UNIVERSIDADE DE CAXIAS DO SUL
ÁREA DE CONHECIMENTO DE CIÊNCIAS EXATAS E ENGENHARIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

GUILHERME HENRIQUE SANTOS ANDREATA

O Uso de Processamento de Linguagem Natural para a Análise de Sentimentos na Rede Social Reddit.

André Luis Martinotto
Orientador

Caxias do Sul, Junho de 2017

O Uso de Processamento de Linguagem Natural para a Análise de Sentimentos na Rede Social Reddit.

por

Guilherme Henrique Santos Andreata

Projeto de Diplomação submetido ao curso de Bacharelado em Sistemas de Informação da área de conhecimento de ciências exatas e engenharia, como requisito obrigatório para graduação.

Projeto de Diplomação

Orientador: André Luis Martinotto
Banca examinadora:

André Gustavo Adami
CCTI/UCS
Carlos Eduardo Nery
CCTI/UCS

Projeto de Diplomação apresentado em
5 de Dezembro de 2013

Daniel Luís Notari
Coordenador

SUMÁRIO

LISTA DE ACRÔNIMOS

NLP	*Natural Language Processing*
NLTK	*Natural Language Toolkit*
MaxEnt	*Maximum Entropy*
RNTN	*Recursive Neural Tensor Networks*
VADER	*Valence Aware Dictionary and sEntiment Reasoner*
JSON	*JavaScript Object Notation*
POJO	*Plain Old Java Objects*
SVM	*Support Vector Machines*

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Palavras-chave: Kinect, Blender, Animação 3D.

1 INTRODUÇÃO

A linguagem é a forma com que nós nos comunicamos, seja ela escrita ou falada. De fato, a linguagem é a forma como expressamos nossas idéias, sentimentos e experiências. O Processamento de Linguagem Natural, é o termo utilizado para descrever um software ou componente de hardware que tem como função analisar a linguagem escrita ou falada (JACKSON; MOULINIER, 2007).

Existem duas abordagens para o Processamento de Linguagem Natural, sendo que a primeira delas é chamada de simbólica (ou racionalista) e a outra de empírica (ou estatística). A primeira abordagem consiste em uma série de regras para a manipulação de símbolos, como as regras gramaticais, que permitem identificar se uma frase está malformada ou não. A abordagem empírica está centrada na análise estatística da linguagem através de grandes quantidades de textos, como por exemplo, a utilização de modelos de Markov para reconhecer padrões na escrita (JACKSON; MOULINIER, 2007).

Existem diversos *frameworks open source* que facilitam o desenvolvimento de *softwares* para o Processamento de Linguagem Natural, sendo que dentre esses se destacam *Stanford's Core NLP Suite* (Stanford CoreNLP, 2017), *Natural Language Toolkit* (Natural Language Toolkit, 2017), *Apache OpenNLP* (Apache OpenNLP, 2017) e *Spacy* (Spacy, 2017). Esses *frameworks* nos permitem, entre outras coisas, efetuar análise de sentimentos, identificar tópicos e conteúdos.

A rede social Reddit é o vigésimo terceiro *website* mais acessado na Internet e o sétimo mais acessado nos Estados Unidos da América (Alexa, 2017). Através deste *website*, seus usuários podem criar ou se inscrever em comunidades, também conhecidas como *subreddits*. Uma vez que as comunidades são criadas pelos próprios usuários, podemos encontrar comunidades sobre todos os assuntos, sejam notícias do mundo, comunidades partidárias, comunidades criadas para pessoas de uma mesma localidade, comunidades de imagens engraçadas, etc.

Nestas comunidades é possível visualizar e comentar *links* enviados por outros usuários. Além disso, o usuário pode efetuar um voto de forma positiva, caso acredite que aquele *link* é útil para a comunidade, ou um voto negativo em caso contrário.

Uma vez que os próprios usuários podem submeter *links*, os eventos e notícias de todo o mundo são reportados no *website*, como exemplo, pode-se citar as eleições ocorridas no ano de 2016 nos Estados Unidos e o tiroteio ocorrido em Paris em 15 de Novembro de 2015.

Dentro deste contexto neste trabalho será desenvolvido um software que permita realizar a análise dos comentários do *website* Reddit. Mais especificamente os comentários do Reddit serão analisados com o objetivo de identificar padrões de sentimentos, ou seja, determinar se a opinião expressada com relação a um determinado tópico é neutra, positiva ou negativa.

1.1 Objetivos do Trabalho

Este trabalho tem como objetivo a análise dos comentários disponíveis no *website* Reddit, identificando padrões de sentimentos entre os usuários de suas comunidades. De forma a atingir o objetivo principal desse trabalho, os seguintes objetivos específicos devem ser realizados:

- Desenvolver uma ferramenta para o Processamento Natural de Linguagem através de *frameworks* já existentes.
- Construção de uma base de dados a partir do *website* Reddit.
- Efetuar o processamento da base de dados criado utilizando-se a ferramenta desenvolvida.

1.2 Estrutura do Trabalho

2 PROCESSAMENTO DE LINGUAGEM NATURAL

O objetivo da área de Processamento de Linguagem Natural é analisar a linguagem natural, ou seja, a linguagem utilizada pelo seres humanos seja ela escrita ou falada (MANNING; SCHüTZE, 1999).

O Processamento de Linguagem Natural é uma área antiga, sendo anterior a invenção dos computadores modernos. De fato, sua primeira grande aplicação foi um dicionário desenvolvido no Birkbeck College em Londres no ano de 1948. Por ser uma área complexa, seus primeiros trabalhos foram notavelmente falhos o que causou uma certa hostilidade por parte das agências fomentadoras de pesquisas.

Os primeiros pesquisadores eram muitas vezes bilíngues, como por exemplo, nativos alemães que imigraram para os Estados Unidos. Acreditava-se que pelo fato desses terem conhecimento de ambas as linguas, Ingles e Alemão, eles teriam capacidade de desenvolver programas de computadores que efetuariam a tradução de modo satisfatório. Uma vez que esses encontraram muitas dificuldades, ficou claro que o maior problema não era o conhecimento das línguas, e sim como expressar esse conhecimento na forma de um programa de computador (HANCOX, 2017).

Para que um computador seja capaz de interpretar uma língua, primariamente necessitamos compreender como nós efetuamos essa interpretação. Por isso, uma parte considerável do Processamento de Linguagem Natural está apoiado na área de Linguística.

2.1 Linguística

O objetivo da Linguística é compreender como os seres humanos adquirem, produzem e entendem as diversas línguas, ou seja, a forma como conversamos, a nossa escrita e outras mídias de comunicação (MANNING; SCHüTZE, 1999).

Na linguagem tanto escrita, como na falada, existem regras que são utilizadas para estruturar as expressões. Uma série de dificuldades no Processamento de Linguagem Natural são ocasionadas pelo fato de que as pessoas constantemente mudam essas regras para satisfazerem suas necessidades de comunicação (MANNING;

SCHüTZE, 1999). Uma vez que as regras são constantemente modificadas pelo locutor, se torna extremamente difícil a criação de um software ou hardware que efetue a interpretação de uma língua.

2.2 Métodos de Processamento de Linguagem Natural

O *Natural Language Processing* (NLP) tem como objetivo a execução de diferentes tarefas, como por exemplo, a categorização de documentos, a tradução e a geração de textos a partir de um banco de dados, etc. Podemos citar duas classes de métodos para a execução deste tipo de tarefas, que são os métodos simbólicos e os métodos estatísticos.

Nos final dos anos 50 e 60, existiam excelentes métodos estatísticos, que foram desenvolvidos durante a segunda guerra mundial, para a solução de problemas Linguísticos (SHANNON; WEAVER, 1948). Porém, no ano de 1957, Chomsky publicou o trabalho intitulado de *"Syntactic Structures"* onde descreve a teoria da gramática gerativa, que é uma teoria que considera a gramática como um conjunto de regras. Essa abordagem através de um conjunto de regras, ao invés de um modelo matemático, entra em conflito com os trabalhos anteriores, criando duas comunidades no campo da Linguística. Como reflexo dessas duas comunidades, a área de NLP que crescia em paralelo, também foi dividida em duas áreas. A primeira dessas áreas que fazia uso de métodos baseados em regras (simbólica) e a segunda que fazia o uso de métodos quantitativos (estatísticas).

Nesta seção será apresentado um exemplo de método simbólico e de um método estatístico. Destaca-se que essa descrição apresenta como objetivo, apenas diferenciar ambas as classes de métodos, através de seus requisitos e forma de execução. Destaca-se ainda que os métodos apresentados nesta seção não são utilizados na análise de sentimentos, sendo que os métodos específicos para essa identificação serão descritos no Capítulo 3.

2.2.1 Método Simbólico

O método simbólico ou racionalista está baseado no campo da Linguística e faz o uso da manipulação dos símbolos, significados e das regras de um texto. Um exemplo simples de um método simbólico é o método de Brill (BRILL, 1992). Por exemplo, no método de Brill a frase "João pintou a casa de branco", será separada em palavras que serão classificadas através de um dicionário pré-definido, como:

Palavra	João	pintou	a	casa	de	branco
Classificação:		Verbo	Artigo	Substantivo	Preposição	Adjetivo

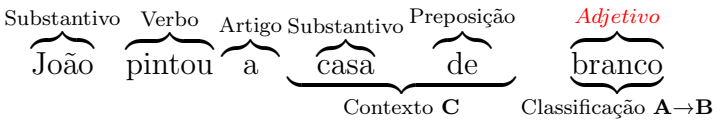
Observa-se que algumas palavras não foram identificadas, como "João", ou classi-

ficadas de forma incorreta, como “branco”. Desta forma, o método de Brill utiliza-se de outras duas regras para a classificação. A primeira regra classifica todas as palavras desconhecidas que iniciam com uma letra em maiúscula como substantivos, por exemplo, a palavra “João”. Já a segunda regra, atribui para a palavra desconhecida a mesma classificação de outras palavras que terminam com as mesmas três letras. Por exemplo, supondo que a palavra “pintou” não fosse encontrada no dicionário, essa seria associada a outras palavras terminadas com o sufixo “tou”, ou seja, essa seria classificada como verbo.

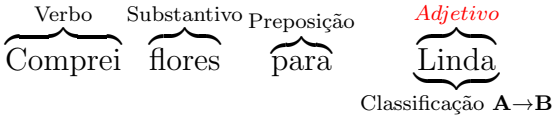
Palavra	João	pintou	a	casa	de	branco
Classificação:	**Substantivo**	Verbo	Artigo	Substantivo	Preposição	Adjetivo

Após essa classificação inicial, o método executa o seguinte conjunto de regras, ou ainda, regras derivadas dessas:

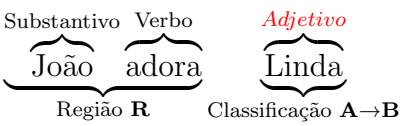
- Se uma palavra tem a classificação **A** e está no contexto **C** então a sua classificação deverá ser mudada para **B**. Por exemplo, se uma palavra **A** (branco no exemplo) é um adjetivo e uma das duas palavras anteriores é uma preposição (“de” no contexto **C**), mude a sua classificação para um substantivo (classificação **B**).



- Se uma palavra tem a classificação **A** e tem uma propriedade **P** então a sua classificação deverá ser alterada para **B**. Por exemplo, se uma palavra **A** (“Linda”) foi classificada como um adjetivo e é iniciada com uma letra maiúscula (propriedade **P**), sua classificação deverá ser alterada para substantivo (classificação **B**).



- Se uma palavra tem a classificação **A** e uma palavra com a propriedade **P** está na região **R**, sua classificação deverá ser **B**. Por exemplo, se uma das duas palavras anteriores à palavra “Linda” (“João adora”na região **R**) iniciam com letra maiúscula (propriedade **P**), sua classificação deverá ser alterada para substantivo (classificação **B**).



2.2.2 Método Estatístico

Um método estatístico utiliza-se de uma grande quantidade de texto, procurando por padrões e associações a modelos, sendo que esses padrões podem ou não estar relacionados com regras sintáticas ou semânticas.

Os métodos estatísticos baseia-se na utilização de um sistema de aprendizado supervisionado, ou seja, a classificação é feita a partir de um conjunto de dados já classificado, que é chamado de *training set*. Um exemplo de método estatístico é a utilização de Modelos de Markov com a aplicação do algoritmo de Viterbi (MANNING; SCHüTZE, 1999).

Em um Modelo de Markov, a classificação da frase "João comprou um carro" é feita a partir de um *training set* que pode, por exemplo, ser composto por textos retirados de *web-sites*, sendo que as palavras destes textos já devem estar classificadas. A partir deste *training set*, as palavras "João", "comprou" e "carro" seriam classificadas como substantivo, verbo e substantivo, respectivamente. Já a palavra "um" apresenta uma ambiguidade uma vez que pode ser classificada como um artigo (ART), ou um substantivo (SM) ou um pronome (PRO). A Figura 2.1 ilustra o conjunto de possibilidades criadas pelo classificador para a classificação completa da frase.

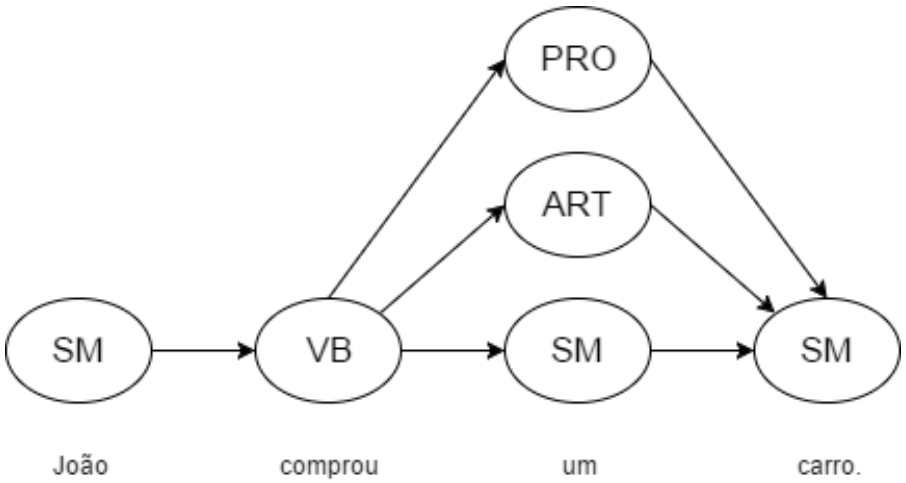


Figura 2.1: Caminhos possíveis de classificação

A idéia central da utilização de Modelos de Markov é escolher, entre os caminhos possíveis (Figura 2.1), o caminho de maior probabilidade. Para tanto, se faz necessário calcular a probabilidade de todos os caminhos através de um Modelo de Markov. Após, utiliza-se o Algoritmo de Viterbi para definir qual o caminho com maior probabilidade (MANNING; SCHüTZE, 1999).

O Modelo de Markov irá utilizar-se do *training set* para inferir a classificação da palavra "um". Por exemplo, considerando-se um *training set* hipotético com as seguintes características: 10000 substantivos aonde 150 são a palavra "um"; 10 são

a palavra “João”; 50 são a palavra “carro”; 20000 artigos aonde 500 são a palavra “um”; 12000 verbos aonde 50 são a palavra “comprou”; 15000 pronomes aonde 50 são a palavra “um”. Neste caso, a probabilidade da palavra “um” ser um substantivo é dada pela Equação 2.1, uma vez que no *training set* temos 150 instâncias da palavra “um” classificadas como substantivo e um total de 10000 substantivos. Ou seja, a probabilidade de “um” ser um substantivo é de 0,015. A Equação 2.1 também é aplicada para as demais possíveis classes da palavra “um”, neste caso, pronome ou artigo. Por exemplo, a probabilidade da palavra “um” ser um pronome seria 0,0033 e a probabilidade da palavra “um” ser um artigo seria 0,025. Esse cálculo de probabilidade é realizado para todas as palavras da frase que está sendo classificada. Na Tabela 2.1 tem-se os resultados obtidos para todas as palavras da frame “João comprou um carro”.

$$\begin{aligned} P(palavra|classe) &= \frac{C(classe, palavra)}{C(classe)} \\ P(um|SM) &= \frac{C(SM, um)}{C(SM)} = \frac{150}{10000} = 0,015. \end{aligned} \tag{2.1}$$

Desta forma tem-se que

	João	comprou	um	carro
Substantivo	0.001	0	0.015	0.005
Verbo	0	0.0042	0	0
Artigo	0	0	0.025	0
Pronome	0	0	0.0033	0

Tabela 2.1: Tabela de Probabilidades de Associação

Além da probabilidade de associação a uma determinada classe, é calculada a probabilidade de transição de uma classe para a outra. Neste caso, o *training set* hipotético apresenta as seguintes características:

- De 20000 frases, 2500 iniciam com um substantivo, 5000 iniciam com um verbo, 5000 iniciam com um artigo e 5000 iniciam com um pronome.
- De 10000 substantivos, os 10000 são seguidos por verbos.
- De 12000 verbos, 3000 são seguidos por um substantivo, 2000 são seguidos por um outro verbo, 5000 são seguidos por um artigo e 2000 são seguidos por um pronome.
- De 20000 artigos, os 20000 são seguidos por um substantivo.
- De 15000 pronomes, 10000 são seguidos por um substantivo e 5000 são seguidos por um verbo.

Neste caso, a probabilidade de transição de um verbo para um substantivo é

dada pela Equação 2.2, uma vez que no *training set* tem-se 12000 verbos, os quais 3000 são seguidos por um substantivo.

$$\begin{aligned} P(transicao|classe) &= \frac{C(classe, transicao)}{C(classe)} \\ P(SM|VB) &= \frac{C(VB, SM)}{C(VB)} = \frac{3000}{12000} = 0,25 \end{aligned} \tag{2.2}$$

Da mesma forma, a probabilidade de transição é cálculada para todas as demais classes. Por exemplo, a probabilidade de transição de um verbo para outro verbo é 0,17, de um verbo para um artigo é 0,42 e de um verbo para um pronome é 0,17. Também, a Equação 2.2 é utilizada também para o cálculo da probabilidade da frase iníciar com determinada classe. A Tabela 2.2 tem-se a probabilidade de transição para todas as classes do *training set* de exemplo.

	Substantivo	Verbo	Artigo	Pronome
Início	0.125	0.25	0.25	0.25
Substantivo	0.0	1.0	0.0	0.0
Verbo	0.25	0.17	0.42	0.17
Artigo	1.0	0.0	0.0	0.0
Pronome	0.67	0.33	0.0	0.0

Tabela 2.2: Tabela de Probabilidade de Transição

A partir das probabilidades calculadas através do Modelo de Markov, é utilizado o algoritmo de Viterbi para determinar o caminho mais provável. O caminho mais provável é obtido através da Equação 2.3, sendo que essa é aplicada a todas as palavras da frase. Na Equação 2.3 os termos v_t, v_{t-1}, a_{ij} e $b_j(o_t)$ correspondem, respectivamente, o caminho mais provável atual, o caminho mais provável anterior, a probabilidade de transição e a probabilidade de associação. Portanto a palavra "João", v_{t-1} é representada pelo valor 1, visto que essa é a primeira palavra da frase: a_{ij} é a probabilidade de transição entre "Início" e um substantivo (Tabela 2.2) e $b_j(o_t)$ é a probabilidade de associação da palavra João com substantivo (Tabela 2.1). Desta forma tem-se que v_t para a palavra João é:

$$v_t(j) = v_{t-1} a_{ij} b_j(o_t) \tag{2.3}$$

$$v_t(j) = 1 * 0,125 * 0,001 = 0,000125. \tag{2.4}$$

Já para a palavra “comprou” tem-se:

$$v_t(j) = 0,000125 * 1 * 0,0042 = 0,000000525. \tag{2.5}$$

Onde os valores 1 e 0,0042 são as probabilidades de transição (Tabela 2.2) e associação (Tabela 2.1) e 0,000125 é o caminho mais provável anterior (Equação 2.3). Após efetuar o cálculo de todos os caminhos, é escolhido o caminho que tem maior probabilidade, sendo que neste caso a palavra “um” é classificada como artigo.

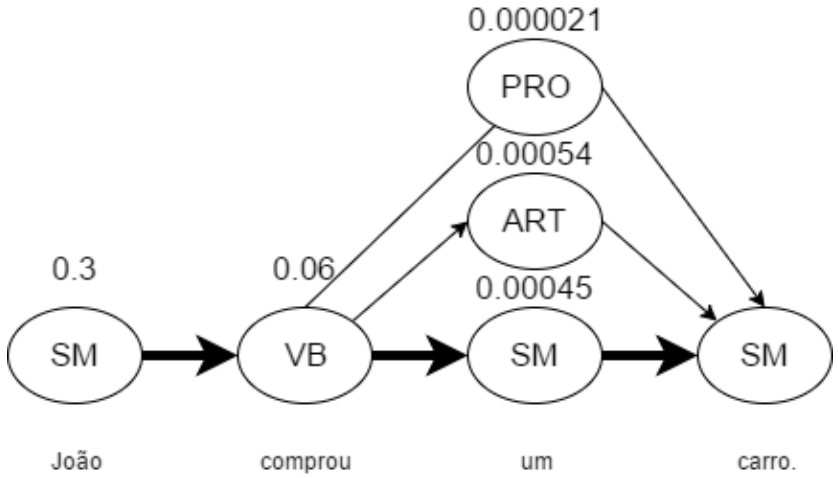


Figura 2.2: Caminhos definidos para a classificação pelo Algoritmo de Viterbi

3 MÉTODOS ESTATÍSTICOS E SIMBÓLICOS APLICADOS NA ANÁLISE DE SENTIMENTOS

Antigamente, para sabermos a opinião de outras pessoas sobre um determinado produto, tinhamos que perguntar diretamente. Com a popularização da Internet e também de redes sociais, milhares de pessoas compartilham para todos as suas opinões sobre produtos, política, serviços e demais assuntos. Porém, muitas vezes essas opiniões acabam por ser esquecidas devido a dificuldade de se analisar uma grande quantidade de textos. De fato, uma das maiores dificuldades reside em como obter a opinião geral sobre determinado produto em uma seção de comentários com mais de 1000 opiniões diferentes. Dentro dessa contexto, a análise de sentimentos, considerada uma tarefa do NLP, tem como função identificar e quantificar esses sentimentos expressos através de textos.

Outros possíveis métodos de análise de sentimentos através de aprendizado de máquina são *Support Vector Machines* (SVM) e *Maximum Entropy* (MaxEnt), os quais possuem performance similar ao Naive Bayes (DOMINGOS; PAZZANI, 1997).

Neste capítulo serão descritos um método estatístico, Naive Bayes e um método simbólico, *Valence Aware Dictionary and sEntiment Reasoner* (VADER), aplicados na análise de sentimentos. Outros possíveis métodos estatísticos para a análise de sentimentos são SVM e MaxEnt, os quais possuem performance similar ao Naive Bayes (PANG; LEE; VAITHYANATHAN, 2002). O método simbólico VADER foi selecionado pois este, segundo seus criadores, apresenta performance superior aos métodos já existentes (HUTTO; GILBERT, 2014).

Ambos métodos selecionados serão descritos para o uso da língua Inglesa, visto que o *Website* analisado (Reddit) possui a maioria de seus comentários em língua Inglesa. Além disso, não foram encontrados métodos que façam uso da língua Portuguesa com similar precisão.

3.1 Naive Bayes

O Naive Bayes é um método estatístico para a classificação o qual podemos aplicar para a análise de sentimento. Esse faz o uso do teorema de Bayes e um *training set* para inferir a classificação de uma frase. Por exemplo, precisamos determinar se a frase "This place is great." demonstra um sentimento negativo ou positivo.

Texto	Categoria
The food was great	Positiva
They are horrible!	Negativa
I love the food here	Positiva
This place is wonderful	Positiva
Forgettable experience	Negativa

Tabela 3.1: *Training Set*

A partir do *training set* hipotético (Tabela 3.1) o método irá calcular a probabilidade da frase "This place is great" ser positiva e também de ser negativa sendo que a partir dessas duas possibilidades, será escolhida a de maior probabilidade.

Para cálcular a probabilidade da frase "This place is great" pertencer a cada categoria é utilizado o teorema de Bayes (MANNING; SCHüTZE, 1999), através da Equação 3.1.

$$P(c|d) = \frac{P(d|c) \times P(c)}{P(d)} \tag{3.1}$$

$$P(Negativa|\,This\ place\ is\ great) = \frac{P(\,This\ place\ is\ great|Negativa) \times P(Negativa)}{This\ place\ is\ great} \tag{3.2}$$

$$P(Positiva|\,This\ place\ is\ great) = \frac{P(\,This\ place\ is\ great|Positiva) \times P(Positiva)}{This\ place\ is\ great} \tag{3.3}$$

Onde o termo P(c|d) é a probabilidade da frase **d** pertencer a classe **c**. Ou seja, a probabilidade de *"This place is great"* ser uma frase positiva ou negativa. O Termo P(d|c) é a probabilidade da classe **c** ser a frase **d**. Ou seja, dentre todas as frases negativas ou positivas, a probabilidade de uma frase ser *"This place is great"*. Já P(c) é a probabilidade da classe **c**. Ou seja, a frequência que frases negativas ou positivas aparecem em nosso *training set*. E, por fim, P(d) é a probabilidade de **d**. Ou seja, a frequência que a frase *"This place is great"* aparece em nosso *training set*. Como ambas as Equações 3.2 e 3.3 terão como divisor P(d), este pode ser

simplificado resultando nas equações 3.4 e 3.5.

$$P(Negativa|\,This\ place\ is\ great) = P(\,This\ place\ is\ great|Negativa) \times P(Negativa) \tag{3.4}$$

$$P(Positiva|\,This\ place\ is\ great) = P(\,This\ place\ is\ great|Positiva) \times P(Positiva) \tag{3.5}$$

Os termos P(Positiva) e P(Negativa) são definidos pela frequência que frases positivas e negativas aparecem no *training set*, sendo determinados através das Equações 3.6 e 3.7. Neste caso, *"The food was great"*, *"I love the food here"*, *"This place is wonderful"* são frases positivas e as demais frases "*They are horrible!*" e "*Forgettable experience*" são negativas.

$$P(Positiva) = \frac{3}{5} = 0,6 \tag{3.6}$$

$$P(Negativa) = \frac{2}{5} = 0,4 \tag{3.7}$$

Uma vez que a frase *"This place is great"* não existe por completo no *training set*, tem-se que o termo P(d|c) da Equação 3.4 e 3.5 é igual a zero (0) (impossibilitando o cálculo de probabilidade para essa frase). Neste caso, se faz o uso do *Naive Bayes*, o qual passa a considerar as palavras ao invés de frases completas, isso elimina o problema com frases que não se encontram no *training sets*. Neste caso, considera-se somente a fraquência que cada palavra aparece em uma frase positiva e em uma negativa. Portanto para o *Naive Bayes* o termo $P(\,This\ place\ is\ great|Positiva)$ visto na Equação 3.5 é dado pela Equação 3.8.

$$\begin{aligned} P(\,This\ place\ is\ great|Positiva) = P(\,This|Positiva) \times P(place|Positiva) \\ \times P(is|Positiva) \times P(great|Positiva) \end{aligned} \tag{3.8}$$

A partir da Equação 3.8 é necessário calcular os termos P(*This*|Positiva), P(*place*|Positiva), P(*is*|Positiva), P(*great*|Positiva). O termo P(*This*|Positiva) é calculado pela razão entre a quantidade de vezes que a palavra *This* foi classificada como positiva em nosso *training set*, e o total de palavras classificadas como positiva (Equação 3.9).

$$P(This|Positiva) = \frac{1}{13} \tag{3.9}$$

Da mesma forma, a Equação 3.9 deve ser aplicada para as demais palavras da frase *"This place is great"*, obtendo-se os valores apresentados na Tabela 3.2.

Uma vez que algumas palavras não encontram-se no *training set* para determinadas situações, elas acabam zerando o resultado final da multiplicação das probabilidades de cada palavra (Equação 3.8). De modo, a evitar que uma única pa-

Palavra	Positiva	Negativa
This	$\frac{1}{13}$	$\frac{0}{5}$
place	$\frac{1}{13}$	$\frac{0}{5}$
is	$\frac{1}{13}$	$\frac{0}{5}$
great	$\frac{1}{13}$	$\frac{0}{5}$

Tabela 3.2: Tabela de Palavras e Probabilidades.

lavra invalide uma frase é utilizado *Laplace smoothing* (MANNING; RAGHAVAN; SCHüTZE, 2008). Neste, é somado 1 a cada palavra e ao total de palavras, são somadas as quantidades de palavras distintas do *training set* (16). Aplicando o *Laplace smoothing* para a Tabela 3.2 é obtida a Tabela 3.3.

Palavra	Positiva	Negativa
This	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$
place	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$
is	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$
great	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$

Tabela 3.3: Tabela de Probabilidades - *Laplace smoothing*.

Utilizando as probabilidades obtidas na Tabela 3.3 na Equação 3.8 tem-se:

$$P(Positiva|\,This\ place\ is\ great) = \frac{1+1}{13+16} \times \frac{1+1}{13+16} \times \frac{1+1}{13+16} \times \frac{1+1}{13+16} = 0,000023. \tag{3.10}$$

Uma vez que o termo P(Positiva | *This place is great*) encontra-se definido através da Equação 3.10, pode-se utilizar a Equação 3.4, onde o termo P(Positiva) é igual a 0,6 definido através da Equação 3.6. Neste caso, tem-se que a probabilidade da frase *"This place is great"* ser classificada como positiva é dada através da Equação 3.11.

$$P(Positiva|\,This\ place\ is\ great) = 0,000023 \times 0,6 = 0,0000138. \tag{3.11}$$

Efetuando o mesmo processo para a probabilidade da frase ser negativa, tem-se:

$$P(Negativa|\,This\ place\ is\ great) = 0,0000049 \times 0,4 = 0,00000196. \tag{3.12}$$

Portanto, tem-se que a frase "This place is great" é positiva, uma vez que a

probabilidade de ser positiva (0,0000138) é maior que a probabilidade dessa frase ser negativa (0,00000196).

3.2 *VADER*

O VADER é um dicionário e classificador de sentimentos que se baseia em regras, portanto, um método de classificação simbólico. Esse foi desenvolvido especificadamente para funcionar em redes sociais onde-se tem um contexto vago, pouca quantidade de texto, gírias e emoticons (HUTTO; GILBERT, 2014).

A classificação do sentimento é feita através da separação da frase em palavras e para cada palavra da frase é atribuída uma pontuação de intensidade em uma escala de -4 até +4. Como por exemplo, a palavra *great* tem a intensidade de 3.1 e *horrible* -2.5. Essa pontuação é obtida através de um dicionário que é construído utilizando o método de *"wisdom of the crowd"* onde um grupo de pessoas atribuiu os valores de intensidade para cada palavra. Por exemplo, a frase *"This place is great"* seria classificada com:

Palavra	*This*	*place*	*is*	*great*
Intensidade:				3,1

As palavras *"This"*, *"place"* e *"is"* são desconsideradas uma vez que não existem no dicionário e não expressam sentimentos. Após, ele faz uso do seguinte conjunto de regras para inferir a intensidade do sentimento:

- A primeira regra consistem em verificar quando uma palavra com pontuação atribuída (uma palavra que expressa sentimentos) é escrita em letras maiúsculas. Neste caso, é aumentada a magnitude da intensidade do sentimento sem modificar a orientação semântica. Para isso, é somado 0,733 a intensidade do sentimento caso este tenha intensidade positiva ou subtraído 0,733 caso este tenha intensidade negativa.

Palavra	*This*	*place*	*is*	*GREAT*
Intensidade:				3,1 →3,833

- A segunda regra verifica se alguma das três palavras anteriores é um advérbio intensificador. Neste caso, estes impactam a intensidade do sentimento aumentando ou diminuindo 0,293 conforme o advérbio.

Palavra	*This*	*place*	*is*	*incredibly*	*great*
Intensidade:				Advérbio	3,1 →3,393

- A terceira regra verifica se a frase contém a palavra *"but"*. Caso encontrada, essa palavra indica uma troca do sentimento da frase, uma vez que o texto

Palavra	*This*	*place*	*is*	*somewhat*	*great*
Intensidade:				Advérbio	3,1 →2,807

seguinte a palavra *"but"* expressa um sentimento mais dominante. Neste caso, o método multiplica a intensidade dos sentimentos expressos até a palavra *"but"* por 0,5 e os sentimentos expressos após a palavra *"but"* por 1,5.

Palavra	*Great*	*place*	***but***	*today*	*the*	*food*	*was*	*horrible*
Intensidade:	3,1 →1,55							-2,5 →-3,75

- A quarta regra verifica se a frase possui pontos de exclamação (!). Este tipo de pontuação aumenta a magnitude da intensidade sem modificar a orientação semântica. 0,292 a cada ponto de exclamação considerando um máximo de 4 pontos de exclamação.

Palavra	*This*	*place*	*is*	*great!*
Intensidade:				3,1 →3,392

- Por fim, a quinta regra examina as três palavras anteriores, procurando a existência de uma negação que inverte a polaridade de um texto. Quando é encontrada uma negação na frase, a intensidade de cada palavra de sentimento é multiplicada por -0,74.

Palavra	*This*	*place*	*wasn't*	*great*
Intensidade:			Negação	3,1 →2,294

Após esse cálculo de pontuação de intensidade, é feita a normalização dessa pontuação através da Equação 3.13.

$$\text{Pontuação Normalizada} = \frac{\text{Pontuação}}{\sqrt{\text{Pontuação}^2 + 15}} \tag{3.13}$$

$$\text{Pontuação Normalizada} = \frac{3{,}1}{\sqrt{3{,}1^2 + 15}} = 0,6249 \tag{3.14}$$

Neste caso para a frase *"This place is great"* (aonde somente a palavra *great* possui pontuação e essa tem o valor de 3,1) será atribuida pontuação final de 0,6249. Caso essa pontuação fosse menor que -1 ou maior que 1, essa seria limitada aos valores de -1 ou 1 respectivamente. Para o VADER, são consideradas frases negativas aquelas com pontuação de -1 até -0,5, frases neutras aquelas com pontuação de -0,5 até 0,5 e frases positivas aquelas com pontuação 0,5 até 1. Portanto a frase *"This place is great"* com pontuação 0,6249 seria classificada como positiva.

3.2.1 Método Selecionado

A partir dos métodos citados anteriormente, foi selecionado o método que demonstra a maior performance reportada por outros artigos. Por estes serem realizados em outras mídias sociais, também foi considerada a similaridade desta mídia social para o Reddit.

A avaliação de performance na classificação de textos é medida através do F_1 Score (ou também *F-Score* ou *F-Measure*). Este é realizado utilizando o valor de revocação, ou seja, a porcentagem do texto classificada corretamente, representado pela Equação 3.15 e também o valor de precisão, representado pela Equação 3.16 aplicados a Equação 3.17.

$$\text{Revocação} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} \tag{3.15}$$

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}} \tag{3.16}$$

$$F_1\text{Score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.17}$$

A utilização do Naive Bayes para classificação de avaliações de filmes enviados por usuários do website IMDB[1] segundo Pang (PANG; LEE; VAITHYANATHAN, 2002), atinge em média a precisão de 80% para análise de sentimentos em determinados casos. Ainda que passando o valor de 50% de precisão apresentado por uma escolha aleatória, este valor está distante do apresentado pelo mesmo método quando utilizado para a classificação de tópicos (NIGAM et al., 2000).

Quando da utilização do método para a análise de *Tweets*, um cenário com maior similaridade ao Reddit, segundo Pålsson e Szerszen (PåLSSON; SZERSZEN, 2016), o método apresenta F_1 Score de 58,2% na análise de *Tweets* e 58,3% na análise de *Tweets* pré processados aonde foram substituídos os *emoticons* por palavras similares como *happy, sad ou neutral*. Em contraste, a utilização do VADER apresenta respectivamente o F_1 Score de 72,3% e 74,9% nos mesmos cenários.

Também, segundo Hutto (HUTTO; GILBERT, 2014), o único cenário no qual o Naive Bayes apresenta melhor precisão utilizando três classes (positivo, neutro e negativo) é o de análise de sentimentos em avaliações de filmes, apresentando F_1 Score de 75% utilizando um *training set* especializado para avaliações de filmes contra 61% apresentado pelo VADER. Nos outros cenários avaliados, *Tweets*, avaliações de produtos da Amazon e editoriais do New York Times, o VADER apresentou melhor performance com o F_1 Score de 96%, 63% e 55% respectivamente, contra 56%, 49%, 44% apresentados pelo Naive Bayes.

[1] http://www.imdb.com

Por não necessitar de um *training set* adequado especificadamente ao tema que está sendo analisado e por consequência apresentar maior precisão em diversos cenários, foi selecionado o VADER para a utilização na análise de sentimentos na rede social Reddit, visto que essa abrange inúmeros assuntos.

4 FRAMEWORKS

Para implementação de um *software* que efetue a análise de sentimentos e auxilie a categorizar extrair as informações de modo a análisar padrões de comportamento, se destacam os seguintes *frameworks*:

- *Stanford's Core NLP Suite.*
- *Natural Language Toolkit.*
- *Apache OpenNLP.*
- *Spacy.*

Essa seção visa analisar os métodos disponíveis nestes para a análise de sentimentos a fim de encontrar *frameworks* que possibilitem a utilização dos métodos apresentados no Capítulo 3 e também possibilitam atender outras necessidades do processamento de linguagem natural.

4.1 Natural Language Toolkit

O *Natural Language Toolkit* (NLTK) é um *Framework* para Python criado em 2001 na Universidade de Pensilvânia. Ele contém mais de 50 dicionários e modelos já treinados incluindo:

- *Sentiment Polarity Dataset Version 2.0* - Conjunto de dados já classificados que contém mais de 1000 filmes avaliados de forma positiva e 1000 filmes avaliados de forma negativa.
- *SentiWordNet* - Provém um dicionário com as palavras extraídas do WordNet já classificadas em positividade, negatividade e objetividade.
- *VADER Sentiment Lexicon* - Dicionário especificamente ajustado para análise de sentimentos expressos em mídias sociais.

4.1.1 Análise de Sentimentos

Para a análise de sentimentos, o NLTK já possui implementado os três classificadores citados anteriormente, *Naive Bayes* e VADER.

Podemos utilizar o classificador Naive Bayes a partir da classe **nltk.classify.naivebayes.NaiveBayesClassifier** através dos seguintes métodos:

- *classify(featureset)* - Classifica a partir de um conjunto de atributos.
- *most_informative_features(n=100)* - A partir de um classificador treinado, retorna os atributos mais relevantes.
- *train(trainingset)* - Treina um classificador a partir de um *training set*.

Para utilização do VADER é utilizada a classe *SentimentIntensityAnalyzer* do módulo *vaderSentiment* através do método *polarity_scores*. Este método recebe uma frase e retorna um objeto contendo a intensidade positiva, neutra e negativa da frase.

O *framework* também contém um pacote contendo classes úteis para a análise de sentimentos chamado de *nltk.sentiment*. Nesse pacote temos os seguintes módulos:

- Classe *nltk.sentiment.sentiment_analyzer.SentimentAnalyzer* - Ferramentas para facilitar e implementar análise de sentimentos, especialmente para demonstrações e ensino.
- Módulo *nltk.sentiment.util* - Contém diversas classes de demonstrações e utilitários como conversão de *json* para *csv*.

4.2 Stanford CoreNLP

O Stanford CoreNLP é um conjunto de ferramentas escrito em Java para processamento de linguagem natural. Dentre essas ferramentas, estão incluídos: *Part-of-Speech Tagging* ou classificação gramatical, reconhecimento de entidade e análise de sentimentos. Também possui suporte a diversas linguas além do inglês, como: árabe, chinês, francês, alemão e espanhol.

4.2.1 Análise de Sentimentos

A análise de sentimentos do Stanford CoreNLP é realizada através de um novo modelo de rede neural construído em cima de estruturas gramaticais chamado de *Recursive Neural Tensor Networks* (RNTN). Seu modelo é treinado a partir do *Sentiment Treebank*, um banco de dados que possui 215.154 orações distribuidas em 11.855 árvores de frases com sentimentos já classificados.

A sua utilização pode ser feita de diversas formas, como linha de comando, através de um servidor *web* e através de sua API java:

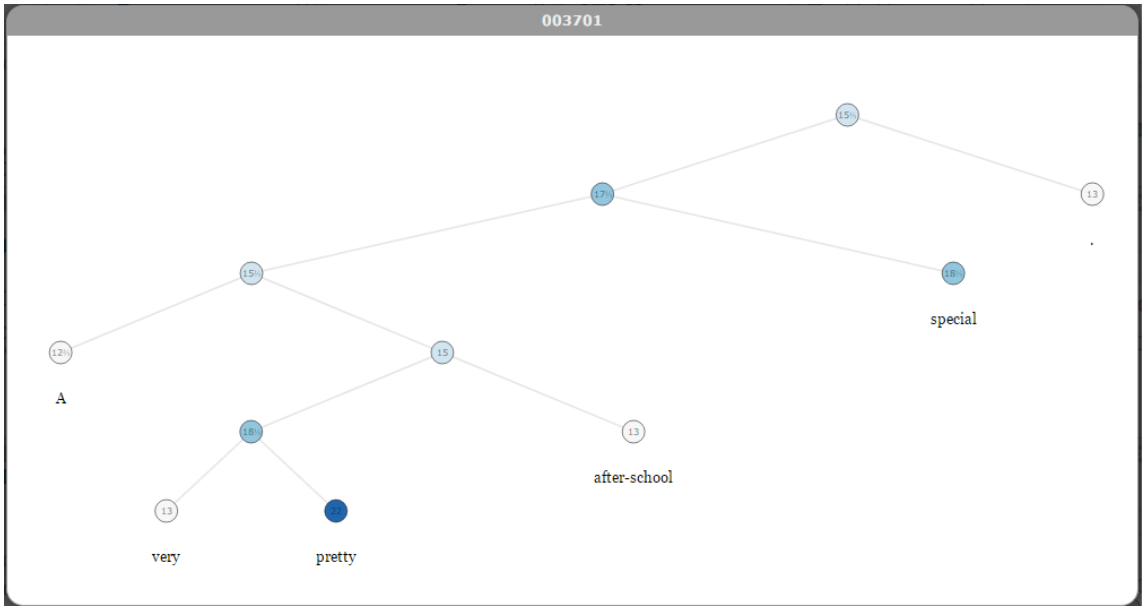


Figura 4.1: Frase já classificada disponível no Sentiment Treebank

```java
public static void main(String[] args) throws IOException {
        String text = "This World is an amazing place";
        Properties props = new Properties();
        props.setProperty("annotators", "sentiment");
        StanfordCoreNLP pipeline = new StanfordCoreNLP(props);

        Annotation annotation = pipeline.process(text);
        List<CoreMap> sentences = annotation.get(CoreAnnotations.SentencesAnnotation.class);
        for (CoreMap sentence : sentences) {
            String sentiment = sentence.get(SentimentCoreAnnotations.SentimentClass.class);
            System.out.println(sentiment + "\t" + sentence);
        }
}
```

Figura 4.2: Exemplo de implementação

Como resultado, o console java irá imprimir que a frase é muito positiva ou *Very positive*.

5 REDE SOCIAL REDDIT

O *website* Reddit teve seu início em 2005 como um agregador de conteúdo e atualmente é o vigésimo terceiro *website* mais acessado na internet e o sétimo mais acessado nos Estados Unidos da América (Alexa, 2017). Seus usuários podem enviar links para conteúdos externos ao Reddit ou também mensagens de texto. A partir desse conteúdo enviado, seja ele uma mensagem de texto no próprio Reddit quanto um link a um *website* externo, seus usuários podem votar para cima (*upvote*) ou para baixo *downvote*, influenciando a sua posição no *website*. Esse algoritmo de ordenação de conteúdo é fechado portanto não está disponível para consulta. Além de votar no conteúdo, seus usuários podem enviar comentários como forma de expressar sua opinião.

Esse conteúdo é distribuido em *subreddits* que funcionam como comunidades que abordam certos assuntos. Os usuários podem se inscrever nesses *subreddits* para que seu conteúdo apareça na página inicial. Dentre os *subreddits* mais notáveis se encontram:

- */r/AskReddit* - Local para fazer perguntas gerais para outros usuários. Atualmente com 16.941.544 de inscritos.
- */r/worldnews* - Notícias do mundo. Atualmente com 16.570.606 de inscritos.
- */r/IAmA* - IAmA é um estilização de 'I am a' ou 'Eu sou um'. Local aonde os usuários podem fazer perguntas e respostas ao criador do tópico que se identifica por algo notável, como uma profissão ou algum feito. Atualmente com 16.941.544 de inscritos.

Dentre esses *subreddits* podemos destacar alguns dos tópicos mais acessados no ano de 2016:

- */r/IAmA - We're NASA scientists & exoplanet experts. Ask us anything about today's announcement of seven Earth-size planets orbiting TRAPPIST-1!* - Tópico de perguntas e respostas com cientistas da NASA após a descoberta dos planetas que orbitavam a estrela TRAPPIST-1.

- */r/IAmA - I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. Ask Me Anything.* - Tópico de perguntas e respostas com Bill Gates.
- */r/worldnews - Fidel Castro is dead at 90.* - Link para anúncio da morte de Fidel Castro.
- */r/AskReddit - [Serious]South Koreans of Reddit, how did they teach you about the existence of North Korea in School when you were young?serious replies only* - Tópico perguntando para os usuários sul coreanos como que foi ensinado para eles sobre a existência da Coreia do Norte.

5.1 API

O *website* possui uma API *open source* localizada em https://github.com/reddit/. Sua documentação é gerada de forma automática a partir do código fonte e podemos encontrar ela em: https://www.reddit.com/dev/api/.

6 EXTRAÇÃO DE DADOS

Para a extração dos dados para a análise de sentimentos foi criado um *crawler* ou robô de navegação. Esse robô tem como objetivo a navegação automática no conteúdo web do Reddit, extraindo os dados referentes a tópicos e a comentários e persistindo esses em um banco de dados.

6.1 *Crawler*

O *Crawler* foi escrito na linguagem Java por se tratar de uma linguagem com uma grande quantidade de bibliotecas disponíveis e também sua facilidade de implementação. A Figura 6.1 representa a arquitetura utilizada para desenvolvimento deste software.

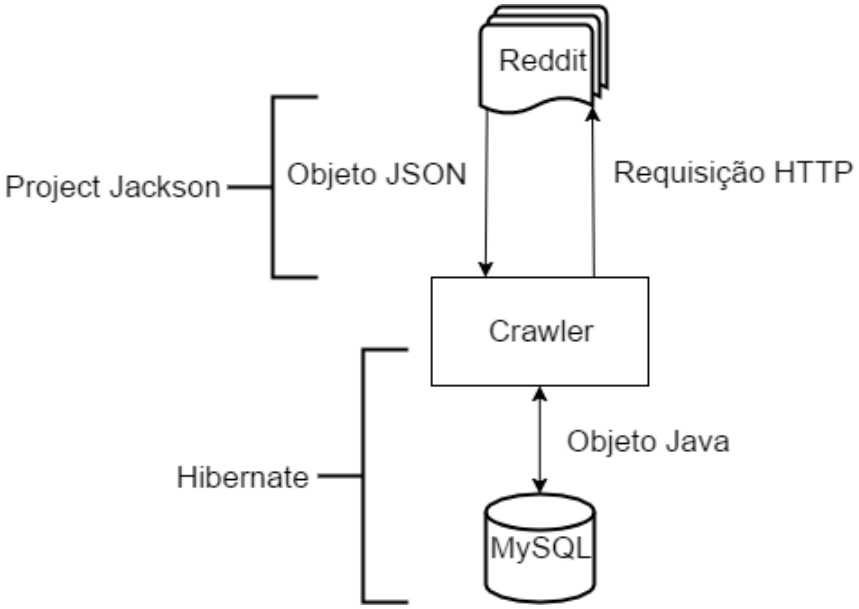


Figura 6.1: Arquitetura do *Crawler*

A partir de um *link* para um tópico, o software tem como tarefa, a extração e pesquisa de dados relacionados com o tópico em questão. Isso se faz da seguinte forma, primeiramente, é enviada uma requisição para o *link* utilizando o sufixo

".json", de forma demonstrada na API, a partir dessa requisição, o *website* retorna um objeto *JavaScript Object Notation* (JSON).

Como o JSON possui 68 campos que assim como seus tipos de dados, não se encontram em nenhuma documentação, foi utilizado o *website* jsonschema2pojo[1] para mapear o JSON retornado em um objeto Java. Esse *website* tem como objetivo a conversão de um esquema JSON ou o próprio JSON para um *Plain Old Java Objects* (POJO) ou Os Singelos Clássicos Objetos Java, permitindo o *download* da classe para utilização, já com as anotações "@JsonProperty" utilizadas na biblioteca Jackson. Este objeto disponibilizado a partir do *website* foi renomeado para *RedditPost* e adicionado ao código fonte da aplicação.

A anotação "@JsonProperty" é relativa ao mapeamento do objeto Java com relação ao JSON e nos permite com intermédio da classe *Object Mapper* instanciar o objeto *RedditPost* a partir de um objeto *String* em formato JSON.

```
RedditPost post =
    objectMapper.readValue(iteratorPost.get("data").toString(),
RedditPost.class);
```

Utilizando o método *readValue* que tem como retorno um *Object*, é informado um objeto *String*, neste caso *"iteratorPost.get("data").toString()"* e uma classe mapeada, neste caso *RedditPost*.

Porém, para o correto funcionamento, deve ser feita a seguinte mudança: O campo *edited* representando se foi editado o comentário, retornado no JSON apresenta um tipo de dado ambíguo aonde que caso o comentário não tenha sido editado, ele apresenta o valor *booleano* de *false*, porém, ao ter sido editado, ele apresenta seu valor em um formato decimal. Este e demais objetos que apresentavam o tipo de dado de forma ambígua foram transformados em objetos *String*.

A partir deste objeto Java, foi utilizado o *framework* Hibernate para a criação do banco de dados, assim como persistência destes. Através da anotação "@Entity", adicionada também na classe RedditPost. O Hibernate mapeia essa entidade junto ao banco de dados, neste caso, MySQL.

Para criação das tabelas do banco de dados, foi utilizada a propriedade *hibernate.hbm2ddl.auto* do Hibernate. Essa propriedade quando instânciada uma nova sessão do *framework* no Java executa as seguintes ações dependendo de seus valores informados:

- *validate*: Não efetua mudanças no banco de dados, somente valida.
- *update*: Atualiza o esquema do banco de dados conforme os objetos mapeados na camada Java.

[1] http://www.jsonschema2pojo.org/

- *create*: Cria o esquema contendo tabelas e campos a partir dos objetos Java, destruindo dados anteriores.
- *create-drop*: Cria o esquema da mesma que o *create*, porém, ao termino da sessão, remove o esquema criado.

No primeiro momento, a propriedade obteve o valor *create*, para fins de criação e validação do esquema criado e após isso, foi informado *update* como seu valor para tornar reflexo as alterações feitas na camada Java.

Portanto, a execução do *Crawler* funciona da seguinte forma, é enviada uma requisição para o *website* através da URL do tópico em questão com o sufixo "*.json*" no final. O *website* retorna um *JSON* com os dados referentes ao tópico solicitado e aos comentários deste tópico. Este objeto JSON é convertido em um POJO através da biblioteca Jackson e persistida no banco de dados através do *framework* Hibernate. Como a API do Reddit possui uma restrição do número de comentários disponibilizados, são efetuadas novas requisições para a seção *"more"* disponível no JSON de retorno.

6.2 Tópicos Selecionados

Para análise de sentimentos e comparação dos resultados obtidos, foram selecionados 25 tópicos controversos, selecionados com o objetivo de encontrar padrões de opinião entre seus comentários, distribuídos da seguinte forma:

Tópicos relacionados com o cenário político nacional:

- *Brazil Seeks To Copy U.S. Gun Culture "to allow embattled citizens the right to defend themselves from criminals"* - Disponível em: `https://www.reddit.com/r/worldnews/comments/36ny58/brazil_blogger_known_for_reporting_on_corruption/`.
- *Brazil descends into chaos as Olympics looms* - Disponível em: `https://www.reddit.com/r/worldnews/comments/4bqcc3/brazil_descends_into_chaos_as_olympics_looms/`.
- *Plane carrying Brazil Supreme Court judge crashes into sea* - Disponível em: `https://www.reddit.com/r/worldnews/comments/5oyz3b/plane_carrying_brazil_supreme_court_judge_crashes/`.
- *Brazil passes Internet governance Bill: Brazil has made history with the approval of a post-Snowden Bill which sets out principles, rights and guarantees for Internet users.* - Disponível em: `https://www.reddit.com/r/worldnews/comments/21f3as/brazil_passes_internet_governance_bill_brazil_has/`.
- *FIFA generated more than $4 billion in sales from the 2014 World Cup, and is Giving Brazil $100 Million After The Country Spent $15 Billion On The World*

Cup - Disponível em: https://www.reddit.com/r/worldnews/comments/2t65ql/fifa_generated_more_than_4_billion_in_sales_from/.

- *Brazil Seeks To Copy U.S. Gun Culture "to allow embattled citizens the right to defend themselves from criminals"* - Disponível em: https://www.reddit.com/r/worldnews/comments/3skpe7/brazil_seeks_to_copy_us_gun_culture_to_allow/.

Tópicos relacionados sobre política internacional.

- *2.6 terabyte leak of Panamanian shell company data reveals "how a global industry led by major banks, legal firms, and asset management companies secretly manages the estates of politicians, Fifa officials, fraudsters and drug smugglers, celebrities and professional athletes.".* - Disponível em: https://www.reddit.com/r/worldnews/comments/4d75i7/26_terabyte_leak_of_panamanian_shell_company_data/.
- *Fidel Castro is dead at 90.* - Disponível em: https://www.reddit.com/r/worldnews/comments/5exz2e/fidel_castro_is_dead_at_90/.
- *Donald Trump to strip all funding from State Dept team promoting women's rights around the world - Leaked plan comes as First Daughter Ivanka defends her father's record with women* - Disponível em: https://www.reddit.com/r/worldnews/comments/67ivae/donald_trump_to_strip_all_funding_from_state_dept/.
- *Manchester Arena 'explosions': Two loud bangs heard at MEN Arena* - Disponível em: https://www.reddit.com/r/worldnews/comments/6cqdye/manchester_arena_explosions_two_loud_bangs_heard/.
- *Sweden asks the U.S. to explain Trump comment on Sweden* - Disponível em: https://www.reddit.com/r/worldnews/comments/5uzetf/sweden_asks_the_us_to_explain_trump_comment_on/
- *"Canada will welcome you," Trudeau invites refugees as Trump bans them* - Disponível em: https://www.reddit.com/r/worldnews/comments/5qqa51/canada_will_welcome_you_trudeau_invites_refugees/

Tópicos diversos.

- *I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. Ask Me Anything.* - Disponível em: https://www.reddit.com/r/IAmA/comments/5whpqs/im_bill_gates_cochair_of_the_bill_melinda_gates/.
- *Hey, it's Lars from Metallica. AMA* - Disponível em: https://www.reddit.com/r/IAmA/comments/1wl9ic/hey_its_lars_from_metallica_ama/.
- *I'm the CEO of Renault and Nissan and we're making autonomous driving*

vehicles happen by 2020. Ask me anything! - Disponível em: `https://www.reddit.com/r/IAmA/comments/2s7obx/im_the_ceo_of_renault_and_nissan_and_were_making/`.

- *I am Julian Assange founder of WikiLeaks – Ask Me Anything* - Disponível em: `https://www.reddit.com/r/IAmA/comments/5n58sm/i_am_julian_assange_founder_of_wikileaks_ask_me/`.

A partir destes tópicos selecionados, foram extraídos seus dados, como conteúdo, usuário de criação e número de *upvotes*. Somente foram extraídos comentários em resposta ao tópico em questão, comentários em resposta a outros comentários foram desconsiderados pois estes podem não estar relacionados com o tópico em questão, o que torna a sua análise de sentimento inaproveitável. Além de não existirem *softwares* ou *frameworks* preparados para a análise de sentimentos de uma conversa sobre um determinado tópico.

REFERÊNCIAS

Alexa. **Alexa Top 500 Global Sites**. <Disponível em: `http://www.alexa.com/topsites/`>. Acesso em: 27 de Fevereiro de 2017.

Apache OpenNLP. **Apache OpenNLP**. <Disponível em: `https://opennlp.apache.org/`>. Acesso em: 27 de Fevereiro de 2017.

BRILL, E. A Simple Rule-based Part of Speech Tagger. In: THIRD CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 1992. p.152–155. (ANLC '92).

DOMINGOS, P.; PAZZANI, M. J. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. **Machine Learning**, [S.l.], v.29, n.2-3, p.103–130, 1997.

HANCOX, P. J. **A brief history of Natural Language Processing**. <Disponível em: `http://www.cs.bham.ac.uk/~pjh/sem1a5/pt1/pt1_history.html/`>. Acesso em: 02 de Abril de 2017.

HUTTO, C. J.; GILBERT, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: ICWSM. **Anais...** The AAAI Press, 2014.

JACKSON, P.; MOULINIER, I. **Natural Language Processing for Online Applications**: text retrieval, extraction and categorization. [S.l.]: John Benjamins Publishing Company, 2007.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008.

MANNING, C. D.; SCHüTZE, H. **Foundations of Statistical Natural Language Processing**. [S.l.]: MIT Press, 1999.

Natural Language Toolkit. **Natural Language Toolkit**. <Disponível em: `http://www.nltk.org/`>. Acesso em: 27 de Fevereiro de 2017.

NIGAM, K. et al. Text Classification from Labeled and Unlabeled Documents Using EM. **Mach. Learn.**, Hingham, MA, USA, v.39, n.2-3, p.103–134, May 2000.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs Up?: sentiment classification using machine learning techniques. In: ACL-02 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING - VOLUME 10, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2002. p.79–86. (EMNLP '02).

PåLSSON, A.; SZERSZEN, D. **Sentiment Classification in Social Media**: an analysis of methods and the impact of emoticon removal (dissertation). 2016.

SHANNON, C. E.; WEAVER, W. A Mathematical Theory of Communication. **The Bell System Technical Journal**, [S.l.], v.27, p.379–423,623–656, July, October 1948.

Spacy. **Spacy**. <Disponível em: `https://spacy.io/`>. Acesso em: 27 de Fevereiro de 2017.

Stanford CoreNLP. **Stanford CoreNLP**. <Disponível em: `http://stanfordnlp.github.io/CoreNLP/`>. Acesso em: 27 de Fevereiro de 2017.