que não seja separada por um delimitador, como aspas, parênteses ou chaves (a lista completa pode ser conferida no APÊNDICE G – ADVERSATIVOS, NEGATIVOS E DELIMITADORES). Caso seja encontrada, ele utilizará os *tokens* adversativos e negativos encontrados pelo caminho para determinar a polaridade final. Caso não sejam encontradas palavras opinativas na sentença atual, o algoritmo realiza uma nova pesquisa, desta vez utilizando uma sentença anterior e uma posterior. Caso sejam localizadas, ele utiliza a mais próxima para determinar a polaridade final do *token*, novamente considerando delimitadores, negações e adversativos.

O processo de definição das polaridades fortes e fracas pode ser exemplificado na figura a seguir.

Figura 26 – Exemplo de definição de polaridades *a posteriori*



Gostei do parque, embora estivesse gelado.

Structural Token				
index	Tipo	polaridade	Passo 1	Passo 2
0	opinativo	1		
4	adversativa			
6	opinativo	0		-1

Não gostei da padaria. Apesar do pãozinho quente!
Mas o serviço é horrivel e não vale a pena.

Structural Token				
index	Tipo	polaridade	Passo 1	Passo 2
0	negativa			
1	opinativo	1	-1	
0	adversativa			
4	opinativo	0		1
0	adversativa			
4	opinativo	-1		
6	negativa			

Não gostei da padaria. Apesar do pãozinho quente!

Structural Token				
index	Tipo	polaridade	Passo 1	Passo 2
0	negativa			
1	opinativo	1	-1	
0	adversativa			
4	opinativo	0		1

Fonte: Próprio autor

A Figura 26 apresenta um exemplo simplificado do processo de descoberta de polaridades para palavras opinativas fortes (passo 1) e palavras opinativas fracas (passo 2). Podem-se observar três comentários, cada um com uma palavra opinativa fraca (em vermelho) para ser detectada.

No primeiro comentário nenhuma tarefa foi executada no passo 1, uma vez que não existem *tokens* estruturais de negação antes da palavra opinativa “gostei” (vide *index* 0). No passo 2 verificou-se a existência do *token* opinativo fraco “gelado” (vide *index* 6). Ao buscar-se *tokens* opinativos fortes na mesma sentença, foi detectada a palavra do índice 0, e como existe uma palavra adversativa no caminho, a polaridade foi invertida. O processo geraria a saída “embora_gelado; -1”. Se existisse uma palavra de negação até três palavras de distância da palavra opinativa

fraca, esta seria indicada na composição da “Evaluation”, gerando algo como “embora_não_gelado; 1”.

No segundo comentário, por outro lado, encontrou-se um *token* de negação antecedendo a palavra opinativa forte da primeira sentença, o que inverteu sua polaridade e gerou a saída “não_gostei;-1”. No passo 2 não foram encontradas palavras opinativas fortes na mesma sentença que a do *token* opinativo fraco (segunda sentença, palavra “quente”, índice 4). Identifica-se a existência de duas palavras opinativas fortes nas sentenças vizinhas (“não_gostei” e “horrível”). Como a palavra “não_gostei” está a seis *tokens* de distância da palavra “quente” e a palavra “horrível” está a cinco *tokens*, o algoritmo utiliza a da sentença posterior, por estar mais próxima. Novamente existe um *token* adversativo no caminho, o qual é utilizado para inverter a polaridade e gerar a saída “mas_quente; 1”.

O terceiro comentário apresenta o mesmo texto do segundo, removendo-se a última sentença. Como só existe palavra opinativa forte na sentença anterior, esta é utilizada para a determinação da polaridade do *token* opinativo fraco. Como pode-se observar, o algoritmo utiliza a nova polaridade definida pelo passo 1, e inverte-a devido à existência do *token* adversativo, gerando a saída “apesar_quente; 1”.

Para manter a rastreabilidade da forma de cálculo das polaridades e ajudar a determinar as viabilidades de cada método, foram utilizados diferentes fatores de confiança nas polaridades, apresentados a seguir:

- Polaridades determinadas pelo dicionário e palavras negativas tem os valores -1 e 1;
- Polaridades determinadas utilizando outras palavras opinativas na mesma sentença, os valores -2 e 2;
- As polaridades determinadas por palavras opinativas de outras sentenças possuem os valores -3 e 3;
- E por fim palavras opinativas com polaridade determinada por outras palavras opinativas fracas (com polaridades -2, -3, 2 ou 3), possuem valores determinados pela multiplicação da polaridade do *token* encontrado pelo fator de confiança (2 ou 3, a depender se foram encontradas na mesma sentença ou não).

As combinações de fatores de confiança podem ser visualizadas na imagem a seguir.

Figura 27 – Combinações de fatores de confiança

sentença 1		sentença 2		sentença 3		Fator de confiança
Forte	Fraco	Forte	Fraco	Forte	Fraco	
x						1
		x				1
				x		1
x	o					2
		x	o			2
				x	o	2
x			o			3
			o	x		3
x			o		o	> 3
	o		o	x		> 3

Fonte: Próprio autor

Observam-se na Figura 27 algumas distribuições de *tokens* opinativos fortes (“x”) e fracos (“o”) em três sentenças, bem como os fatores de confiança do *tokens* marcados com a cor de fundo escura. Inicialmente trata-se de *tokens* fortes, determinados pelo dicionário, e portanto o fator é um. O fator de confiança dois aparece nos *tokens* fracos, determinados por uma palavra opinativa forte na mesma sentença. Se a determinação da polaridade ocorre baseando-se em *tokens* opinativos fortes de outra sentença, o fator de confiança é três. No último nível do exemplo estão as palavras opinativas fracas, que foram determinadas por outras palavras opinativas fracas. Seu valor é determinado pela multiplicação dos fatores de confiança dos outros *tokens*, gerando um valor maior que três.

Aplicando os fatores de confiança às polaridades dos exemplos apresentados na Figura 26, obtêm-se a análise apresentada a seguir.

Figura 28 – Polaridades definidas com fatores de confiança

Gostei do parque, embora estivesse gelado.	
Palavra opinativa	**Polaridade**
gostei	1
embora_gelado	-2

Não gostei da padaria. Apesar do pãozinho quente! Mas o serviço é horrivel e não vale a pena.	
Palavra opinativa	**Polaridade**
não_gostei	-1
mas_quente	3
horrível	-1

Não gostei da padaria. Apesar do pãozinho quente!	
Palavra opinativa	**Polaridade**
não_gostei	-1
apesar_quente	3

Fonte: Próprio autor

A Figura 28 apresenta as polaridades com valores multiplicados pelo fator de confiança. As polaridades das palavras opinativas fracas aparecem destacadas e com os valores maiores que um, uma vez que foram determinados por outras palavras opinativas.

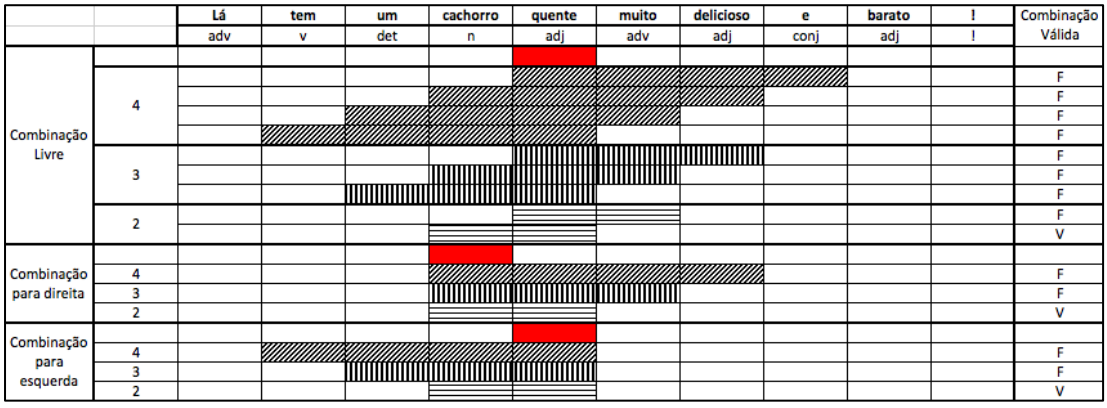
No último passo, após o cálculo das polaridades *a posteriori*, as palavras opinativas são adicionadas a uma lista de objetos do tipo “Evaluation”. Para cada *token* opinativo encontrado, é criada uma “Evaluation” para que a próxima tarefa (identificação de aspectos) possa preencher a lista de aspectos. Antes de explicar o processo de extração de aspectos, porém, faz-se necessário o correto entendimento da identificação de Expressões Multipalavras, o qual será apresentado na seção a seguir.

4.3.3. Identificação de Expressões Multipalavras (MWE)

Para descobrir se uma palavra é parte integrante de uma MWE, o algoritmo verifica as possíveis combinações entre o *token* e as palavras vizinhas, para tentar formar expressões válidas, e caso encontre, calcula o *Pointwise Mutual Information* (PMI) das mesmas. Para filtrar combinações válidas, adotaram-se os padrões apresentados em Boos, Prestes e Villavicencio (2014) (citados anteriormente no referencial teórico; conforme Tabela 2 – Padrões morfológicos MWE). Substantivo e substantivo (n,n); substantivo e adjetivo (n,adj); três substantivos em sequência (n,n,n); dois substantivos seguidos de adjetivo (n,n,adj); substantivo seguido de dois adjetivos (n,adj,adj); ou substantivo preposição substantivo (n,prp,n).

Para a elaboração das combinações de possíveis palavras, o algoritmo adota três abordagens distintas, a depender da situação. As combinações podem ser observadas na imagem a seguir.

Figura 29 – Combinações de palavras na pesquisa de MWE



Fonte: Próprio autor

A Figura 29 demonstra um exemplo de verificação por Expressões Multipalavras para a sentença “Lá tem um cachorro quente muito delicioso e barato!”. Para os casos em que a palavra pode estar localizada em qualquer parte da expressão, ele efetua uma combinação livre; para os casos em que o código está percorrendo os *tokens* da sentença da esquerda para a direita, ele realiza uma combinação para direita; e para os casos em que o código estiver percorrendo da direita para a esquerda, é realizada uma combinação para esquerda.

Em todos os casos, o algoritmo procura formar expressões maiores (quatro palavras) e vai reduzindo a quantidade de palavras até encontrar um padrão morfológico válido (combinação de etiquetas morfológicas apresentadas por Boos, Prestes e Villavicencio (2014), Tabela 2 – Padrões morfológicos MWE) e que possua PMI suficiente, ou chegar a uma única palavra.

Por exemplo, supondo que a sentença está sendo percorrida para frente (Figura 29 – Combinação para direita). Ao encontrar o substantivo “cachorro”, o algoritmo verifica se a palavra faz parte de uma MWE. Para isto inicialmente são selecionados os próximos quatro *tokens* “cachorro quente muito delicioso” e é verificado se a combinação de etiquetas morfológicas obtém um padrão válido. Como o padrão “n;adj;adv;adj” é inválido, o algoritmo reduz a quantidade de *tokens* para três e obtém o padrão “n;adj;adv”, o qual também não é válido. Porém, ao reduzir novamente a quantidade de *tokens* obtém-se o padrão válido “n;adj” e a expressão “cachorro quente” é pesquisada quanto ao seu valor de PMI.

Por outro lado, supondo que a frase esteja sendo percorrida do final para o início, o algoritmo encontraria o adjetivo “barato”. Seria formada a combinação de quatro palavras “muito delicioso e barato”, com a combinação morfológica “adv;adj;conj;adj”. Observa-se que, independente da redução realizada, esta combinação não gera padrões morfológicos válidos, e portanto esta expressão não será verificada quanto a suas frequências, o que representa um considerável ganho de desempenho no tempo de processamento. O processo se repetiria para o adjetivo delicioso (“cachorro quente muito delicioso”, “n;adj;adv;adj”), também sem a obtenção de um padrão morfológico válido. Por fim, ao verificar o adjetivo quente (“tem um cachorro quente”, “v;det;n;adj”), o algoritmo adquire a expressão “cachorro quente” ao reduzir a quantidade de tokens da expressão para dois, e o valor de PMI é calculado.

Para o cálculo do PMI, utilizou-se a fórmula de generalização para mais de duas variáveis, apresentadas em Cruys (2011): ($SI_{(x_1,x_2,\ldots,x_n)} = \log_2 \frac{P(x_1,x_2,\ldots,x_n)}{\prod_{i=1}^{n} P(x_i)}$). Entretanto, a fórmula utiliza um cálculo fundamentado na probabilidade, e fez-se necessária uma maneira de associar a probabilidade à observância da expressão. Para tal, definiu-se a probabilidade de uma expressão como:

$$p_{(x_1,x_2,\ldots,x_n)} = \frac{n(x_1, x_2, \ldots, x_n)}{N}$$

Onde $n(x_1, x_2, \ldots, x_n)$ corresponde ao número de sentenças, para o tipo de objeto atual, onde se observa a ocorrência da expressão pesquisada, e N corresponde ao total de sentenças para o tipo de objeto atual.

Da mesma forma foram calculadas as probabilidades de cada um dos elementos da expressão, e foram aplicados os valores na fórmula anterior. Para obter o quantitativo de sentenças utilizou-se a classe “FrequencyCounter”. Como explicado anteriormente, é a classe que implementa a interface “IFrequency” e é responsável por fazer chamadas às classes de acesso ao banco de dados para consultar as frequências das expressões solicitadas.

O PMI é dito válido caso seja maior que um limite mínimo (configurado como 5, conforme apontado por Antunes e Mendes (2014)) ou caso o PMI esteja dentro de uma faixa de tolerância. Esta faixa é determinada pela combinação PMI/frequência, em termos que, para cada grau de redução no valor de PMI, é exigido um valor parametrizado para a frequência. Por exemplo, uma expressão com PMI 4,8 pode ser considerada como MWE, desde que tenha uma frequência mínima de 1000 ocorrências e uma expressão com PMI 3,2 pode ser considerada MWE desde que ocorra no mínimo 2000 vezes.

Os valores configurados para as tarefas deste trabalho foram: uma tolerância de PMI mínimo de 3, com um mínimo de 1000 ocorrências por nível de tolerância. Desta forma, aceitaram-se MWE com PMI mínimo de 3, desde que possuíssem frequência mínima de 2000. Para a obtenção destes valores foram observadas amostras aleatórias geradas com expressões multipalavras de padrões morfológicos válidos e seus respectivos valores de PMI e frequência. Verificou-se que estas eram as faixas de frequência com as melhores MWE válidas, porém com PMI abaixo do limite mínimo, e portanto necessitavam ser recuperadas. Entretanto são necessários

estudos futuros para determinar corretamente estes parâmetros e maximizar os resultados.

4.3.4. Identificação de aspectos

A tarefa objetiva a correta identificação dos aspectos referenciados pelas palavras opinativas e o devido preenchimento da lista de aspectos em cada "Evaluation". Esta tarefa é desenvolvida pela classe "SentimentAnalyzer.Core.AspectTagger", método "Execute".

Para demonstrar o processo de Identificação de Aspectos será utilizado o exemplo a seguir.

Figura 30 – Exemplo de Análise de Sentimentos

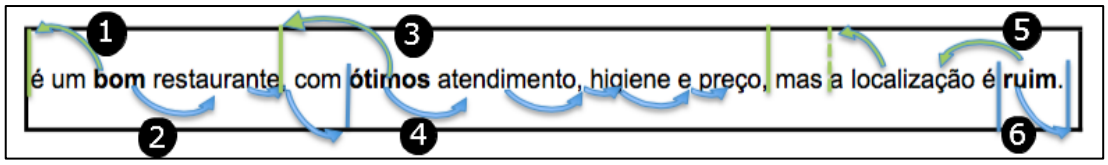
Sentença:
é um bom restaurante, com ótimos atendimento, higiene e preço, mas a localização é ruim.

Evaluations		
palavra opinativa	**polaridade**	**aspectos**
bom	1	restaurante
ótimos	1	atendimento; limpeza; preço
ruim	-1	localização

Fonte: Próprio autor

A Figura 30 apresenta o resultado de uma Análise de Sentimentos efetuada em uma sentença. A lista de "Evaluations", com as palavras opinativas e polaridades detectadas na fase anterior, deve ser preenchida com os aspectos relacionados. Para que esta tarefa seja possível, o método desenvolvido delineia agrupamentos de aspectos. Para o exemplo apresentado na Figura 30: "é um bom restaurante", "com ótimos atendimento, higiene e preço" e "mas a localização é ruim". Neste sentido os *tokens* "e" e ",", e as listas de adversativos e delimitadores são utilizados, porém, nem toda ocorrência dos *tokens* "e" e "," se configuram como delimitadores de grupo. Eles podem representar uma lista de aspectos, como se observa no trecho "com ótimos atendimento, higiene e preço", e portanto devem ser devidamente tratados. A figura a seguir apresenta uma sistematização desses processos.

Figura 31 – Descoberta de aspectos



Fonte: Próprio autor

O processo de descoberta de aspectos referente ao exemplo anterior está representado na Figura 31. Para cada “Evaluation” (palavras opinativas polarizadas) da sentença, o sistema pesquisa por aspectos anteriores e posteriores à palavra opinativa atual. No passo 1 da figura o algoritmo encontra o limite inferior (início da sentença) e como não existem aspectos, passa a procurar para frente (passo 2), no qual encontra o substantivo “restaurante”. O algoritmo então procura um *token* para frente e detecta a presença de uma vírgula. Para determinar se esta vírgula é um indicador de uma lista de substantivos, o algoritmo tenta verificar até três *tokens* à frente se existem outros substantivos. Entretanto, neste caso, é encontrado o limite superior, definido pela presença da próxima “Evaluation” (“ótimos”). A palavra “restaurantes” é então adicionada a lista de aspectos da palavra opinativa atual e o algoritmo passa para o próximo item da lista de “Evaluations”.

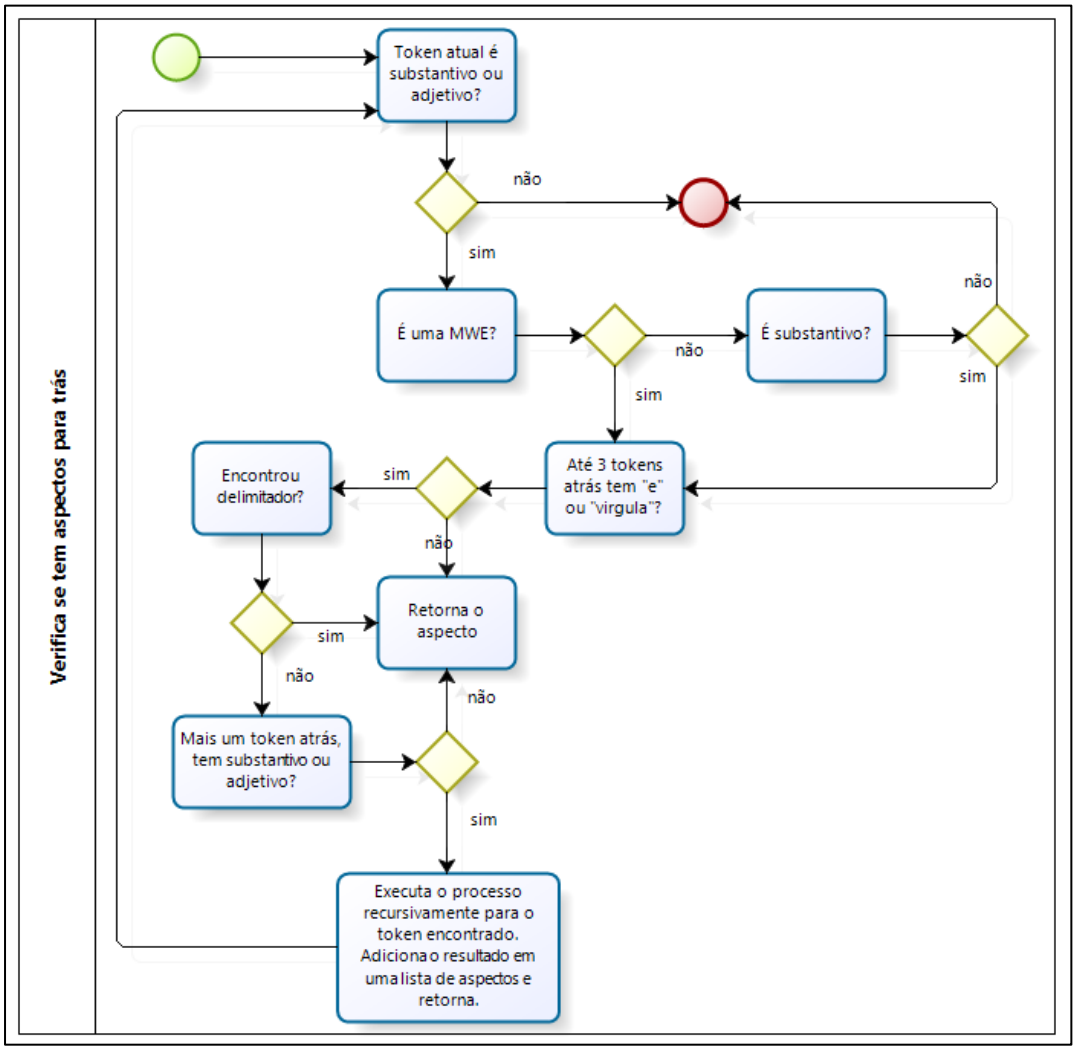
Para a palavra opinativa “ótimos”, o algoritmo procura, no passo 3 por aspectos anteriores e encontra o limite inferior, determinado pelo *token* após o último aspecto da palavra opinativa anterior (“,”). Como não existem aspectos encontrados neste intervalo, o algoritmo passa a procurar para frente (passo 4), no qual encontra o substantivo “atendimento”. O algoritmo então procura um *token* para frente e detecta a presença de uma vírgula. Para determinar se trata-se de uma lista de substantivos, ele tenta verificar até três *tokens* à frente se existem outros substantivos. É encontrado o substantivo “higiene” e um *token* a frente a conjunção “e”. Nesta situação o algoritmo pega somente o substantivo da frente (“preço”) e a lista de aspectos é adicionada à “Evaluation”. Destaca-se que o algoritmo não chega ao limite superior (definido pelo *token* que antecede a próxima palavra opinativa “ruim”), devido a presença da conjunção “e” que indica o final de uma lista de aspectos.

Para o *token* “ruim”, o algoritmo encontra, no passo 5, o substantivo “localização” na pesquisa por aspectos anteriores, e tenta localizar em até três *tokens* anteriores uma vírgula ou “e” para determinar se este aspecto faz parte de uma lista. Entretanto, encontra um *token* adversativo (“mas”) e interrompe a pesquisa. O substantivo “localização” é adicionado a lista de aspectos e como não são encontrados aspectos para frente (passo 6), o processo é concluído.

Para verificar se o *token* atual é um aspecto ou não, as rotinas que foram desenvolvidas estão representadas nos fluxogramas a seguir, o primeiro para

quando a rotina está percorrendo a sentença em direção ao início, e o segundo para quando o método está percorrendo em direção ao final.

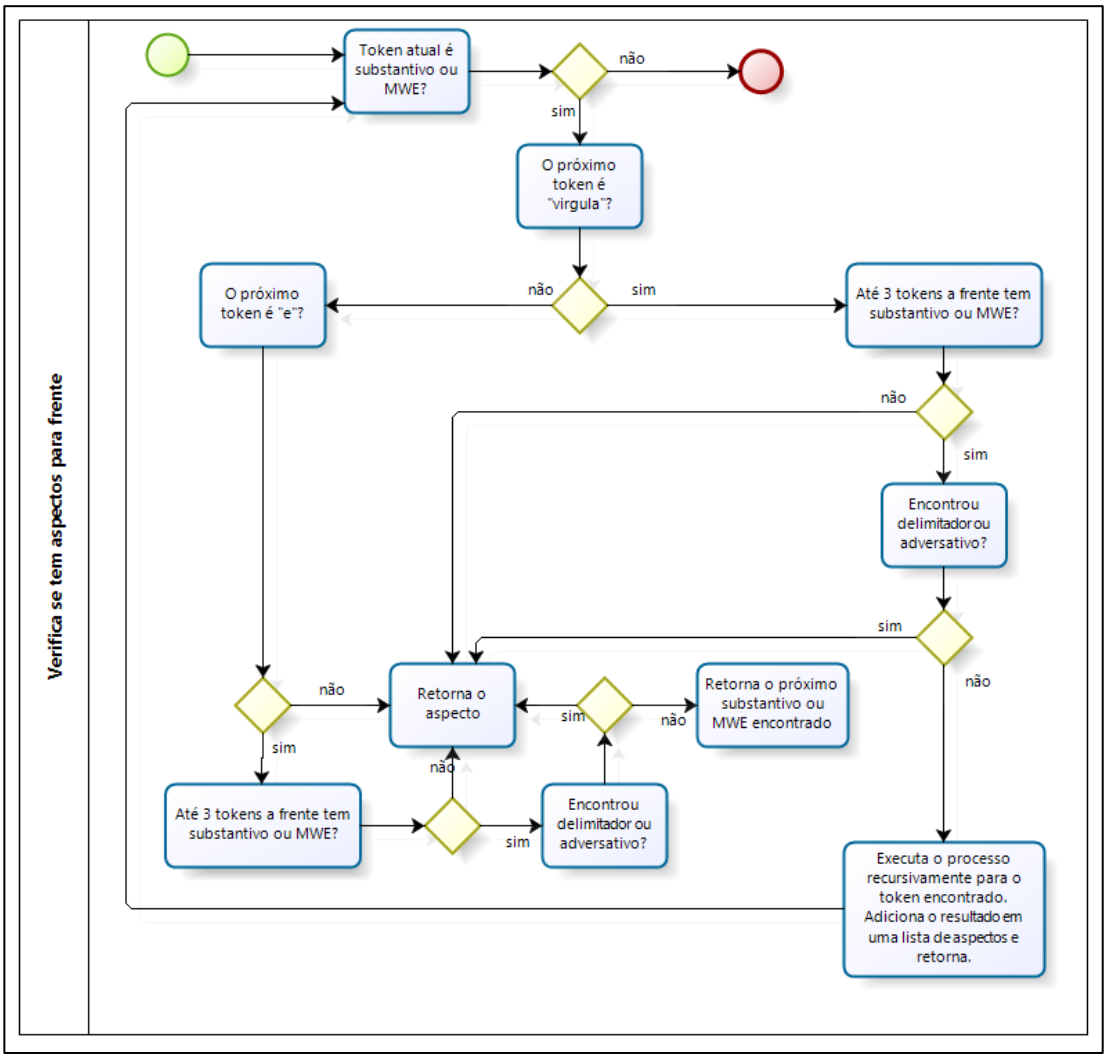
Figura 32 – Fluxo do método de verificação de aspectos para trás



Fonte: Próprio autor

A Figura 32 apresenta o fluxo de decisão para determinar se o *token* atual é um aspecto. Um aspecto pode ser um substantivo ou uma Expressão Multipalavra. Em ambos os casos é verificada a possibilidade de se tratar de uma lista de substantivos separados por vírgula e/ou “e”. E.g. “O café da manhã, o atendimento e o preço são excepcionais!” no qual a palavra opinativa positiva “excepcionais” qualifica a lista de aspectos encontrados para trás “preço, atendimento, café da manhã”. As mesmas regras são aplicadas no processo de pesquisa para frente, demonstrado a seguir.

Figura 33 – Fluxo do método de verificação de aspectos para frente



Fonte: Próprio autor

Demonstra-se na Figura 33 o processo de descoberta se o *token* atual é um aspecto, quando a sentença estiver sendo percorrida para frente. Os métodos são executados recursivamente para recuperar os substantivos que poderão estar contidos em uma lista.

Um ponto a ser destacado nos métodos apresentados são os valores dos parâmetros. E.g. “Até três *tokens* a frente tem substantivo” ou “até três *tokens* a frente tem ‘e’ ou ‘,’”. Os valores de distâncias entre os *tokens* foram definidos arbitrariamente e necessitam de trabalhos futuros avaliando os resultados, de modo a otimizá-los.

Apresentados os métodos desenvolvidos para a realização da Análise de Sentimentos, na seção a seguir serão apresentados alguns dados e discussões desta etapa.

4.3.5. Discussões

Objetivando a comparação de resultados entre os dois dicionários de sentimentos utilizados, foram realizadas três análises. A análise 1 contém os dois dicionários em conjunto, sendo pesquisado inicialmente o SentiLex, e no LIWC somente caso a palavra não fosse encontrada no primeiro. A análise 2 foi realizada exclusivamente no dicionário SentiLex e a análise 3 foi realizada exclusivamente no dicionário LIWC.

O tempo gasto para a realização de cada análise de sentimentos é apresentado na tabela a seguir.

Tabela 8 – Tempos de processamento da Análise de Sentimentos

Análise	**Tempo**
Análise 1 – SentiLex + LIWC	11 horas e 30 minutos
Análise 2 – SentiLex	11 horas
Análise 3 – LIWC	10 horas e 20 minutos

Fonte: Próprio autor

A Tabela 8 apresenta os tempos aproximados gastos por cada análise realizada. As análises foram executadas em dois computadores Intel Core i5-4570 de 3,20 GHz, com 8 GB de memória RAM e sistema operacional Windows 8, sendo um para a ferramenta e outro para a base de dados. Para a base de dados foi utilizado um HD SSD de 500 GB. Observa-se que não houve alteração substancial no tempo ao adicionar dicionários à lista de execução, sendo a análise 3 a mais rápida, e um incremento de aproximadamente 6% ao utilizar-se o dicionário SentiLex, e de 4% entre esta segunda análise e a análise com a combinação dos dois dicionários. A seguir apresentam-se os quantitativos das análises.

Tabela 9 – Quantitativos de “Evaluations” para as análises realizadas

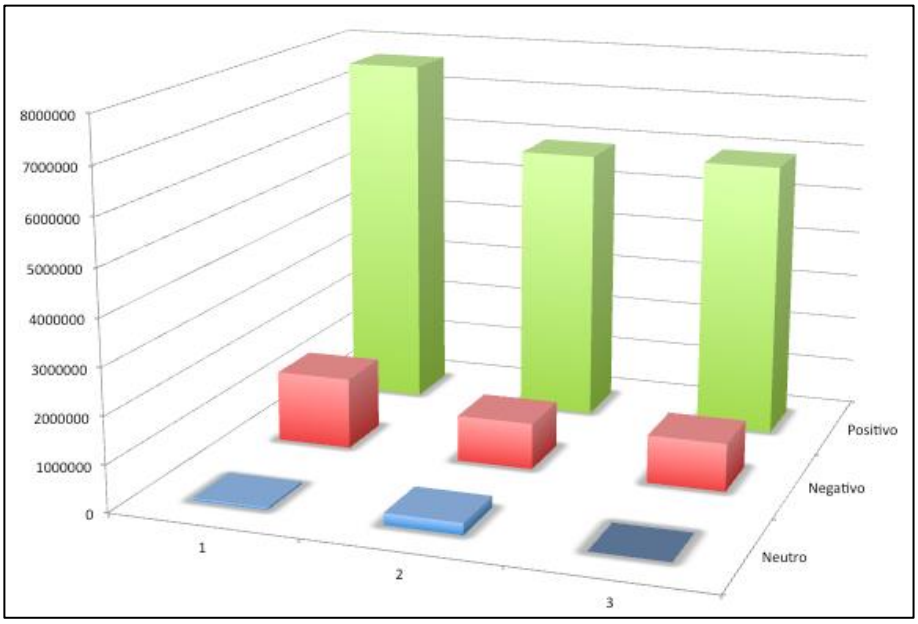
Análise	**Quantitativo**
Análise 1 – SentiLex + LIWC	9.245.488
Análise 2 – SentiLex	7.141.816
Análise 3 – LIWC	6.961.749

Fonte: Próprio autor

Conforme observa-se na Tabela 9, a combinação dos dicionários representou um elevado ganho em termos do quantitativo de palavras opinativas e aspectos

detectados. A imagem a seguir detalha as detecções de palavras opinativas efetuadas.

Figura 34 – Distribuição de polaridades por análise

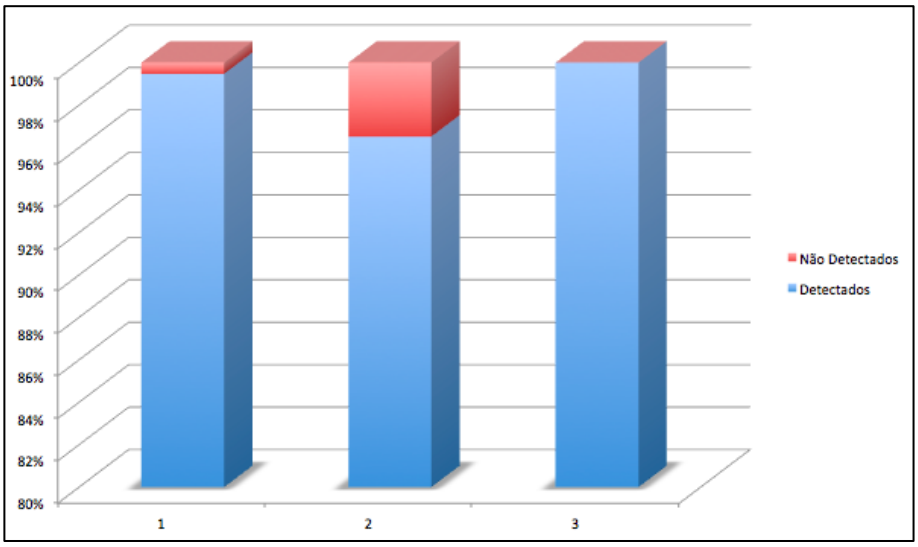


Fonte: Próprio autor

Observa-se na Figura 34 a distribuição das polaridades analisadas em cada experimento realizado. Destaca-se a grande diferença no quantitativo de palavras opinativas positivas em relação às demais, mantendo um percentual de 83% em todas as análises. Este fato comprova a maior tendência por comentários positivos em relação aos demais.

As palavras neutras aqui apresentadas referem-se a palavras que o sistema não conseguiu determinar sua polaridade, uma vez que identificar opiniões neutras de fato estava fora do escopo deste trabalho. A análise 3 não apresenta palavras neutras pois o dicionário LIWC é composto somente por palavras nas categorias positivo e negativo. A relação entre polaridades, detectadas ou não, está representada na imagem a seguir.

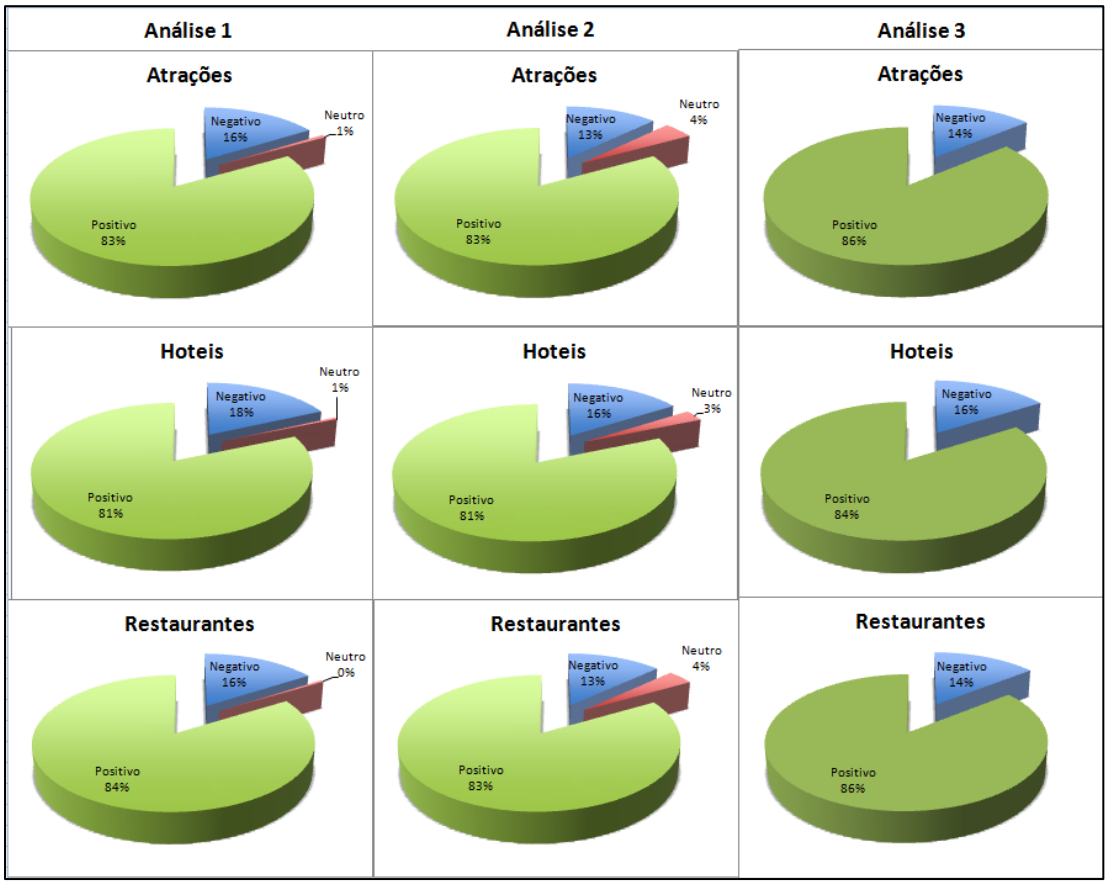
Figura 35 – Proporção de palavras opinativas com polaridades detectadas



Fonte: Próprio autor

A Figura 35 apresenta a proporção entre as palavras opinativas identificadas no texto com e sem polaridade detectada. Como citado anteriormente, a análise 3 não apresenta situações de não detecção de polaridades pela composição do dicionário. Observa-se que a junção do dicionário LIWC na análise 1 contribuiu significativamente para a determinação da polaridade das palavras opinativas. A imagem a seguir demonstra detalhadamente a distribuição das polaridades por tipo de objeto.

Figura 36 – Proporção de polaridades por tipo de objeto



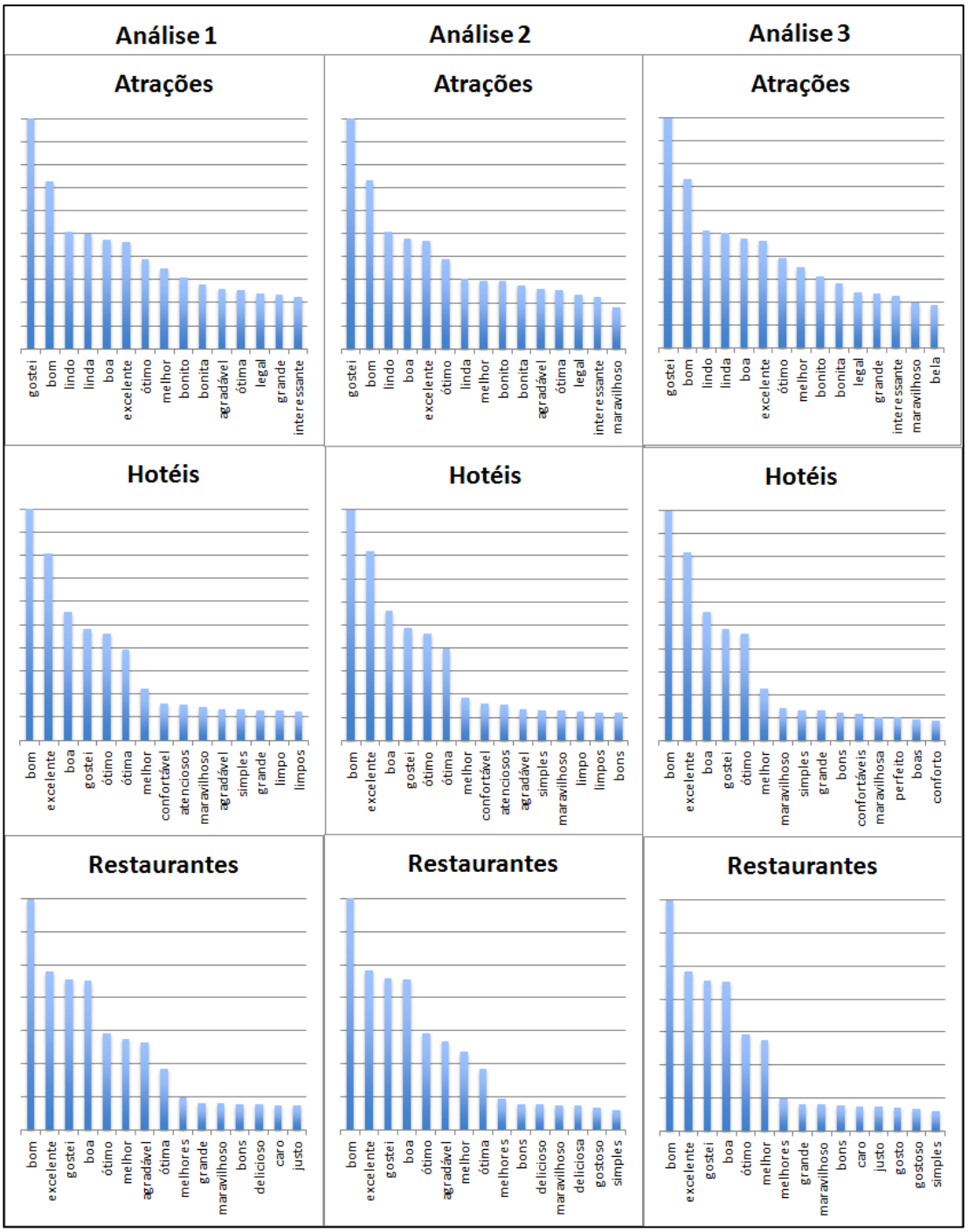
Fonte: Próprio autor

Observa-se na Figura 36 a distribuição das polaridades de palavras opinativas entre os tipos de objetos, sendo que cada coluna da imagem representa uma análise. Como pode ser observado, os percentuais não apresentam grande variação entre os tipos de objetos em uma mesma análise, e em todas as três análises a categoria hotéis obteve um percentual de comentários positivos moderadamente inferior às demais categorias.

Outro ponto a ser observado é o percentual de comentários negativos e neutros. Observa-se que a proporção de redução na categoria neutra da análise 2 pode ser constatada como acréscimo na categoria negativa da análise 1. Isto pode ser um indicativo de que a maioria das polaridades determinada pelo dicionário LIWC na análise 1 foram identificadas como negativas. Entretanto são necessários mais estudos para comprovação.

A seguir são apresentadas as palavras opinativas mais frequentes em cada tipo de objeto.

Figura 37 – Palavras opinativas mais frequentes por tipo de objeto



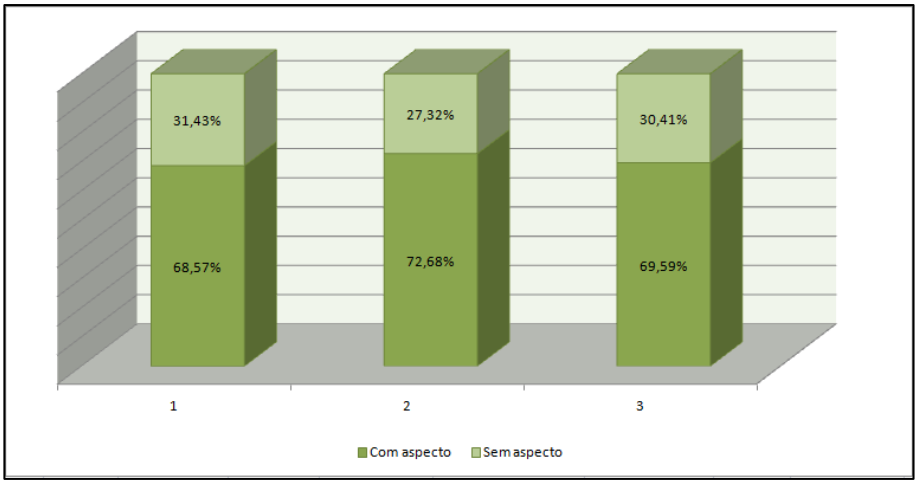
Fonte: Próprio autor

Observa-se na Figura 37 a distribuição de frequência das 15 palavras opinativas mais detectadas nos comentários de cada tipo de objeto. Cada coluna da imagem representa uma análise distinta. Pode-se verificar que, dentre as palavras mais frequentes das três análises, muitas são iguais, embora algumas variações ocorram em função das composições dos dicionários. Estas variações são mais

acentuadas, embora não muito, entre os diferentes tipos de objetos, o que se justifica justamente devido à mudança de contexto dos comentários.

A seguir é apresentada a proporção entre análises com e sem aspectos detectados.

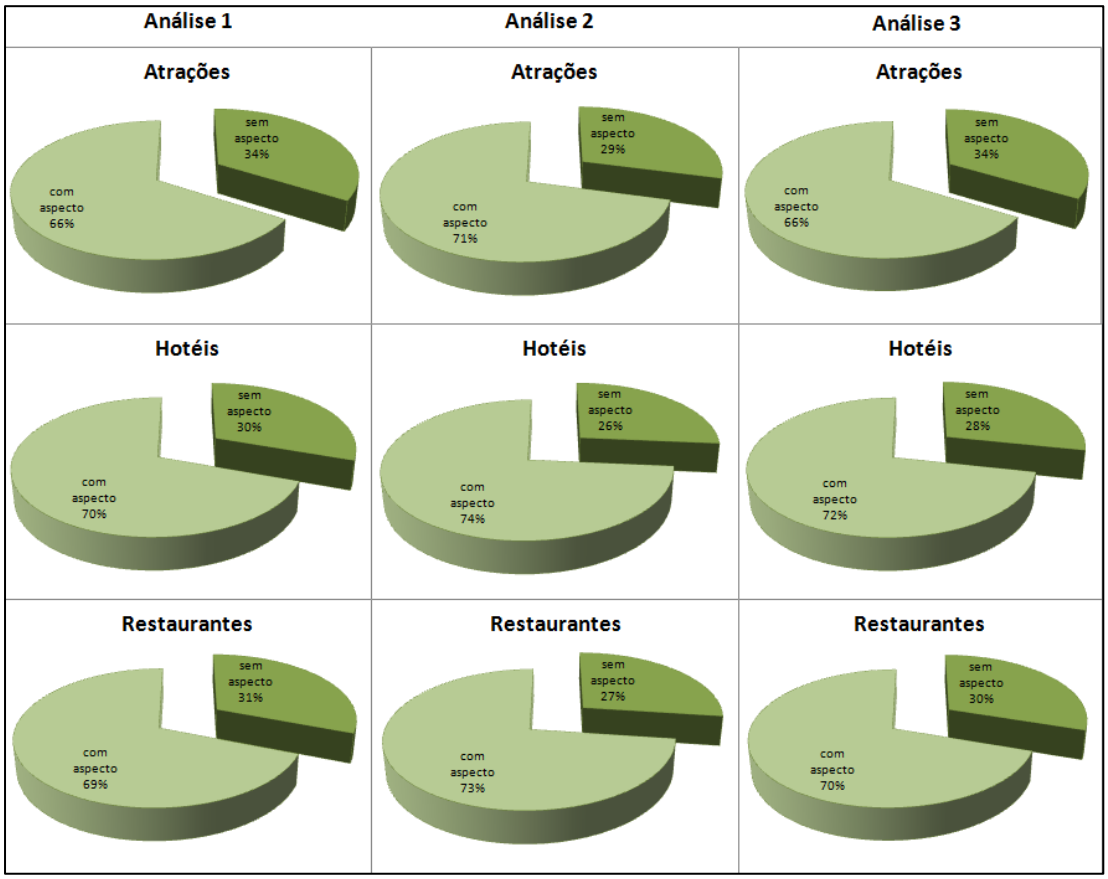
Figura 38 – Proporção de análises com aspectos



Fonte: Próprio autor

Observa-se na Figura 38 o percentual de cobertura da descoberta de aspectos nas três análises efetuadas. Os percentuais de análise com aspectos revelam uma boa taxa de descoberta de aspectos, pois o percentual sem aspectos não representa necessariamente uma falha. Muitas palavras opinativas referem-se ao objeto avaliado e não a um aspecto específico, e outras se referem a aspectos implícitos, não possuindo, de fato, o aspecto explicitamente referenciado na sentença. A seguir estes dados serão detalhados por tipo de objeto.

Figura 39 – Proporção de análises com aspectos por tipo de objeto

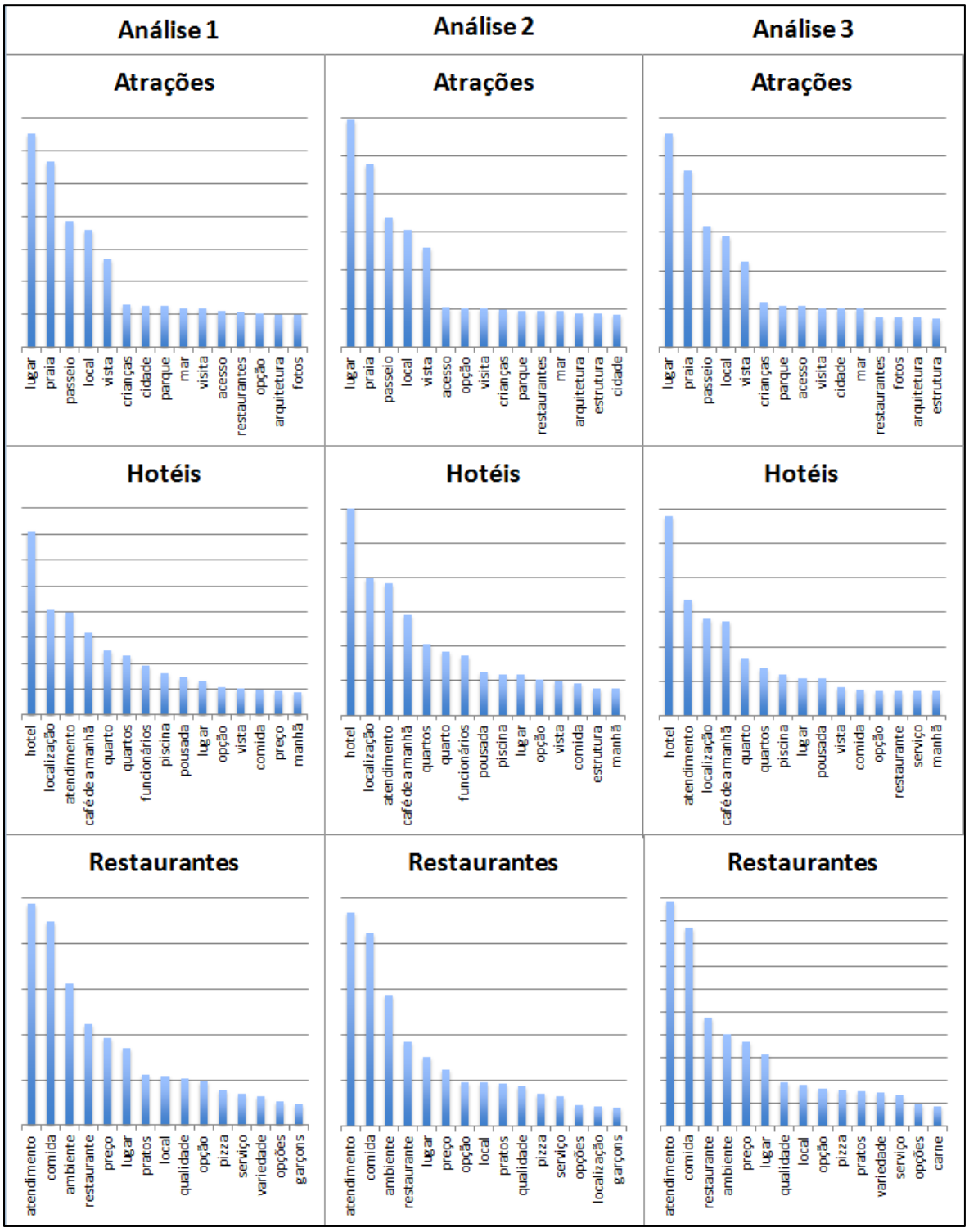


Fonte: Próprio autor

A Figura 39 apresenta o detalhamento dos percentuais de análises com e sem aspectos, por tipo de objeto e análise efetuada. Observa-se que os menores índices de detecção de aspectos ocorrem nas avaliações da análise 1. São necessários mais estudos para determinar se o método proposto possui alguma falha ou se o percentual está relacionado às novas palavras opinativas detectadas pela junção dos dicionários, e não ao método em si.

A seguir são apresentados os aspectos mais frequentemente identificados para cada tipo de objeto.

Figura 40 – Aspectos mais frequentes por tipo de objeto



Fonte: Próprio autor

A Figura 40 descreve a distribuição das frequências dos aspectos mais identificados pela ferramenta. Destaca-se a grande diferença entre as palavras de cada tipo de objeto em uma mesma análise. Entretanto as alterações não são tão acentuadas entre as diferentes análises.

5 CONSIDERAÇÕES FINAIS

O crescimento e popularização da internet têm propiciado um aumento constante no volume de dados *online*. A maioria destes dados está na forma de textos não estruturados, escritos em linguagem natural, sendo grande parte constituída de textos subjetivos, expressando opiniões sobre os mais diversos produtos e serviços.

Para transformar estes dados em informação, são utilizadas técnicas de Extração da Informação, Processamento de Linguagem Natural e Análise de Sentimentos, para coletar, processar e sintetizar os sentimentos coletivos disponíveis. Dentre as diversas técnicas possíveis para a AS, este trabalho tratou da análise ao nível de aspectos, valendo-se da abordagem léxica. Foram processados textos em português com avaliações sobre destinos turísticos brasileiros.

Para a tarefa de Extração da Informação, utilizou-se o “import.IO”, uma ferramenta de grande auxílio nesta tarefa. Como trabalhos futuros relacionados a esta etapa, entretanto, sugere-se o estudo e implementação de sistemas e APIs para *Web Crawling* e *Web Scraping*, de modo a permitir o desenvolvimento de soluções completas e integradas no contexto da Análise de Sentimentos e da Extração da Informação de um modo geral.

Para a tarefa de Processamento de Linguagem Natural utilizou-se a API disponível no corretor ortográfico CogrOO. Ela demonstrou-se capaz de efetuar a análise morfossintática nos comentários. Entretanto, não existe disponível funcionalidade para a análise da gramática de dependência.

A elaboração de uma árvore de dependência para os termos das sentenças é condição necessária para técnicas como a *Double Propagation*, para a elaboração e expansão dos dicionários de sentimentos. Até onde se tem conhecimento não existe disponível uma API para efetuar tal análise em textos escritos em português. Deste modo, recomenda-se para trabalhos futuros a elaboração de uma API que seja capaz de efetuar tal análise.

Outro problema encontrado refere-se ao informalismo utilizado na escrita dos comentários. Com efeito, os casos mais frequentes foram tratados utilizando-se

expressões regulares durante a etapa de pré-processamento. Entretanto, muitos casos de erros ortográficos e gramaticais não foram abrangidos por esta abordagem. Deste modo, como trabalhos futuros, sugere-se a adição de uma nova etapa de pré-processamento, responsável pela correção ortográfica e gramatical automática dos comentários. Para tal tarefa pode ser estudada a viabilidade de utilização da própria API do CogrOO.

O método demonstrou-se promissor, apresentando resultados iniciais favoráveis. São necessários, entretanto, trabalhos futuros para que avaliadores efetuem a análise manual de um conjunto amostral dos comentários, e para o desenvolvimento de um sistema que efetue a comparação dos resultados. Através deste trabalho poderiam ser obtidos dados em termos das medidas comumente utilizadas na área, como *F-Measure, Precision e Recall*.

Observações iniciais demonstram que o método consegue analisar bem frases como:

- “eles nunca pecam pelo atendimento, sucos naturais, ambiente e preço.” Gerando como saída <nunca_pecam; +1; [atendimento | sucos naturais | ambiente | preço];
- “excelente atendimento, vista excepcional” gerando como saídas <excelente; +1; [atendimento]> e <excepcional; +1; [vista]>

Entretanto foram observadas falhas em sentenças como:

- “O lugar é bonito, a comida é boa e o preço é alto.”. Como a palavra opinativa “alto” não possui polaridade definida, o método busca o contexto semântico da frase. Entretanto, a frase apresenta uma contradição de opiniões positivas e negativas sem o emprego de uma palavra adversativa. De todo modo, o sistema gera a saída <alto; +1; [preço]>.
- "o bacana dos pratos é que são criados por pessoas famosas." A falta de um processo anterior a AS para filtrar frases objetivas, faz com que sentenças como esta gerem análises como <bacana; +1; [pratos]>, o que não é verdade necessariamente.
- “Possui uma área externa que é uma delícia se o clima estiver bacana.” Tal como o exemplo anterior, a falta de uma etapa de filtragem de sentenças ocasiona a análise <bacana; +1; [clima]>. Enquanto que na realidade trata-se de uma opinião presa a um condicional.

- “O ambiente é bonito e agradável.” Quando existem dois ou mais adjetivos para um único aspecto, o método gera como saída duas análises, uma com aspecto e outra sem. <bonito; +1; [ambiente]> e <agradável; +1; [ ]>. Em melhorias futuras pode ser aplicada uma lógica para detectar listas de palavras opinativas para um mesmo aspecto.
- “O problema é a espera de mais de uma hora.” A falta de um dicionário específico de domínio ocasionou análises como esta <espera; +1; [problema]>. O método desenvolvido permite a fácil adição de um dicionário de domínio, mas são necessários trabalhos futuros para desenvolvê-los.

De toda forma, os casos de sentenças analisadas erroneamente como apresentados acima não são frequentes, o que não invalida o método. Um conjunto de análises foi selecionado aleatoriamente na base e encontra-se disponível no APÊNDICE H.

Apesar de o trabalho ter sido desenvolvido para o contexto do turismo brasileiro, a ferramenta foi projetada de forma a ser independente de contexto. Deste modo, com as devidas parametrizações a ferramenta pode ser empregada para outros contextos, desde que na língua portuguesa.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ANTUNES, Sandra; MENDES, Amália. An evaluation of the role of statistical measures and frequency for MWE identification. In: Language Resources And Evaluation Conference (LREC), 9., 2014, Reykjavik (iceland). **Proceedings…** Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/518_Paper.pdf>. Acesso em: 09 jun. 2015.

ARAÚJO, Matheus; GONÇALVES, Pollyanna; BENEVENUTO, Fabrício. Métodos para Análise de Sentimentos no Twitter. In: Simpósio Brasileiro De Sistemas Multimídia E Web (WEBMEDIA), 19, 2013, Salvador. **Proceedings**... Disponível em: <http://homepages.dcc.ufmg.br/~fabricio/download/webmedia13.pdf> Acesso em: 08 jun. 2015.

BALAGE FILHO, Pedro P.; PARDO, Thiago A. S.; ALUÍSIO, Sandra M.. An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. In: Brazilian Symposium In Information And Human Language Technology, 9., 2013, Fortaleza. **Proceedings..**. p. 215 - 219.

BALDWIN, Timothy; KIM, Su Nam. Multiword Expressions. In: INDURKHYA, Nitin; DAMERAU, Fred J... **Handbook of Natural Language Processing.** 2. ed. Boca Raton, Fl: Chapman And Hall/crc, 2010. Cap. 12. p. 267-292.

BECKER, Karin; TUMITAN, Diego. Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 28, 2013, Recife. **Minicursos.** Recife: Sbbd, 2013. Disponível em: <http://www.inf.ufrgs.br/~kbecker/lib/exe/fetch.php?media=minicursosbbd_versaosubmetida.pdf>. Acesso em: 18 set. 2014.

BORTH, Damian et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 21., 2013, New York, Ny, Usa. **Proceedings.** ACM, 2013. p. 223 - 232. Disponível em: < http://www.ee.columbia.edu/ln/dvmm/vso/download/visual_sentiment_ontology_FINAL.pdf >. Acesso em 23 set. 2014.

BOOS, Rodrigo Augusto Scheller; PRESTES, Kassius Vargas; VILLAVICENCIO, Aline. Identification of Multiword Expressions in the brWaC. In: Language Resources And Evaluation Conference (LREC), 9., 2014, Reykjavik (iceland). **Proceedings…** Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/518_Paper.pdf>. Acesso em: 09 jun. 2015.

BRASIL, Ministério do Turismo. **Plano Nacional do Turismo 2013-2016**: O Turismo fazendo muito mais pelo Brasil. Brasília: Ministério do Turismo, 2013.

BRASIL, Ministério do Turismo; SEBRAE; VARGAS, Fundação Getúlio. **Índice de competitividade do turismo nacional** : destinos indutores do desenvolvimento turístico regional : relatório Brasil 2013. – Brasília, DF : Ministério do Turismo, 2013. 92 p.

CALZOLARI, Nicoletta et al. Towards best practice for multiword expressions in computational lexicons. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 3., 2002. **Proceedings...** p. 1934 - 1940.

CHURCH, Kenneth Ward; HANKS, Patrick. Word association norms, mutual information, and lexicography. **Computational Linguistics.** Cambridge, MA, EUA, p. 22-29. mar. 1990. Disponível em: <http://dl.acm.org/ft_gateway.cfm?id=89095&ftid=248552&dwn=1&CFID=518013958&CFTOKEN=59567655>. Acesso em: 09 jun. 2015.

COHEN, Andrew R.; VITÁNYI, Paul M.b.. **Normalized Google Distance of Multisets with Applications.** Corr Abs/1308.3177, 2013.

CORTES, Corinna; VAPNIK, Vladimir. Support-Vector Networks. **Machine Learning.** Hingham, MA, USA, p. 273-297. set. 1995.

CRUYS, Tim van de. Two multivariate generalizations of pointwise mutual information. In: WORKSHOP ON DISTRIBUTIONAL SEMANTICS AND COMPOSITIONALITY DISCO, 11, 2011, Stroudsburg, Pa, Eua. **Proceedings...** Stroudsburg, Pa, Eua: Association For Computational Linguistics, 2011. p. 16 - 20. Disponível em: <http://dl.acm.org/ft_gateway.cfm?id=2043124&ftid=1040296&dwn=1&CFID=518013958&CFTOKEN=59567655>. Acesso em: 09 jun. 2015.

DOMINGUES, Miriam Lúcia Campos Serra. **Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o português do Brasil.** 2011. 140 f. Tese (Doutorado) - Curso de Programa de Pós-graduação em Engenharia Elétrica, Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2011. Disponível em: <http://www.miriam.ufpa.br/arquivos/TesedeDoutorado_Miriam_PPGEE_10_2011.pdf>. Acesso em: 01 nov. 2014.

FELDMAN, Ronen. Techniques and Applications for Sentiment Analysis. **Communications Of The Acm,** v. 56, n. 4, p.82-89, abr. 2013. Disponível em: <http://dl.acm.org/citation.cfm?id=2436274>. Acesso em: 18 set. 2014.

GÜNGÖR, Tunga. Part-of-Speech Tagging. In: INDURKHYA, Nitin; DAMERAU, Fred J.. **Handbook of Natural Language Processing.** 2. ed. Boca Raton, Fl: Chapman And Hall/crc, 2010. Cap. 10. p. 205-235.

HAHN, Sangyun; OSTENDORF, Mari. A Comparison of Discriminative EM-Based Semi-Supervised Learning algorithms on Agreement/Disagreement classification. In: NIPS - SPEECH AND LANGUAGE: LEARNING-BASED METHODS AND SYSTEMS, 22. 2008, Whistler, British Columbia, Canada. **WORKSHOP.**

HIPPISLEY, Andrew. Lexical Analysis. In: INDURKHYA, Nitin; DAMERAU, Fred J.. **Handbook of Natural Language Processing.** 2. ed. Boca Raton, Fl: Chapman And Hall/crc, 2010. Cap. 3. p. 31-58.

JETBRAINS. **IntelliJ.** 2000. Disponível em: <https://www.jetbrains.com/idea/>. Acesso em: 13 mar. 2015.

KOHLSCHÜTTER, Christian; FANKHAUSER, Peter; NEJDL, Wolfgang. Boilerplate Detection using Shallow Text Features. In: THE THIRD ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, 3., 2010, New York City, NY. **Proceedings..**. Disponível em: <http://www.l3s.de/~kohlschuetter/publications/wsdm187-kohlschuetter.pdf>. Acesso em: 20 mar. 2015.

LIDDY, Elizabeth D.. Enhanced Text Retrieval Using Natural Language Processing. **Bulletin Of The American Society For Information Science**, Maryland, v. 24, n. 4, p.14-16, 1998.

LIMA, Edirlei Soares de. **Inteligência Artificial:** Rio de Janeiro: PUC, 2014. Disponível em: <http://slideplayer.com.br/slide/1868847/>. Acesso em: 16 nov. 2014.

LIU, Bing. Sentiment analysis and subjectivity. In: INDURKHYA, Nitin; DAMERAU, Fred J.. **Handbook of Natural Language Processing.** 2. ed. Boca Raton, Fl: Chapman And Hall/crc, 2010. Cap. 26. p. 627-666.

LIU, Bing. **Sentiment analysis and opinion mining.** Chicago: Morgan & Claypool Publishers, 2012. 167 p. Disponível em: <http://www.morganclaypool.com/doi/pdf/10.2200/S00416ED1V01Y201204HLT016> . Acesso em: 22 out. 2014.

LJUNGLÖF, Peter; WIRÉN, Mats. Syntactic Parsing. In: INDURKHYA, Nitin; DAMERAU, Fred J.. **Handbook of Natural Language Processing.** 2. ed. Boca Raton, Fl: Chapman And Hall/crc, 2010. Cap. 4. p. 59-81.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de. Uma Introdução às Support Vector Machines. **Rita - Revista de Informática Teórica Aplicada,** Porto Alegre, v. 14, p.43-67, 2007. Semestral.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **An Introduction to Information Retrieval.** Cambridge, England: Cambridge University Press, 2009. 544 p.

MICROSOFT. **SQL Server 2008.** 2015. Disponível em: <http://www.microsoft.com>. Acesso em: 13 mar. 2015.

MITCHELL, T. **Machine Learning**. McGraw Hill. 1997. New York, USA. 414 p.

NADKARNI, Prakash M; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. Natural language processing: an introduction. **Journal Of The American Medical Informatics Association,** [s.l], v. 18, n. 5, p.544-551, set. 2011.

NEWMAN, Mark. E. J. Power laws, Pareto distributions and Zipf's law. **Contemporary Physics.** p. 323-351. 2005.

NIGAM, Kamal et al. Text Classification from Labeled and Unlabeled Documents using EM. **Machine Learning - Special Issue On Information Retrieval**. Hingham, Ma, p. 103-134. maio 2000.

NORTHWOOD, Chris. COM6170: **Machine Learning Foundations**. 2009. Disponível em: <http://www.pling.org.uk/cs/com6170.html>. Acesso em: 16 nov. 2014.

PENTAHO. **Kettle - Spoon.** 2004. Disponível em: <http://www.pentaho.com>. Acesso em: 13 mar. 2015.

POZO, Aurora Trinidad Ramirez. **Classificação**. Curitiba: UFPR, 2006. 34 slides, color. Disponível em: < http://www.inf.ufpr.br/aurora/disciplinas/topicosia/Classifica%E7%E3o.ppt>. Acesso em: 09 dez. 2014.

QIU, Guang et al. Expanding Domain Sentiment Lexicon through Double Propagation. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 21. 2009, Pasadena, California. **Proceedings.**

RUSSELL, Matthew A.. **Mining the Social Web.** 2. ed. Sebastopol: O'reilly Media, Inc., 2013. 421 p.

SAG, Ivan A. et al. Multiword Expressions: A Pain in the Neck for NLP. In: CICLING INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 3., 2002, Mexico-city. **Proceeding.** London, Uk: Springer-verlag, 2002. p. 1 - 15.

SILVA, Mário J.; CARVALHO, Paula; SARMENTO, Luís. Building a Sentiment Lexicon for Social Judgement Mining. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE (PROPOR), 2012, Coimbra. **Lecture Notes in Computer Science (LNCS)**.

SILVA, Nelson Rocha; LIMA, Diego; BARROS, Flávia. SAPair: Um Processo de Análise de Sentimento no Nível de Característica. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEM, 2012, Curitiba. **Anais.** Disponível em: <http://www.ppgia.pucpr.br/~enia/anais/wti/artigos/105283.pdf>. Acesso em: 18 set. 2014.

SILVA, William Daniel Colen de Moura. **Aprimorando o Corretor Gramatical CoGrOO.** 2013. 166 f. Dissertação (Mestrado) - Curso de Mestrado em Ciência da Computação, Departamento de Instituto de Matemática e Estatística, USP, São Paulo, 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45134/tde-02052013-135414/publico/WilliamColen_Dissertation.pdf>. Acesso em: 18 set. 2014.

TORRES, Carlos Eduardo Atencio. **Uso de informação linguística e análise de conceitos formais no aprendizado de ontologias.** 2012. 67 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Departamento de Instituto de Matemática e Estatística, USP, São Paulo, 2012. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45134/tde-11022013-152711/publico/teseCarlosTorres.pdf>. Acesso em: 18 set. 2014.

VISL. **PALAVRAS.** 1996. Disponível em: <http://beta.visl.sdu.dk/visl/pt/parsing/automatic/>. Acesso em: 17 nov. 2014.

VITÁNYI, Paul M.b.; CILIBRASI, Rudi L.. Normalized Web Distance and Word Similarity. In: INDURKHYA, Nitin; DAMERAU, Fred J.. **Handbook of Natural Language Processing.** 2. ed. Boca Raton, Fl: Chapman And Hall/crc, 2010. Cap. 13. p. 293-314.

WANG, Dong; LIU, Yang. A cross-corpus study of unsupervised subjectivity identification based on calibrated EM. In: Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2., 2011, Portland. **Proceedings.** Stroudsburg, Pa: Association For Computational Linguistics, 2011. p. 161 - 167.

WHITE, David; PAINTER, Matthew; FOGG, Andrew. 2015. **Import IO.** Disponível em: <http://import.io>. Acesso em: 13 mar. 2015.

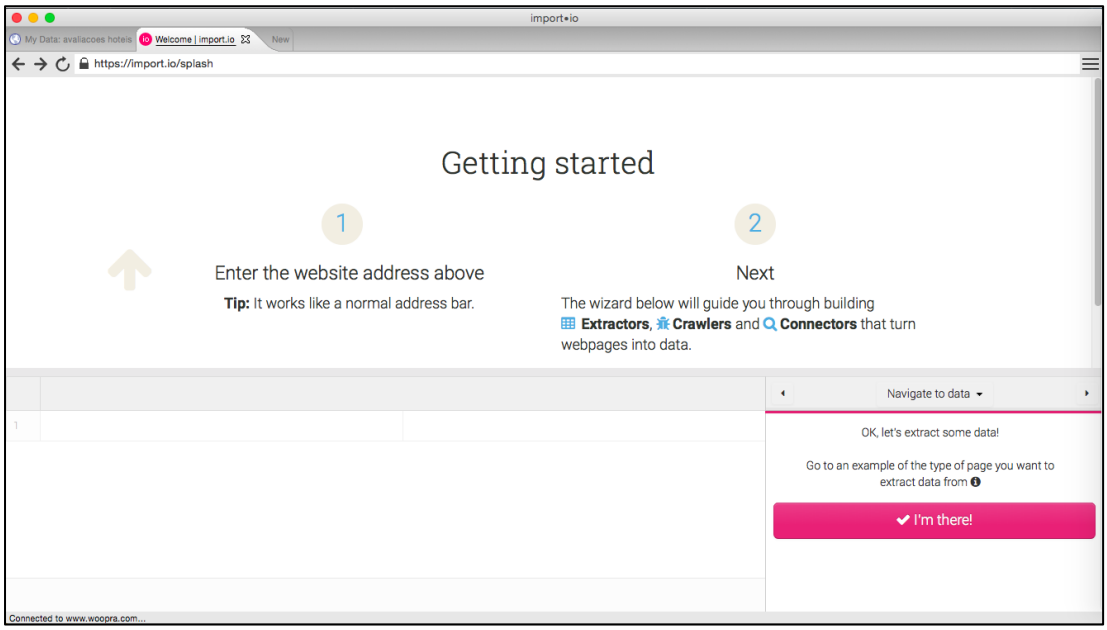
XIAO, Richard. Corpus Creation. In: INDURKHYA, Nitin; DAMERAU, Fred J.. **Handbook of Natural Language Processing.** 2. ed. Boca Raton, Fl: Chapman And Hall/crc, 2010. Cap. 7. p. 147-165.

7 APÊNDICES

APÊNDICE A – CRIANDO UM NOVO CRAWLER

Para a extração das informações existentes no site “Trip Advisor” utilizou-se o *web crawler* “import.io”. A opção “*new crawler*” no menu principal da ferramenta permite configurar uma nova tarefa de extração. Através desta é exibida a página inicial do assistente, apresentada a seguir:

Figura 41 – Tela Inicial Import.io



Fonte: Próprio autor

Observa-se na Figura 41 o assistente de criação de um novo *crawler*. Para acessar a página a ser utilizada no processo de treinamento da ferramenta é utilizada a barra de endereços na parte superior, como um *browser* comum. Ao acessá-la, a opção “*I’m there!*” prossegue com o assistente, que conduz o usuário pelos passos necessários para a criação.

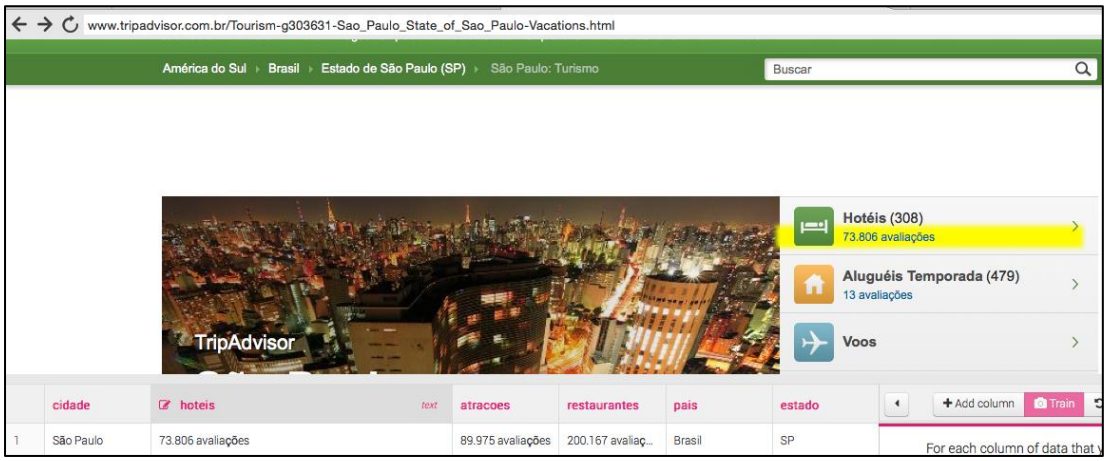
Inicialmente são ajustadas as configurações de otimização das páginas, de modo a verificar se será necessário o carregamento de estilos e scripts, ou somente o conteúdo HTML. Quanto menos scripts e estilos forem necessários, menos dispendioso é o processo de extração, uma vez que menos informação precisa ser trafegada e processada.

Em seguida, o usuário pode escolher se a página desejada apresenta um ou muitos registro por página. Caso seja escolhida a opção de mais de um registro por página, o assistente solicita que o usuário selecione cada uma das linhas

apresentadas e clique no botão “*Train rows*”. A opção “*I’m got all x rows*” indica que todas as linhas já foram treinadas.

Após selecionarem-se todas as linhas, ou caso seja escolhida a opção de um registro por página, é apresentada a etapa de localização dos dados, demonstrada na figura a seguir:

Figura 42 – Seleção de dados a serem extraídos

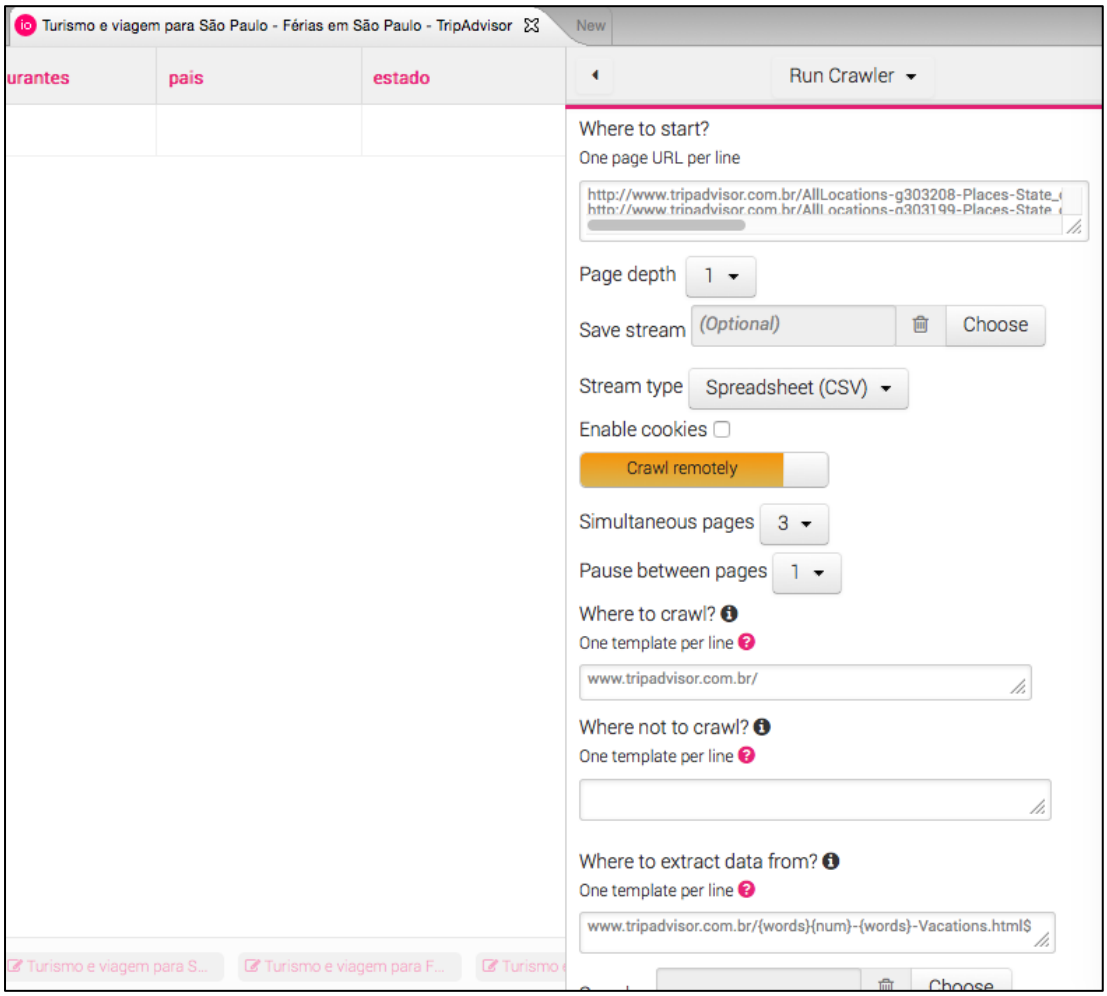


Fonte: Próprio autor

A Figura 42 apresenta, como exemplo da etapa de seleção dos dados a serem extraídos, o *crawler* criado para a extração da quantidade de comentários de cada cidade. Através da opção “*Add column*” são adicionadas colunas e seus tipos, podendo ser textos, números, imagens, marcações HTML, dentre outros. O usuário seleciona para cada coluna o local da página que contém a informação para o seu preenchimento, conforme pode ser observado, por exemplo, em amarelo na Figura 42 para a coluna hotéis. Com a informação selecionada, utiliza-se a opção “*Train*” para efetuar o treinamento da coluna. Após treinar todas as colunas, a opção “*I’ve got what I need!*” conclui o processo de treinamento da página.

Precisam ser treinadas pelo menos cinco páginas para que a ferramenta consiga inferir o padrão de informações a serem extraídas, sendo que em casos mais complexos podem ser necessárias mais páginas. O processo de treinamento pode ser terminado quando for observado que a ferramenta está conseguindo identificar todas as informações desejadas de forma automática para as páginas que forem adicionadas ao treinamento. Ao terminar o processo, o assistente solicita o envio dos dados para os servidores da empresa e é apresentada a tela de início do processo de extração, apresentada na figura a seguir:

Figura 43 – Parametrização do crawler



Fonte: Próprio autor

A Figura 43 demonstra alguns dos parâmetros da extração avançada. Nela o campo “*Where to start*” permite configurar os endereços iniciais, nos quais o sistema começará o processo de extração. “*Page depth*” refere-se a profundidade, ou quantidade de links que a ferramenta poderá seguir para procurar páginas que respeitem o modelo treinado. A ferramenta permite no máximo cinquenta níveis de profundidade. Portanto, devem ser adotadas estratégias para iniciar a extração a partir de páginas mais centrais ou que tenham um elevado grau de links de saída para as páginas a serem extraídas.

“*Save stream*” permite que o *stream* de extração dos dados seja salvo em um arquivo csv local durante o processo de extração. Esta opção pode demonstrar-se particularmente interessante nos casos de grandes volumes de dados extraídos, nos quais a ferramenta apresenta um problema no momento de salvar e não consegue enviar corretamente os registros para o servidor. Também é importante para os

casos de interrupção inesperada do processo, para que possam ser aproveitados resultados parciais da extração.

"*Crawl remotely*" ou "*Crawl localy*" indica se o processo de extração deve ser executado nos servidores da ferramenta ou no computador local onde o software foi instalado. Se a velocidade da internet disponível para o processo for boa, a extração local pode ser mais rápida, devido à inexistência de concorrência. Entretanto, o que se observou foi que colunas do tipo HTML só tem o conteúdo extraído caso esta opção esteja marcada como extração remota.

"*Simultaneos pages*" e "*Pause between pages*" controlam a velocidade da extração. Quanto maior o número de páginas simultâneas e menor o tempo de espera entre as páginas, menor o tempo gasto para recuperar as informações. Entretanto, parâmetros muito exigentes podem sobrecarregar o site de origem dos dados, causando uma indisponibilidade do serviço. Portanto estes parâmetros devem ser utilizados com cautela.

"*Where to crawl*" refere-se às URLs das páginas que serão visitadas pela ferramenta. "*Where not to crawl*" por outro lado refere-se a páginas que não devem ser visitadas. "*Where to extract data from*" consiste nas páginas onde a ferramenta deve extrair dados. Estes três campos devem ser configurados de modo a manter o extrator "no caminho" correto. Devido aos diversos links existentes nas páginas, pode ocorrer do *crawler* ser conduzido para sites e páginas indesejadas, o que pode incorrer em extração de dados indesejados, ou em um processo muito mais lento do que o necessário. Estes campos permitem a especificação de endereços específicos ou de *templates* através de palavras chave como "{any}, {num}, {words}, {not}, etc" para especificar padrões de endereço.

Terminado o processo de configuração dos parâmetros, a ferramenta está pronta para extrair os dados desejados, bastando para tal escolher a opção "*Go*".

APÊNDICE B – DESTINOS BRASILEIROS IMPORTADOS

Nome	Estado	Quantidade
Rio de Janeiro	RJ	348534
São Paulo	SP	320137
Curitiba	PR	99713
Fortaleza	CE	98510
Brasília	DF	92464
Natal	RN	91735
Gramado	RS	90158
Salvador	BA	86333
Foz do Iguaçu	PR	79009
Belo Horizonte	MG	74110
Maceió	AL	64902
Búzios	RJ	63957
Porto Alegre	RS	62840
Recife	PE	59976
Campos do Jordão	SP	42064
João Pessoa	PB	41738
Porto de Galinhas	PE	37804
Manaus	AM	35138
Porto Seguro	BA	29732
Aracaju	SE	29284
Jericoacoara	CE	28918
Campinas	SP	28475
Belém	PA	28032
Canela	RS	24971
Goiânia	GO	24757
Vitória	ES	22087
Fernando de Noronha	PE	21626
Praia da Pipa	RN	21002
Balneário Camboriú	SC	20896
Bonito	MS	19839
São Luís	MA	19287
Arraial d'Ajuda	BA	18346
Ilhabela	SP	17431
Santos	SP	17226
Ribeirão Preto	SP	17019
Ubatuba	SP	14923
São Sebastião	SP	14398
Praia do Forte	BA	13755
Bento Gonçalves	RS	13641
Caldas Novas	GO	13024
Tiradentes	MG	13003
Monte Verde	MG	12618

Cabo Frio	RJ	11878
Londrina	PR	11349
Guarujá	SP	11220
Bombinhas	SC	11025
Poços de Caldas	MG	10741
São José dos Campos	SP	10431
Itacaré	BA	10231
Joinville	SC	9715
Aquiraz	CE	9545
Blumenau	SC	9528
Pirenópolis	GO	9186
Guarulhos	SP	9172
Arraial do Cabo	RJ	8928
Trancoso	BA	8498
Uberlândia	MG	8466
Vila Velha	ES	8189
Canoa Quebrada	CE	8129
Teresina	PI	8082
Barreirinhas	MA	7778
Teresópolis	RJ	7777
Juiz de Fora	MG	7429
Penedo	RJ	7366
Ilhéus	BA	7112
Penha	SC	7014
Sorocaba	SP	7009
Maringá	PR	6823
Caxias Do Sul	RS	6726
São José do Rio Preto	SP	6421
Santo André	SP	6404
Jundiaí	SP	6297
Olinda	PE	6051
Tibau do Sul	RN	5632
Piracicaba	SP	5530
Rio Quente	GO	5505
Olímpia	SP	5432
Lençóis	BA	5178
São Bernardo do Campo	SP	5006
Barueri	SP	4925
Brumadinho	MG	4730
São Lourenço	MG	4611
Porto Velho	RO	4562
Caraguatatuba	SP	4417
Atibaia	SP	4361
Rio Branco	AC	4183
Palmas	TO	4154

Águas de Lindóia	SP	4153
Macapá	AP	4057
Visconde de Mauá	RJ	4003
Campina Grande	PB	3999
Mossoró	RN	3909
Nova Petrópolis	RS	3879
Serra Negra	SP	3807
Mata de São João	BA	3762
Itu	SP	3682
Taubaté	SP	3640
Praia Imbassaí	BA	3639
Praia dos Carneiros	PE	3630
Vinhedo	SP	3625

APÊNDICE C - EXPRESSÕES REGULARES PROCESSO DE NORMALIZAÇÃO

Expressão	Substituição
(?i)(^|[(\[{“']](https?://|ww)[0-9a-z.-]+\.[a-z.]{2,6}[/\w.-]*/?\??[^]*([.!?,;\)\]\}]*|$)	$1$3
(?i)(^|[(\[{“'])[A-Za-z0-9._-]+[A-Za-z0-9]@[A-Za-z0-9._-]+[.][A-Za-z]{2,4}([.!?,;\)\]\}]*|$)	$1$2
remove emoticons substituídos duplicados	
(^|[(\[{“'])+((\(feliz\))|(\(tristeza\))|(\(neutro\)))(*\2)+	$1$2
remove datas	
(^|[(\[{“'])\d{1,2}([/.-])\d{1,2}\2\d{2,4}([- .!?,;\)\]\}”'_:]|$)	$1data$3
remove horas 10h15min	
(?i)(^|[(\[{“'])[0-9]{1,2}:?h[0-9]{1,2}(m|(min))?([- .!?,;\)\]\}”'_:]|$)	$1hora$4
remove valores US$345,99	
(^|[(\[{“'])[a-zA-Z]{0,2}\$ *\d+([.,]\d*)*([- .!?,;\)\]\}”'_:]|$)	$1valor$3
remove espaço antes de pontuações	
+([,.;!?ºª:])	$1
remove pontuações repetidas	
([^.!?,;:])([.!?,;:])([.!?;,:]+)	$1$2
remover aspas erradas	
[´`]	
[“”]	"
adiciona espaço depois e antes de parêntesis chaves e colchetes	
([^])([(\[{])	$1 $2
([\[{])([^])	$1 $2
([)\]}])([^ .!?,;\)\]\}”'])	$1 $2
([^])([\]}])	$1 $2
adiciona espaço depois de pontuação	
([!?;:\/]+)([^])	$1 $2
([.,])([^ 0-9])	$1 $2
([^ 0-9A-ZÀ-Ü]|[A-ZÀ-Ü]{2,})([\/]+)([^])	$1 $2 $3
acerta não em NÃO Não ñ NAO Nao nao	
(^|[(\[{“'])(((NÃ|NA|Nã|Na|na)(O|o))|ñ)([- .!?,;\)\]\}”'_:]|$)	não$6
acerta não em N e n	
(^|[(\[{“'])(n|N)([^0-9])	não$2
remove letras repetidas menos rr, ss, cc, pp e as vogais	
([!-/<=>bdfghj-nqtv-zà-õ])\1+	$1
remove letras repetidas vogais (volta para primeira ocorrência)	
([aeiou])\1{2,}	$1
remove letras repetidas considerando rr, ss	

(([rscp])\2)\2+	$1
substituir risos rsrsrsrs	
(?i)(^|[(\[{“']))(([rs]{3,}|[k]{2,}|[ha]{4,}|[he]{4,}|[hua]{3,})+ *)+([- .!?,;\)\]\}”'_:]|$)	$1sorridente$4
substituir notas ruins, nota zero, etc.	
(?i)(^|[(\[{“'])nota([:])?(é|foi)? ?(([01234](,\d+)?)|zero|m[íi]nima|um|dois|tr[eê]s|quatro)([- .!?,;\)\]\}”'_:]|$)	$1ruim$7
substituir notas boas, nota dez, mil, 10, etc.	
(?i)(^|[(\[{“'])nota[:]?(é|foi)? ?((([56789]|1\.?0+)(\.\d{3}|,\d)*)|dez|mil|m[aá]xima|cem|cinco|seis|sete|oito|nove)([- .!?,;\)\]\}”'_:]|$)	$1ótimo$7
substituir recomend(ad)o.	
(?i)(^|[(\[{“'])r+-?e+-?c+-?o+-?m+-?e+-?n+-?d+-?(a+-?d+-?)?o([- .!?,;\)\]\}”'_:]|$)	$1gostei$3
corrige variações do ótimo e ótimoo.	
(?i)(^|[(\[{“'])[aeiouà-ü]*-?t+-?i+-?m+-?([oa])+-?(s)*([- .!?,;\)\]\}”'_:]|$)	$1ótim$2$3$4
corrige variações do excelente.	
(?i)(^|[(\[{“'])e+-?x+-?c+-?e+-?l+-?e+-?n+-?t+-?e+-?(s)*([- .!?,;\)\]\}”'_:]|$)	$1excelente$3
corrige variações do fantástico	
(?i)(^|[(\[{“'])f+-?a+-?n+-?t+-?[aeiouà-ü]*-?s+-?t+-?i+-?c+-?([oa])+-?(s)*([- .!?,;\)\]\}”'_:]|$)	$1fantástic$2$3$4
corrigir ma-ra-vi-lho-so	
(?i)ma-ra-vi-lho-s([oa])(s)*	maravilhos$1$2
(?i)(^|[(\[{“'])m+-?a+-?r+-?a+-?v+-?i+-?l+-?h+-?o+-?s+-?([oa])+-?(s)*([- .!?,;\)\]\}”'_:]|$)	$1maravilhos$2$3$4
corrige variações do bom	
(?i)(^|[(\[{“'])b+-?o+-?[mn]+-?([- .!?,;\)\]\}”'_:]|$)	$1bom$2
corrige variações do agradável	
(?i)(^|[(\[{“'])a+-?g+-?r+-?a+-?d+-?[aeiouà-ü]*v+-?e+-?(l|is)+-?([- .!?,;\)\]\}”'_:]|$)	$1agradáve$2$3
corrige variações do indispensável	
(?i)(^|[(\[{“'])i+-?n+-?d+-?i+-?s+-?p+-?e+-?n+-?s+-?[aeiouà-ü]*v+-?e+-?(l|is)+-?([- .!?,;\)\]\}”'_:]|$)	$1indispensáve$2$3
corrige variações do incrível	
(?i)(^|[(\[{“'])[aeiouà-ü]*-?n+-?c+-?r+-?[aeiouà-ü]*v+-?e+-?(l|is)+-?([- .!?,;\)\]\}”'_:]|$)	$1incríve$2$3
corrige variações do impecável(mente)	
(?i)(^|[(\[{“'])i+-?m+-?p+-?e+-?c+-?[aeiouà-ü]*-?v+-?e+-?(l|is)+-?(mente)?([- .!?,;\)\]\}”'_:]|$)	$1impecáve$2$4
variações do superlativo (tentativa de redução que funciona para a maioria das vezes)	

(?i)[ií]+-?(s-?)+[ií]+-?m+-?([oa])+-?(s)*([- .!?,;\)\]\}”'_:]|$)	$2$3$4
corrige variações do péssimo	
(?i)(^|[(\[{“'])p+-?[aeiouà-ü]*-?(s-?)+i+-?m+-?([oa])+-?(s)*([- .!?,;\)\]\}”'_:]|$)	$1péssim$3$4$5
corrige variações do ridículo	
(?i)(^|[(\[{“'])r+-?i+-?d+-?[aeiouà-ü]*-?c+-?u+-?l+-?([oa])+-?(s)*([- .!?,;\)\]\}”'_:]|$)	$1ridícul$2$3$4
corrige variações do horrível	
(?i)(^|[(\[{“'])h+-?o+-?(r-?)+[aeiouà-ü]*-?v+-?e+-?(l|is)+-?([- .!?,;\)\]\}”'_:]|$)	$1horríve$3$4
corrige variações do terrível(mente)	
(?i)(^|[(\[{“'])t+-?e+-?(r-?)+[aeiouà-ü]*-?v+-?e+-?(l|is)+-?(mente)?([- .!?,;\)\]\}”'_:]|$)	$1terríve$3$5
corrige variações do decepção	
(?i)(^|[(\[{“'])d+-?e+-?c+-?e+-?p+-?[cç]+-?[aeiouà-ü]*-?o+-?([- .!?,;\)\]\}”'_:]|$)	$1decepção$2
corrige variações do (des)ilusão	
(?i)(^|[(\[{“'])(d+-?e+-?s+-?)?i+-?l+-?u+-?s+-?[aeiouà-ü]*-?o+-?([- .!?,;\)\]\}”'_:]|$)	$1ilusão$3
corrige variações do chateação	
(?i)(^|[(\[{“'])c+-?h+-?a+-?t+-?e+-?a+-?[cç]+-?[aeiouà-ü]*-?o+-?([- .!?,;\)\]\}”'_:]|$)	$1chateação$2
corrige são	
(?i)(^|[(\[{“'])s+-?[aã]+-?o+-?([- .!?,;\)\]\}”'_:]|$)	$1são$2
adicionar espaço entre hashtags	
(?i)([0-9a-zà-õ])(#[^])	$1 $2
remover tralhas (e asteriscos) das hashtags	
(?i)(^|[(\[{“'])[#*]([0-9a-zà-õ])	$1$2
corrigir eh	
(?i)eh([- .!?,;\)\]\}”'_:]|$)	é$1
remover prefixo super (cogroo concatena no token)	
(?i)(^|[(\[{“'])super	$1
adiciona espaço em palavras começadas com não* **	
(?i)(^|[(\[{“'])não([^ .!?,;\)\]\}”'-_:])	$1não $2
trata a expressão vale a pena	
(?i)(^|[(\[{“'])vale([uw]|r[{aá])?(+(muito)|(realmente))? +[àa] +pena([- .!?,;\)\]\}”'_:]|$)	$1gostei$6
trata abreviações	
(?i)(^|[(\[{“'])vo?ce?([- .!?,;\)\]\}”'_:]|$)	$1você$2
(?i)(^|[(\[{“'])q([- .!?,;\)\]\}”'_:]|$)	$1que$2
(?i)(^|[(\[{“'])pq([- .!?,;\)\]\}”'_:]|$)	$1por que$2

trata variações de substantivos	
(?i)(^|[(\[{“'])c+-?o+-?m+-?i+-?d+-?a+-?([- .!?,;\)\]\}”'_:]|$)	$1comida$2
(?i)(^|[(\[{“'])a+-?t+-?e+-?n+-?d+-?i+-?m+-?e+-?n+-?t+-?o+-?([- .!?,;\)\]\}”'_:]|$)	$1atendimento$2
(?i)(^|[(\[{“'])p+-?i+-?(z-?)+a+-?([- .!?,;\)\]\}”'_:]|$)	$1pizza$3
(?i)(^|[(\[{“'])garçon([- .!?,;\)\]\}”'_:]|$)	$1garçom$2
(?i)(^|[(\[{“'])c+-?h+-?o+-?p+-?e*([- .!?,;\)\]\}”'_:]|$)	$1chope$2

Emoticons positivos

(^|[(\[{])(:\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:)
(^|[(\[{])(:\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:]
(^|[(\[{])(:})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:}
(^|[(\[{])(:o\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:o)
(^|[(\[{])(:o\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:o]
(^|[(\[{])(:o})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:o}
(^|[(\[{])(:-\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-]
(^|[(\[{])(:-\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-)
(^|[(\[{])(:-})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-}
(^|[(\[{])(=\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=)
(^|[(\[{])(=\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=]
(^|[(\[{])(=})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=}
(^|[(\[{])(=\^\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^]
(^|[(\[{])(=\^\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^)
(^|[(\[{])(=\^})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^}
(^|[(\[{])(:B)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:B
(^|[(\[{])(:-D)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-D
(^|[(\[{])(:-B)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-B
(^|[(\[{])(:\^D)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:^D
(^|[(\[{])(:\^B)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:^B
(^|[(\[{])(:\^D)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:^D
(^|[(\[{])(:\^B)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:^B
(^|[(\[{])(=B)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=B
(^|[(\[{])(=\^B)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^B
(^|[(\[{])(=\^D)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^D
(^|[(\[{])(:’\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:’)
(^|[(\[{])(:’\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:’]

(^|[(\[{])(:'})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:'}
(^|[(\[{])(='\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=')
(^|[(\[{])(='\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=']
(^|[(\[{])(='})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	='}
(^|[(\[{])(<3)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	<3
(^|[(\[{])(\^\.\^)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	^.^
(^|[(\[{])(\^-\^)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	^-^
(^|[(\[{])(\^_\^)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	^_^
(^|[(\[{])(\^\^)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	^^
(^|[(\[{])(:*)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:*
(^|[(\[{])(=*)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=*
(^|[(\[{])(:-*)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-*
(^|[(\[{])(;\))(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	;)
(^|[(\[{])(;\])(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	;]
(^|[(\[{])(;})(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	;}
(^|[(\[{])(:-p)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-p
(^|[(\[{])(:-P)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-P
(^|[(\[{])(:-b)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:-b
(^|[(\[{])(:\^p)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:^p
(^|[(\[{])(:\^P)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:^P
(^|[(\[{])(:\^b)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:^b
(^|[(\[{])(=P)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=P
(^|[(\[{])(=p)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=P
(^|[(\[{])(\\o\\)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	\o\
(^|[(\[{])(/o/)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	/o/
(^|[(\[{])(:P)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:P
(^|[(\[{])(:p)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:p
(^|[(\[{])(:D)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:D
(^|[(\[{])(:d)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:d
(^|[(\[{])(:b)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	:b
(^|[(\[{])(=b)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=B
(^|[(\[{])(=\^p)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^p
(^|[(\[{])(=\^P)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^P
(^|[(\[{])(=\^b)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=^b

(^|[(\[{])(\\o/)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	\o/
(^|[(\[{])(*-*)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	*-*
(^|[(\[{])(=D)(\2+|[.!?,;\)\]\}]+|$)	$1feliz$3	=D

Emoticons negativos

(^|[(\[{])(D:)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	D:
(^|[(\[{])(D=)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	D=
(^|[(\[{])(D-:)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	D-:
(^|[(\[{])(D\^:)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	D^:
(^|[(\[{])(D\^=)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	D^=
(^|[(\[{])(:\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:(
(^|[(\[{])(:\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:[
(^|[(\[{])(:\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:{
(^|[(\[{])(:o\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:o(
(^|[(\[{])(:o\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:o[
(^|[(\[{])(:\^\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:^(
(^|[(\[{])(:\^\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:^[
(^|[(\[{])(:\^\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:^{
(^|[(\[{])(=\^\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=^(
(^|[(\[{])(=\^\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=^{
(^|[(\[{])(>=\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>=[
(^|[(\[{])(>=\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>={
(^|[(\[{])(>=\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>=(
(^|[(\[{])(>:-\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>:-{
(^|[(\[{])(>:-\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>:-[
(^|[(\[{])(>:-\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>:-(
(^|[(\[{])(>=\^\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>=^[
(^|[(\[{])(:-\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:-[
(^|[(\[{])(:-\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:-(
(^|[(\[{])(=\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=(
(^|[(\[{])(=\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=[
(^|[(\[{])(=\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	={
(^|[(\[{])(=\^\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=^[
(^|[(\[{])(>:-=\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>:-=(
(^|[(\[{])(>=\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>=[
(^|[(\[{])(:'\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:'(
(^|[(\[{])(:'\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:'[
(^|[(\[{])(:'\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:'{

(^|[(\[{])(=’\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=’{
(^|[(\[{])(=’\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=’(
(^|[(\[{])(=’\[)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=’[
(^|[(\[{])(=\\)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=\
(^|[(\[{])(:\\)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:\
(^|[(\[{])(=/)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=/
(^|[(\[{])(:/)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:/
(^|[(\[{])(=\$)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	=$
(^|[(\[{])(o\.O)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	o.O
(^|[(\[{])(O_o)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	O_o
(^|[(\[{])(Oo)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	Oo
(^|[(\[{])(:\$:-\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:$:-{
(^|[(\[{])(>=\^\()(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>=^(
(^|[(\[{])(>=\^\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	>=^{
(^|[(\[{])(:o\{)(\2+|[.!?,;\)\]\}]+|$)	$1tristeza$3	:o{

Emoticons neutros

(^|[(\[{])(:|)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:|
(^|[(\[{])(=|)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	=|
(^|[(\[{])(:-|)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:-|
(^|[(\[{])(>\.<)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	>.<
(^|[(\[{])(><)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	><
(^|[(\[{])(>_<)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	>_<
(^|[(\[{])(:o)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:o
(^|[(\[{])(:O)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:o
(^|[(\[{])(:0)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:0
(^|[(\[{])(=O)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	=O
(^|[(\[{])(:@)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:@
(^|[(\[{])(=@)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	=@
(^|[(\[{])(:\^o)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:^o
(^|[(\[{])(:\^@)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:^@
(^|[(\[{])(-\.-)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	-.-
(^|[(\[{])(-\.-’)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	-.-’
(^|[(\[{])(-_-)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	-_-
(^|[(\[{])(-_-’)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	-_-’
(^|[(\[{])(:x)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:x
(^|[(\[{])(=X)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	=X
(^|[(\[{])(=#)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	=#
(^|[(\[{])(:-x)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:-x

(^|[(\[{])(:-@)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:-@
(^|[(\[{])(:-#)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:-#
(^|[(\[{])(:\^x)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:^x
(^|[(\[{])(:\^#)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:^#
(^|[(\[{])(:#)(\2+|[.!?,;\)\]\}]+|$)	$1neutro$3	:#

APÊNDICE D – LISTA DE SUBSTANTIVOS MAIS UTILIZADOS

Palavra	Frequência
qualidade	151392
variedade	83093
pizza	73198
limpeza	34405
padrão	32432
massa	31673
gosto	30970
sabores	30198
experiência	30006
delícia	29870
problema	28668
linda	25808
espera	24283
pé	23374
conforto	21503
aconchegante	19068
beleza	17748
cuidado	16000
surpresa	15562
barulho	14558
problemas	13952
variedades	13470
gostos	13420
tranquilidade	12748
maioria	12136
águas	12125
marido	11905
compras	11798
risoto	11797
ingredientes	11791
porta	11734
oportunidade	11563
diversão	11505
acompanhamentos	11480
churrasco	11409
verdade	11355
leite	11340
pastel	11301
grupo	11282
época	11175
opinião	11153
barracas	11098
apresentação	11025
%	10911
varanda	10854
negócios	10848
média	10840
loja	10835
atividades	10801
prédio	10766
som	10751
sábado	10700
luxo	10681
detalhes	10626
lago	10609
diária	10602
trilha	10594
academia	10574
atração	10379
salão	10379
fome	10377
parada	10358
gastronomia	10352
nível	10351
mercado	10334
toalhas	10318
finais	10199
sistema	10156
cadeiras	10153
animais	10129
graça	10109
eventos	10098
charme	10094
arte	10055
camas	10051
jardim	9956
drinks	9930
infraestrutura	9908
sorvetes	9808
sushi	9662
estrelas	9627
tempero	9624
carte	9587
apartamento	9570
rodizio	9539

Exposições	9510
apartamentos	9485
poder	9481
espetáculo	9368
simpatia	9355
ambientes	9350
estação	9327
casais	9269
camarões	9266
criança	9240
frigobar	9232
gerente	9200
feijoada	9186
bebida	9085
ok	9028
filas	9024
filhos	9018
roupas	9002
shows	8994
demora	8985
manutenção	8826
bolos	8750
visão	8746
entradas	8688
férias	8687
quiosques	8665
movimento	8633
adultos	8607
calor	8589
padaria	8543
sanduíches	8511
couvert	8508
site	8485
maravilhosa	8453
Metrô	8418
ondas	8371
avenida	8350
barco	8311
temporada	8288
filho	8283
estádio	8213
turismo	8208
cinema	8183
paraíso	8161
obras	8157

Porto	8139
diversidade	8091
expectativas	8027
impressão	7980
bolinho	7928
direito	7924
turista	7888
ilha	7880
happy	7875
guia	7873
dinheiro	7870
taxi	7856
itens	7837
feirinha	7833
fondue	7829
paz	7808
quilo	7793
dono	7760
sinal	7739
moqueca	7662
pedidos	7597
encontro	7595
frios	7589
decepção	7557
coco	7507
diferença	7484
queijos	7452
cima	7447
filha	7433
estabelecimento	7423
chef	7413
noites	7382
costela	7352
batatas	7302
data	7272
período	7272
táxi	7255
salmão	7243
futebol	7235
domingos	7231
sanduíche	7217
programa	7201
sensação	7182
jogos	7180
tradição	7170

lanchonete	7165
boca	7164
contato	7151
carnaval	7113
banana	7085
saída	7065
peça	7059
calçadão	7041
justo	7029
maravilhoso	7022
aniversário	6980
cortesia	6905
caminho	6895
momentos	6866
transporte	6862
unidade	6855
barraca	6817
descanso	6813
sorte	6811
meia	6781
chegada	6774
obra	6765
áreas	6721
dúvidas	6718
lagoa	6707
Estado	6696
funcionarios	6692
interior	6669
metros	6622
avaliação	6573
redor	6572
capital	6565
iluminação	6532
vila	6524
degustação	6500
alto	6500
detalhe	6479
cardapio	6464
resto	6464
informações	6462
temperatura	6455
balcão	6440
exposição	6433
luz	6408
elevador	6374

nordeste	6358
Búzios	6356
roupa	6290
famílias	6277
questão	6261
sugestão	6254
bolo	6253
hambúrguer	6250
si	6193
recheio	6178
estrada	6160
maré	6148
coração	6143
peças	6130
ruas	6112
limpa	6107
carros	6101
pedras	6089
quadras	6056
morro	6054
evento	6025
jardins	6011
público	5954
alimentos	5934
caminhadas	5875
sauna	5874
Cantina	5853
molhos	5828
festa	5815
feijão	5797
proposta	5795
chefe	5787
organização	5773
recreação	5761
brinquedos	5759
maravilha	5752
construção	5734
medo	5705
casas	5694
calma	5673
disposição	5655
cortes	5643
manobrista	5629
pista	5625
staf	5609

lindas	5609
mergulho	5600
cebola	5592
horários	5580
dicas	5572
centavo	5564
piso	5555
foto	5549
fundo	5533
comentários	5530
valores	5524
fama	5523
hamburguer	5517
farofa	5512
torta	5507
início	5505
lua	5502
donos	5499
funcionário	5482
alho	5460
passagem	5441
pedra	5410
visitantes	5396
tortas	5388
mofo	5386
cobertura	5348
frio	5339
sorveteria	5237
so	5221
parmegiana	5171
chá	5165
ponta	5155
janeiro	5146
exemplo	5141
chão	5136
hóspede	5130
jeito	5115
olhos	5080
shoppings	5077
hoteis	5076
madeira	5060
medida	5014
cabo	5011
preco	4998
experiencia	4955

respeito	4944
prazer	4939
linha	4925
requinte	4908
dificuldade	4898
trilhas	4892
programação	4891
papo	4881
ida	4878
bife	4867
quadra	4856
creme	4836
sushis	4835
service	4803
mês	4788
cafés	4784
box	4776
toque	4775
tomate	4774
indicação	4771
cachoeira	4766
mirante	4750
terra	4748
higiene	4745
árvores	4740
amo	4732
país	4730
boteco	4726
carinho	4726
sábados	4716
número	4711
esportes	4694
maneira	4689
ha	4684
metro	4653
cordeiro	4638
andares	4620
tapiocas	4609
acompanhamento	4608
caipirinha	4597
paciência	4597
ai	4593
chocolates	4583
wifi	4581
boas	4574

paladar	4559
ponte	4549
pressa	4519
pacote	4518
salas	4518
lagosta	4517
fã	4508
promoção	4507
faixa	4506
sombra	4472
min	4451
crepes	4450
preparo	4437
idades	4427
expectativa	4415
visitação	4411
verão	4398
viagens	4397
crepe	4390
uso	4389
mão	4359
pais	4358
recepcionista	4348
amigo	4338
area	4329
chuva	4328
madrugada	4314
alta	4288
zona	4277
chapa	4276
comércio	4264
proprietário	4262
pimenta	4247
limão	4246
grupos	4236
Dunas	4233
Polvo	4231
copa	4219
belas	4209
bicicleta	4204
mel	4194
acomodação	4175
excelência	4173
espaços	4158
top	4153

desconto	4141
Campinas	4127
telefone	4124
aspecto	4109
vantagem	4107
lanchonetes	4101
localizacao	4094
atendente	4093
janela	4086
bola	4080
namorado	4074
guias	4056
empresa	4053
sono	4043
jogo	4010
fins	3993
estada	3981
fim_de_semana	3978
rs	3970
pousadas	3970
bonito	3966
manteiga	3953
ingresso	3942
acervo	3930
metade	3912
lojinhas	3911
cidades	3896
garrafa	3893
bela	3892
especialidade	3891
caixa	3885
paredes	3876
acústica	3864
artistas	3850
proprietários	3850
termos	3844
litoral	3841
meses	3833
cartões	3831
máximo	3829
forno	3823
taxa	3818
ducha	3814
self-service	3807
caranguejo	3803

altura	3799
monitores	3794
encantador	3793
calçada	3789
conjunto	3775
porco	3774
educação	3769
Km	3762
passado	3753
pagamento	3747
nota	3745
spa	3729
delícias	3700
conservação	3693
sossego	3683
cordialidade	3677
retorno	3668
namorada	3667
bancos	3663
tratamento	3658
tempos	3657
energia	3657
trem	3648
possibilidade	3640
capricho	3640
palco	3630
pedaço	3629
pastéis	3628
acordo	3623
casamento	3620
barzinho	3615
categoria	3612
amantes	3609
canto	3597
paisagens	3585
fast	3580
povo	3575
Cataratas	3573
artesanatos	3573
banheira	3571
refrigerante	3552
compra	3550
morango	3546
casquinha	3533
conversa	3528
facilidade	3527
visitas	3520
prédios	3508
sal	3506
legumes	3505
maionese	3488
troca	3478
infra	3456
humor	3453
amiga	3450

APÊNDICE E – LISTA DE ADJETIVOS MAIS UTILIZADOS

Palavra	Frequência
bom	585068
excelente	400515
boa	351759
ótimo	246011
melhor	188168
agradável	179171
ótima	168765
grande	90401
maravilhoso	77869
melhores	76079
lindo	74490
bons	60280
simples	59965
caro	58587
bonito	56377
maravilhosa	55280
legal	50006
atenciosos	49506
próximo	48859
delicioso	47829
perfeito	47652
linda	46758
gostoso	46476
aconchegante	45794
boas	44829
bonita	42083
confortável	41614
razoável	40280
deliciosa	39783
pequeno	37221
interessante	36758
ruim	36514
limpo	36484
fácil	32464
rápido	31085
ótimos	30892
saborosa	30801
tradicional	30127
incrível	30010
especial	29438
impecável	28958
tranquilo	28246
ótimas	27320
bela	27231
deliciosos	26253
atencioso	26216
gostosa	25965
limpos	25214
sensacional	24500
ideal	24372
amplo	24279
confortáveis	24227
pequena	23840
principal	23771
imperdível	22931
barato	22927
perfeita	21867
bacana	21724
péssimo	21356
novo	20969
saboroso	20737
quente	20508
espetacular	19944
belo	19539
enorme	18902
alto	18610
fantástico	17799
saborosos	16687
deliciosas	16602
simpáticos	16356
diferentes	16339
antigo	15976
prestativos	15533
difícil	14993
simpático	14777
familiar	14314
tranquila	13422
negativo	13299
acessível	12224
quentes	12154
farto	11856
excepcional	11847
completo	11840
maravilhosos	11762
pequenos	11742

romântico	11653
acolhedor	11586
espaçoso	11357
fraco	11350
igual	11007
fantástica	10987
alta	10877
eficiente	10599
charmoso	10509
maravilhosas	10425
livre	10312
horrível	9917
famoso	9894
caros	9717
central	9712
feliz	9546
pequenas	9470
certo	9443
limpa	9426
típica	9401
histórico	9219
segunda	9095
sorridente	9067
externa	9066
comum	9063
naturais	8957
leve	8885
saborosas	8856
baixa	8814
diversas	8724
pior	8667
gostosos	8659
cordial	8659
tradicionais	8658
obrigatória	8647
natural	8614
péssima	8605
turísticos	8470
moderno	8457
nova	8342
pertinho	8110
inesquecível	8068
regional	8068
antiga	8033
especiais	7902

agradáveis	7877
interessantes	7847
claro	7834
necessário	7794
última	7677
atenciosa	7654
distante	7578
seguro	7524
turístico	7517
importante	7492
impossível	7413
diferencial	7398
frita	7355
calmo	7318
razoáveis	7289
famosa	7201
típico	7103
cultural	7066
limpas	7034
novos	7001
espaçosos	6996
surpreendente	6955
italiano	6941
bonitas	6797
gratuito	6747
antigos	6730
típicas	6652
deslumbrante	6575
divino	6528
Proximo	6487
altos	6451
rápida	6340
barulhento	6321
regionais	6298
gentis	6093
preciso	6084
superior	6015
fina	5992
ampla	5925
farta	5918
maiores	5915
fria	5899
lindos	5884
velho	5767
menor	5758

própria	5715
inteiro	5704
barata	5703
justos	5674
próximos	5650
típicos	5650
sujo	5633
vazio	5603
último	5588
frescos	5578
gostosas	5535
positivo	5499
completa	5499
atrativo	5496
disponível	5494
disponíveis	5493
branco	5483
saudável	5456
gentil	5390
acessíveis	5385
paulista	5323
nordestina	5269
impressionante	5251
silencioso	5233
segundo	5232
novas	5229
simpática	5180
razoavel	5163
japonês	5117
lento	5093
legais	5040
baixo	5023
frio	4989
total	4946
incríveis	4919
crocante	4913
caseira	4884
lindas	4877
suficiente	4817
prestativo	4812
obrigatório	4803
negativos	4785
fresco	4745
enormes	4681
honesto	4665

solícitos	4661
charmosa	4609
infantil	4569
bonitos	4483
brasileira	4467
internacional	4465
certa	4462
interna	4429
atrativos	4399
árabe	4396
executivo	4363
incluso	4347
básico	4331
seguinte	4315
compatível	4314
complexo	4250
rústico	4219
longo	4197
elegante	4181
grátis	4145
antigas	4127
próximas	4101
alemã	4096
absurdo	4083
próprio	4080
verdadeiro	4071
tb	3979
longa	3965
comuns	3956
moderna	3948
carioca	3917
direto	3905
meia	3895
impecáveis	3886
interno	3848
azul	3815
aconchegantes	3809
positivos	3754
fartos	3721
prático	3703
verdadeira	3687
gigante	3633
postal	3626
original	3619
belas	3603

confortavel	3602
sul	3598
macia	3534
pronto	3528
individuais	3429
rápidos	3364
hour	3345
médio	3339
perfeitos	3331
mau	3330
imensa	3330
alegre	3322
sozinho	3320
artesanais	3315
colonial	3298
gratuita	3290
brasileiro	3274
baratos	3270
generosas	3251
imperdivel	3249
cordiais	3244
famosos	3200
divina	3187
norte	3186
confuso	3182
oriental	3172
fortes	3159
térreo	3144
mineira	3140
exuberante	3133
relaxante	3133
suja	3130
grata	3129
emocionante	3120
satisfatório	3112
inúmeras	3108
puro	3100
mexicana	3097
artesanal	3089
seca	3081
desagradável	3050
correto	3049
noturna	3045
culturais	3036
perigoso	3007

seco	3000
má	2982
externo	2975
público	2975
ruins	2964
panorâmica	2960
negra	2907
externas	2893
gastronômica	2834
justo	2825
portuguesa	2824
recomendável	2824
individual	2822
fraca	2805
ambulantes	2802
pura	2790
profissional	2787
inteira	2780
nobre	2767
comercial	2751
cortês	2742
fritas	2730
escuro	2724
extra	2705
aquático	2693
imponente	2666
mediano	2657
caseiro	2610
perfeitas	2603
baiana	2584
convidativo	2577
sensacionais	2547
luxuoso	2537
mto	2527
frios	2481
francês	2455
inigualável	2451
famosas	2441
atento	2431
al	2429
calmas	2405
atenciosas	2398
urbana	2390
ágil	2379
velhos	2374

temático	2349
rústica	2348
intimista	2348
presente	2343
dificil	2343
jovem	2335
frito	2333
quentinho	2327
extensa	2310
velha	2307
cremoso	2296
astral	2289
funcional	2284
menores	2276
possivel	2261
nobres	2253
so	2252
terceira	2246
limpinho	2243
popular	2229
indescritível	2213
belos	2211
regular	2209
catedral	2207
fresca	2207
prestativa	2202
romantico	2181
frescas	2180
nacional	2170
caras	2168
Americano	2158
curta	2158
importantes	2149
informal	2147
turística	2143
histórica	2137
terrível	2134
fininha	2121
simpáticas	2112
idosos	2110
atraente	2110
celular	2106
fresquinho	2096
top	2087
modernos	2075

anterior	2069
duro	2059
prontos	2059
branca	2053
kids	2037
aprazível	2035
últimos	2027
inferior	2026
eficientes	2016
lamentável	2008
japoneses	1993
sujos	1985
mágico	1982
vermelhas	1977
fantásticos	1972
rico	1964
cristalina	1957
condizente	1954
atentos	1949
primoroso	1939
perigosa	1935
unico	1933
italianos	1925
bacanas	1924
breve	1923
imperdíveis	1923
fixo	1922
curto	1915
confortaveis	1912
leves	1900
nacionais	1886
novinho	1879
repleto	1876
desconfortável	1872
segura	1864
fartas	1849
anexo	1848
chamado	1841
divinos	1837
apto	1833
espetaculares	1822
food	1821
baratas	1817
honesta	1807
digno	1806

chato	1795
musical	1793
francesa	1789
imenso	1784
saudáveis	1782
indispensável	1776
exemplar	1773
particular	1769
rapido	1759
repelente	1747
vermelha	1734
argentino	1731
física	1719
facil	1716
incomparável	1715
responsável	1713
cansativo	1712
exclusivo	1708
municipal	1708
fenomenal	1689
sofrível	1685
sujas	1680
varios	1679
tipica	1678
constante	1677
fino	1676
receptivos	1675
generosa	1672
rodoviária	1670

acolhedora	1663
simpaticos	1659
simpatico	1656
brasileiros	1650
inúmeros	1644
transparente	1644
urgente	1631
amplas	1624
decepcionante	1615
independente	1611
históricos	1610
suave	1603
internacionais	1602
cristalinas	1600
romântica	1599
vip	1598
cru	1594
econômico	1590
x	1587
acústico	1583
ímpar	1580
macio	1577
português	1574
negro	1569
rica	1569
pronta	1568

APÊNDICE F – LISTA DE STOP-WORDS UTILIZADA

1	2	3	4	5
6	7	8	9	10
-	,	:	!	?
.	=-se	a	acesso	achei
acho	acomodações	adoro	aeroporto	agora
água	ainda	além_de	além_disso	além_do
algo	algum	algumas	alguns	ali
alimentação	almoçar	almoço	ambiente	amigos
amplos	andar	ano	anos	antes
apenas	após	aproveitar	aquele	aqui
ar	área	areia	arquitetura	arroz
artesanato	as	assim	até	atenção
atendentes	atendimento	Atrações	bacalhau	bairro
banheiro	banheiros	banho	bar	bares
bastante	batata	Bebidas	beira	bem
benefício	Brasil	Brasília	bufet	cada
cafe	café	cama	camarão	caminhada
cara	cardápio	carne	carnes	carro
carta	cartão	casa	casal	caso
cedo	centro	certeza	cerveja	cervejas
check	Chega	chegamos	chegar	cheia
cheio	cheiro	chocolate	chope	churrascaria
chuveiro	cidade	cliente	clientes	clima
coisa	coisas	com	comemos	comer
comi	comida	comidas	como	comprar
condicionado	conferir	conhecer	conta	cozinha
crianças	culinária	cultura	curitiba	curtir
custo	custo-benefício	dá	dar	de
decoração	Decorado	deixa	deixar	deixe
deixou	demais	dentro_do	depois	desde
desejar	Destaque	deve	devido	dia
dias	dica	diferente	diversos	dizer
doce	doces	dois	domingo	duas
durante	dúvida	e	é	ela

ele	eles	em	encontra	Encontrar
enfim	então	entrada	entrar	entre
equipe	era	escolha	escolher	espaço
esperava	esposa	essa	esse	esta
está	estacionamento	Estadia	estão	estar
estava	estavam	este	estilo	estive
estiver	estrutura	etc	eu	existe
existem	experimentar	extremamente	falar	falta
Familia	família	fato	faz	fazem
fazer	feira	feito	fica	ficam
ficamos	ficar	ficou	fila	filé
fim	final	fiquei	foi	fomos
For	fora	foram	forma	Fortaleza
forte	Fotos	frango	frente	frutas
frutos	fui	funcionários	garçom	garçons
gente	geral	Gramado	grandes	há
havia	história	hoje	hora	horário
horas	Hospedagem	Hospedar	hóspedes	hostel
hotéis	hotel	Ibis	igreja	inclusive
indico	infelizmente	instalações	internet	ir
isso	italiana	ja	já	jantar
japonesa	la	lá	lado	lanche
lanches	lazer	levar	Locais	local
localização	localizada	localizado	logo	Lojas
lugar	lugares	maior	mais	manha
manhã	mar	massas	me	meio
menu	mesa	mesas	mesma	mesmo
meu	meus	mim	minha	minutos
molho	Momento	muita	muitas	muito
muitos	mundo	museu	música	natal
natureza	nenhum	noite	nome	normal
nos	nossa	nosso	novamente	o
o_que	oferece	onde	ônibus	opção
opções	orla	os	ou	outra
outras	outro	outros	pães	pagar

paisagem	pão	para	parabéns	Parece
parque	parte	passar	passear	passeio
Passeios	pedi	pedida	pedido	pedimos
pedir	Pegar	peixe	peixes	pena
perto	perto_do	pessoa	pessoal	pessoas
petiscos	picanha	piscina	piscinas	piza
pizaria	pizas	pode	poderia	pois
ponto	pontos	por	porção	porções
porque	Possível	possui	pouco	pousada
pra	praça	praia	praias	prato
pratos	precisa	preço	preços	primeira
primeiro	principais	principalmente	produtos	provar
próxima	qualquer	quando	quantidade	quanto
quarto	quartos	quase	que	queijo
quem	quer	quiser	reais	realmente
recepção	rede	refeição	refeições	reforma
região	relação	reserva	Resort	restaurante
restaurantes	rio	Rio_de_Janeiro	rodízio	rua
sabor	Sair	sala	salada	saladas
salgados	Salvador	são	se	segurança
sei	seja	sem	semana	sempre
sendo	ser	seria	serve	servem
serviço	serviços	servida	servido	servidos
seu	Seus	shopping	show	sim
simplesmente	só	sobre	sobremesa	sobremesas
sol	somente	sorvete	sou	SP
sua	suas	suco	sucos	talvez
tamanho	também	tanto	tão	tapioca
tarde	te	teatro	tem	tempo
tenho	ter	tinha	tipo	tipos
tirar	tive	tivemos	toda	todas
todo	todos	tomar	Trabalho	três
tudo	turistas	tv	um	um_pouco
uma	umas	única	único	uns
vai	vale	valor	várias	vários

veio	vem	ver	verde	vez
vezes	vi	viagem	vida	vinho
vinhos	visita	visitar	vista	visual
vivo	você	volta	voltar	voltarei
voltaria	vontade	vou		

APÊNDICE G –ADVERSATIVOS, NEGATIVOS E DELIMITADORES

Palavras negativas:

não; tampouco; nem; nunca; jamais

Palavras adversativas:

mas; porém; contudo; todavia; entretanto; embora; apesar

Delimitadores:

(;); {; }; "; '; <; >;

APÊNDICE H – AMOSTRA DE ANÁLISE REALIZADA PELA FERRAMENTA

Sentença	Palavra opinativa	Aspecto	Polaridade
piscinas só tem urina com cloro, preços absurdos, se precisar beber água morre de sede, pois a mesma e somente vendida a 5 reais achei um absurdo.	absurdos	preços	-1
piscinas só tem urina com cloro, preços absurdos, se precisar beber água morre de sede, pois a mesma e somente vendida a 5 reais achei um absurdo.	absurdos	urina com cloro	-1
piscinas só tem urina com cloro, preços absurdos, se precisar beber água morre de sede, pois a mesma e somente vendida a 5 reais achei um absurdo.	absurdo	reais	-1
ainda te fará voltar enriquecido com tanto conhecimento!	enriquecido	conhecimento	1
O caminho para o paraíso não é fácil.	paraíso	caminho	1
O caminho para o paraíso não é fácil.	não_fácil		-1
na maré baixa, inclusive, é possível ir andado, ou então pagar procurar as pequenas embarcações por lá.	pequenas	embarcações	-1
a praia mais linda de Alagoas que conheci.	linda	praia	1
a agência fica dentro do hotel 10 de julho que facilitou nós fecharmos o passeio.	facilitou	julho	1
a beleza da Amazônia e o tratamento que me foi dispensado deixaram uma marca indelével, tornando inesquecível aquele fim de semana em Manaus.	beleza		1
a beleza da Amazônia e o tratamento que me foi dispensado deixaram uma marca indelével, tornando inesquecível aquele fim de semana em Manaus.	inesquecível	fim_de_semana	4
a beleza da Amazônia e o tratamento que me foi dispensado deixaram uma marca indelével, tornando inesquecível aquele fim de semana em Manaus.	beleza		1
a beleza da Amazônia e o tratamento que me foi dispensado deixaram uma marca indelével, tornando inesquecível aquele fim de semana em Manaus.	inesquecível	fim_de_semana	4
fizemos mais um passeio de lancha maravilhoso onde ele nos mostrou Manaus de diversas formas.	maravilhoso	lancha	1
a mistura perfeita entre exclusividade e simplicidade.	perfeita	mistura	1
na cabeça do manauara, você é turista e está trazendo muito dinheiro. e nesta de te indicarem barcos contratados para este passeio, estão ganhando parte do valor que contratam com você.	ganhando	parte	-1
ao longo do caminho, pelo Rio são Francisco, podem ser vistas formações rochosas muito bonitas, garantindo a bela vista.	bonitas	formações rochosas	1
ao longo do caminho, pelo Rio são Francisco, podem ser vistas formações rochosas muito bonitas, garantindo a bela vista.	garantindo		1
ao longo do caminho, pelo Rio são Francisco, podem ser vistas formações rochosas muito bonitas, garantindo a bela vista.	bela	vista	1

o custo benefício é enorme em relação ao que pude observar nos outros passeios, o preço é muito justo.	enorme	custo benefício	-1
o custo benefício é enorme em relação ao que pude observar nos outros passeios, o preço é muito justo.	justo	preço	1
o custo benefício é enorme em relação ao que pude observar nos outros passeios, o preço é muito justo.	justo	passeios	1
engraçado que a Casa do Papai Noel em si eu não visitei, já tinha visto fotos pela internet e achei que não valria a pena visitar, mas claro, passeei pelo shopping, fui a várias lojinhas e experimentei o delicioso churros de nutela!	claro		1
engraçado que a Casa do Papai Noel em si eu não visitei, já tinha visto fotos pela internet e achei que não valria a pena visitar, mas claro, passeei pelo shopping, fui a várias lojinhas e experimentei o delicioso churros de nutela!	delicioso	churros de nutela	1
o café da manhã de ótima qualidade e com variedade.	ótima	qualidade	1
o café da manhã de ótima qualidade e com variedade.	ótima	variedade	1
o café é excelente e os chalés muito limpos.	excelente	café	1
o café é excelente e os chalés muito limpos.	limpos	chalés	1
senti que fui assaltado, pois paguei valor por dia somente com café da manhã e depois fui informado que poderia alugar do proprietário por valor menor (valor).	assaltado		6
senti que fui assaltado, pois paguei valor por dia somente com café da manhã e depois fui informado que poderia alugar do proprietário por valor menor (valor).	menor	valor	-1
as mesas do café da manhã estavam sempre sujas. em alguns dias tive que deixar o carro na rua, pois não havia vaga no estacionamento (se não me engano há 3 vagas).	sujas	café de a manhã	-1
as mesas do café da manhã estavam sempre sujas. em alguns dias tive que deixar o carro na rua, pois não havia vaga no estacionamento (se não me engano há 3 vagas).	não_vaga	estacionamento	1
as mesas do café da manhã estavam sempre sujas. em alguns dias tive que deixar o carro na rua, pois não havia vaga no estacionamento (se não me engano há 3 vagas).	não_engano		1
as mesas do café da manhã estavam sempre sujas. em alguns dias tive que deixar o carro na rua, pois não havia vaga no estacionamento (se não me engano há 3 vagas).	vagas		-1
fica a poucas quadras de um ótimo shopping, transporte, comércio etc. tem agência de viagens no hotel, um café da manhã delicioso, com muita variedade, e um ótimo restaurante para as demais refeições.	ótimo	shopping	1
fica a poucas quadras de um ótimo shopping, transporte, comércio etc. tem agência de viagens no hotel, um café da manhã delicioso, com muita variedade, e um ótimo restaurante para as demais refeições.	ótimo	transporte	1
fica a poucas quadras de um ótimo shopping, transporte, comércio etc. tem agência de viagens no hotel, um café da manhã delicioso, com muita variedade, e um ótimo restaurante para as demais refeições.	ótimo	comércio	1
fica a poucas quadras de um ótimo shopping, transporte, comércio etc. tem agência de viagens no hotel, um café da manhã delicioso, com muita variedade, e um ótimo restaurante para as demais refeições.	delicioso	café de a manhã	1
fica a poucas quadras de um ótimo shopping, transporte, comércio etc. tem agência de viagens no hotel, um café da manhã delicioso, com muita variedade, e um ótimo restaurante para as demais refeições.	delicioso	hotel	1
fica a poucas quadras de um ótimo shopping, transporte, comércio etc. tem agência de viagens no hotel, um café da manhã delicioso, com muita variedade, e um ótimo restaurante para as demais refeições.	ótimo	restaurante	1

8 ANEXOS

ANEXO A – ETIQUETAS MORFOSSINTÁTICAS COGROO

1 Etiquetas de Função[13]

1.1. Argumentos ao nível da oração (governados por valência)	
símbolo	categoria
SUBJ	sujeito (incluindo sujeitos impessoais se)
ACC	objeto direto (incluindo alguns tipos de se)
ACC-PASS	função do clítico se numa oração passiva (partícula apassivante)
DAT	objeto indireto pronominal (incluindo se)
PIV	objeto preposicional
SA	complemento adverbial [pode ser substituído por um pronome adverbial] (relativo ao sujeito)
OA	complemento adverbial (relativo ao objeto)
SC	predicativo do sujeito
OC	predicativo do objeto
COM	complementador em estruturas de comparação (como, (do) que)
P	predicador
PMV	em contexto de coordenação, verbo principal coordenado com os seus próprios constituintes
PAUX	em contexto de coordenação, verbo auxiliar partilhado por verbos principais com os seus próprios constituintes
MV	verbo principal
AUX	verbo auxiliar
AUX<	em contexto de coordenação, partícula de ligação entre o auxiliar partilhado e verbos coordenados
PRT-AUX	partícula de ligação verbal
1.2 Adjuntos oracionais (não governados por valência)	
ADVL	adjunto adverbial
PRED	adjunto predicativo
PASS	agente da passiva
S<	aposto da oração
>S	dependente de complementador
VOC	constituinte vocativo
FOC	marcador de foco
TOP	constituinte de tópico
1.3 Subordinação e coordenação	
SUB	subordinador
CO	coordenador
CJT	elemento conjunto
PCJT	preposição conjunta (de/desde.....a/até/para)

[13] Fonte: http://www.linguateca.pt/floresta/BibliaFlorestal/anexo1.html

1.4 Função de um constituinte a nível do grupo	
H	núcleo
>N	adjunto adnominal (antecede o núcleo)
N<	adjunto adnominal (segue o núcleo)
N<ARG	complemento nominal (complementa um substantivo não deverbal)
N<ARGS	complemento nominal (complementa um substantivo deverbal, relativo ao sujeito)
N<ARGO	complemento nominal (complementa um substantivo deverbal, relativo ao objeto)
APP	aposição (do substantivo) [epíteto de identidade]
N<PRED	adjeto predicativo [epíteto predicativo]
>A	dependente em adjp ou advp (antecede o núcleo)
A<	dependente em adjp ou advp (segue o núcleo)
NUM<	dependente de numeral
KOMP<	complemento comparativo
>P	dependente da preposição
P<	argumento de preposição
1.5 Tipos de frase (função)	
STA	enunciado declarativo
QUE	enunciado interrogativo
CMD	enunciado imperativo
EXC	enunciado exclamativo

2 Etiquetas de Forma[14]

2.1. Formas de grupo	
símbolo	categoria
np	sintagma nominal (H: nome ou pronome)
adjp	sintagma adjetival (H: adjetivo ou determinante)
advp	sintagma adverbial (H: advérbio)
vp	sintagma verbal (contém sempre MV e poderá exibir AUX)
pp	sintagma preposicional (H: preposição)
cu	sintagma evidenciador de relação de coordenação
sq	sequência de funções discursivas; sequência de elementos identificadores do falante, tema, etc. e do discurso propriamente dito
2.2. Formas oracionais	
fcl	oração finita
icl	oração não-finita
acl	oração adverbial
2.3. Categoria gramatical (nós terminais)	

[14] Fonte: http://www.linguateca.pt/floresta/BibliaFlorestal/anexo1.html

n	nome, substantivo
prop	nome próprio
adj	adjetivo
n-adj	flutuação entre substantivo e adjetivo
v-fin	verbo finito
v-inf	infinitivo
v-pcp	particípio
v-ger	gerúndio
art	artigo
pron-pers	pronome pessoal
pron-det	pronome determinativo
pron-indp	pronome independente (com comportamento semelhante ao nome)
adv	advérbio
num	numeral
prp	preposição
intj	interjeição
conj-s	conjunção subordinativa
conj-c	conjunção coordenativa