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*

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 de 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, é possível efetuar 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 no dia 15 de Novembro de 2015.

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 utilizando 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, necessitamos antes entender 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 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 lou-

cutor, se torna extremamente difícil a criação de um software ou hardware 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 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 a linguística 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 principal objetivo 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 identificação de sentimentos 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 de 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

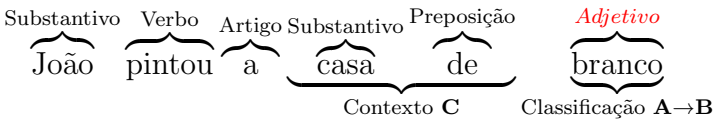
Observa-se que algumas palavras não foram identificadas, como "João", ou classificadas de forma incorreta, como "branco". Desta forma, o método de Brill utiliza-se

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.

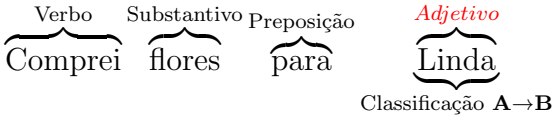
Palavra	João	pintou	a	casa	de	branco
Classificação:	**Substantivo**	Verbo	Artigo	Substantivo	Preposição	Adjetivo

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

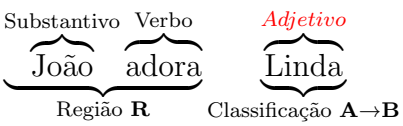
- Se uma palavra tem a classificação **A** e está no contexto **C** então a sua classificação deverá ser mudada para **B**. Por exemplo, se uma palavra **A** (branco no exemplo) é um adjetivo e uma das duas palavras anteriores é uma preposição ("de" no contexto **C**), mude para sua classificação para 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 representa o conjunto de possibilidades que o classificador apresenta para uma classificação completa da frase.

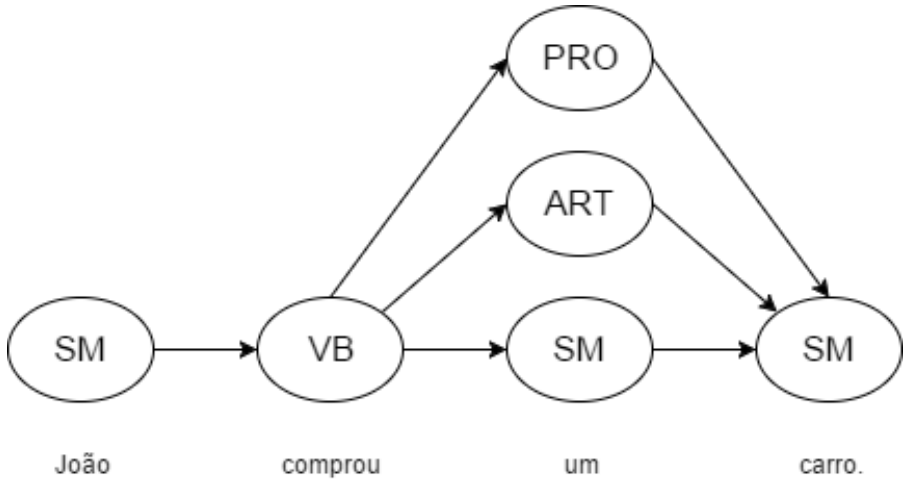


Figura 2.1: Caminhos possíveis de classificação

A idéia central da utilização de Modelos de Markov é escolher, entre esses caminhos (Figura 2.1), pelo 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 caminho com maior probabilidade.

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 ser um substantivo é dada pela Equação 2.1, onde no *training set* temos 150 instâncias da palavra “um” classificadas como substantivo e um total de 10000 substantivos.

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

Desta forma tem-se que a probabilidade de “um” ser um substantivo é de 0,015. A Equação 2.1 também é aplicada para as demais classes, 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”.

	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, 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, 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, onde no *training set* existem 12000 verbos, os quais 3000

são seguidos por um substantivo, desta forma para a transição de um verbo para substantivo tem-se:

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

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 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.

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

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 a probabilidade do caminho atual, resultado do caminho 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, ou seja, não foram calculados os valores de v_t para as palavras anteriores, 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:

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

Já para a palavra “comprou” tem-se:

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

Aonde além das probabilidades de transição e associação respecitvamente retirados das Tabelas 2.2 e 2.1, v_{t-1} é representado pelo cálculo do caminho anterior, ou seja, 0,000125. Ao efetuar o cálculo de todos os caminhos, para determinar qual a classificação correta de uma palavra, é escolhido o caminho que tem maior probabilidade, no caso apresentado, a palavra “um” é classificada como artigo.

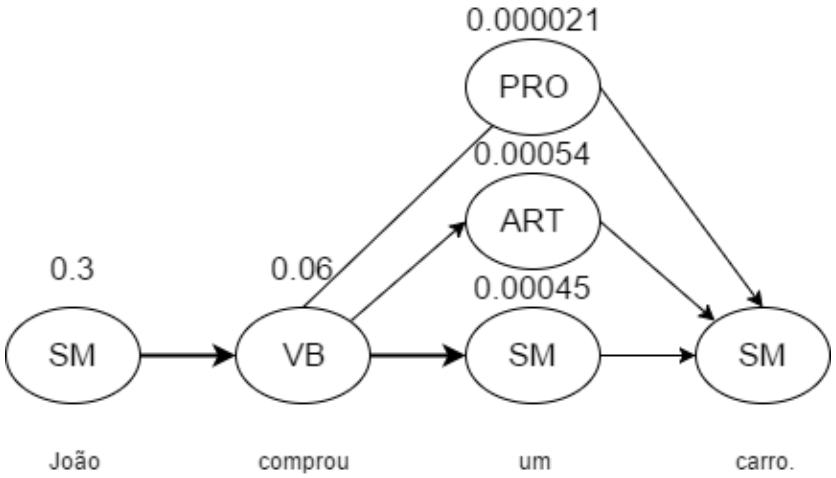


Figura 2.2: Caminhos já decididos de classificação

Como visto, o método simbólico para resolver problemas de Processamento de Linguagem Natural faz uso da criação de regras baseadas no conhecimento humano, enquanto o método estatístico, decide através de cálculos probabilísticos apoiados em estatísticas de um banco de dados para a resolução correta do problema.

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 itens sujeitos a nossa crítica. Porém, muitas vezes essas opiniões acabam por ser esquecidas pela dificuldade de se analisar uma grande quantidade de textos. Como saber a opinião geral sobre determinado produto em uma seção de comentários com mais de 1000 opiniões diferentes? A análise de sentimentos, considerada uma tarefa do NLP, tem como função identificar e quantificar esses sentimentos expressos através de textos.

Neste capítulo serão descritos um método estatístico e um método simbólico aplicados na análise de sentimentos. Ambos considerando o uso da língua Inglesa, visto que o *Website* analisado (Reddit) possui a maioria de seus comentários nessa língua e 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. Ele faz uso do teorema de Bayes para que a partir de um *training set* se possa inferir uma classificação.

Por exemplo, precisamos determinar se a seguinte frase demonstra um sentimento negativo ou positivo: “This place is great.”, como este método faz uso de um *training set* para a classificação, será considerado o seguinte *training set* hipotético:

A partir do *training set* representado pela tabela 3.1 o método irá calcular a probabilidade da frase “This place is great” ser positiva e também a possibilidade dela ser uma frase negativa e a partir dessas duas possibilidades, será escolhida a maior.

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*

3.1.1 Teorema de Bayes

Para cálcular a probabilidade da frase "This place is great" pertencer a cada categoria utilizando o teorema da Bayes, é aplicada a Equação 3.1:

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

Os termos da Equação 3.1 são determinados da seguinte forma:

- P(c|d) é a probabilidade de **d** pertencer a classe **c**. Ou seja, a probabilidade de *"This place is great"* ser uma frase positiva ou negativa.
- P(d|c) é a probabilidade da classe **c** ser **d**. Ou seja, dentre todas as frases negativas ou positivas, a probabilidade de uma frase ser *"This place is great"*.
- P(c) é a probabilidade da classe **c**. Ou seja, a frequência que frases negativas ou positivas aparecem em nosso *training set*.
- 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 terão como divisor P(d), este é removido da equação, tendo como comparação entre as probabilidades negativas e positivas as Equações 3.2 e 3.3:

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

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

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 3.4 e 3.5. 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, do conjunto de 5 frases do *training set*.

$$P(Positiva) = \frac{3}{5} = 0,6 \tag{3.4}$$

$$P(Negativa) = \frac{2}{5} = 0,4 \tag{3.5}$$

Porém, a frase *"This place is great"* não existe por completo no *training set* tornando o termo P(d|c) da Equação 3.1 0 e impossibilitando o cálculo de probabilidade para essa frase. Neste caso, se faz o uso do *Naive Bayes*, o qual passa a considerar palavras ao invés de frases completas, solucionando o problema de frases não encontradas no *training sets*.

3.1.2 *Naive Bayes* aplicado ao Teorema

Bayes como visto, é relacionado com o teorema utilizado para cálculo da probabilidade, já a palavra *naive*, ou ingênuo, é relacionada uma outra característica, que é a independência entre atributos. Esses, para o aprendizado de máquina, são características da informação que estamos classificando. Por exemplo, ao efetuar uma classificação relacionada com medicina, atributos que seriam considerados poderiam ser histórico de doença, altura da pessoa e peso. No caso da análise de sentimentos, os atributos são as próprias palavras do texto, ou seja, em sua classificação, ele ignora a ordem das palavras e somente considera a frequência na qual elas aparecem. Portanto para o *Naive Bayes* o termo P(*This place is great*| Positiva) visto na Equação 3.3 é dado pela Equação 3.6:

$$\begin{aligned} P(\textit{This place is great}|Positiva) = P(\textit{This}|Positiva) \times P(\textit{place}|Positiva) \\ \times P(\textit{is}|Positiva) \times P(\textit{great}|Positiva) \end{aligned} \tag{3.6}$$

A partir da Equação 3.6 se faz necessário cálcular os termos P(*This*|Positiva), P(*place*|Positiva), P(*is*|Positiva), P(*great*|Positiva). Aonde que, por exemplo, P(*This*|Positiva) é a quantidade de vezes que a palavra *This* foi classificada como positiva em nosso *training set*, dividido pelo total de palavras classificadas como positiva:

$$P(\textit{This}|Positiva) = \frac{1}{13} \tag{3.7}$$

Da mesma forma, a Equação 3.7 é aplicada para as demais palavras da frase que está sendo classificada obtendo os valores representados pela Tabela 3.2:

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.

Como algumas palavras não existem em nosso *training set* para determinadas situações, elas acabam tornando o resultado final da multiplicação das probabilidades de cada palavra (Equação 3.6) como 0, para evitar que uma única palavra invalide uma frase é utilizado *Laplace smoothing*. Neste, é somado 1 a cada palavra e ao total de palavras, são somadas as quantidades de palavras diferentes do *training set*, que neste caso são as seguintes 16 palavras distintas: *"The"*, *"food"*, *"was"*, *"great"*, *"They"*, *"are"*, *"horrible!"*, *"I"*, *"love"*, *"here"*, *"This"*, *"place"*, *"is"*, *"wonderful"*, *"Forgettable"*, *"experience"*. Neste caso, 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 sob os termos da Equação 3.6 o classificador obtem a seguinte Equação:

$$P(Positiva|\,This\ place\ is\ great) = \frac{1+1}{13+16} \times \frac{1+1}{13+16} \times \frac{1+1}{13+16} \times \frac{1+1}{13+16} = 0,000023. \tag{3.8}$$

Com o termo P(Positiva | *This place is great*) definido através da Equação 3.8, este é aplicado a Equação 3.2, junto com o termo P(Positiva) definido através da Equação 3.4, obtendo a probabilidade da frase *"This place is great"* ser classificada como positiva através da Equação 3.9:

$$P(Positiva|\,This\ place\ is\ great) = 0,000023 \times 0,6 = 0,0000138. \tag{3.9}$$

Efetuando o mesmo processo para a probabilidade da frase ser negativa, obtemos:

$$P(Negativa|\,This\ place\ is\ great) = 0,0000049 \times 0,4 = 0,00000196. \tag{3.10}$$

Portanto, a partir dessas duas possibilidade, utilizando o método de classificação Naive Bayes para efetuar a análise de sentimentos, ele iria classificar a frase "This place is great" como positiva por essa probabilidade (0,0000138) ser maior que a probabilidade dessa frase ser negativa (0,00000196).

3.2 *VADER*

O *Valence Aware Dictionary and sEntiment Reasoner* (VADER) é um dicionário e classificador de sentimentos que se baseia em regras, portanto, um método de classificação simbólico. Ele é especialmente ajustado para funcionar em redes sociais aonde temos um contexto vago e pouca quantidade de texto, nesse contexto, ele é extremamente eficaz, podendo se comparar a classificação feita por humanos (HUTTO; GILBERT, 2014).

Esse método faz uso de um dicionário que foi construído levando em consideração gírias e emoticons utilizados em redes sociais. Neste dicionário as palavras estão previamente associadas a uma polaridade de sentimento (positivo e negativo) e 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 associação foi construída utilizando o método de *"wisdom of the crowd"* aonde um grupo de pessoas atribuiu os valores para cada palavra ao invés de somente uma pessoa especializada ou uma classificação automática através de estatística.

Ele faz uso de cinco regras gerais:

- Pontuação. O ponto de exclamação (!) aumenta a magnitude da intensidade sem modificar a orientação semântica. Como por exemplo, *"This place is great!!!"* é mais intenso que *"This place is great"*.
- Capitalização. Especificamente, uma palavra que é relevante para a análise de sentimentos, quando essa é escrita em letras maiúsculas, é aumentada a magnitude da intensidade do sentimento sem modificar a orientação semântica. Como por exemplo, na frase *"This place is GREAT"*, temos a palavra *"GREAT"* (Ótimo) que está relacionada com o sentimento positivo. Neste caso aonde ela está escrita em letras maiúsculas, ela é mais intensa que *"This place is great"*.
- Advérbios intensificadores. Estes impactam a intensidade do sentimento aumentando ou diminuindo a intensidade do sentimento. Na frase *"This place is extremelly good"* o advérbio *extremelly* (extremamente) aumenta a intensidade do sentimento expresso pela frase (*good* ou bom), enquanto na frase

"This place is marginally good", a palavra *"marginally" ou marginalmente* acaba diminuindo a intensidade do sentimento expresso.

- A palavra *"but"*. Essa palavra indica uma troca no sentimento da frase expressa aonde que o texto seguinte a ela expressa um sentimento mais dominante. Por exemplo, a frase *"This place is great but today, the service was horrible"* convém um sentimento misto.
- Por fim, ao examinar as três palavras anteriores, o método consegue identificar 90% dos casos aonde uma negação inverte a polaridade de um texto. Como por exemplo, na frase *"This place isn't that great"*, a palavra *great* demonstra um sentimento positivo, porém, ao analisar as três palavras anteriores *"place isn't that"* encontramos uma negação, mudando o sentimento expresso da frase de positivo para negativo.

4 FRAMEWORKS

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*, *Maximum Entropy* (MaxEnt) e também 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*.

Podemos utilizar o classificador MaxEnt a partir do módulo **nltk.classify.maxent** através dos seguintes métodos:

- *train(train_toks, algorithm=None, trace=3, encoding=None, labels=None, gaussian_prior_sigma=0, **cutoffs)* - Treina um classificador MaxEnt a partir de

um *training set*.

  - *train_toks* - *Training set*.
  - *algorithm* - Algorítmo a ser usado para treinar o classificador.
  - *trace* - Nível de detalhe utilizado no log.
  - *encoding*
  - *labels* - Uma lista de possíveis rótulos, se nenhuma for especificada, todos os labels do *training set* serão utilizados.
  - *gaussian_prior_sigma=0* - Somente utilizado no LM-BFGS.
  - *cutoffs* - Argumentos que especificam condições em que o processo será terminado.

- *classify(featureset)* - Classifica a partir de um conjunto de atributos.
- *explain(featureset, columns=4)* - Mostra uma tabela demonstrando os efetiso de cada atributo e como eles combinam para determinar a probabilidade de cada rótulo.
- *show_most_informative_features(n=10, show='all')* - A partir de um classificador treinado, retorna os atributos mais relevantes.

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.

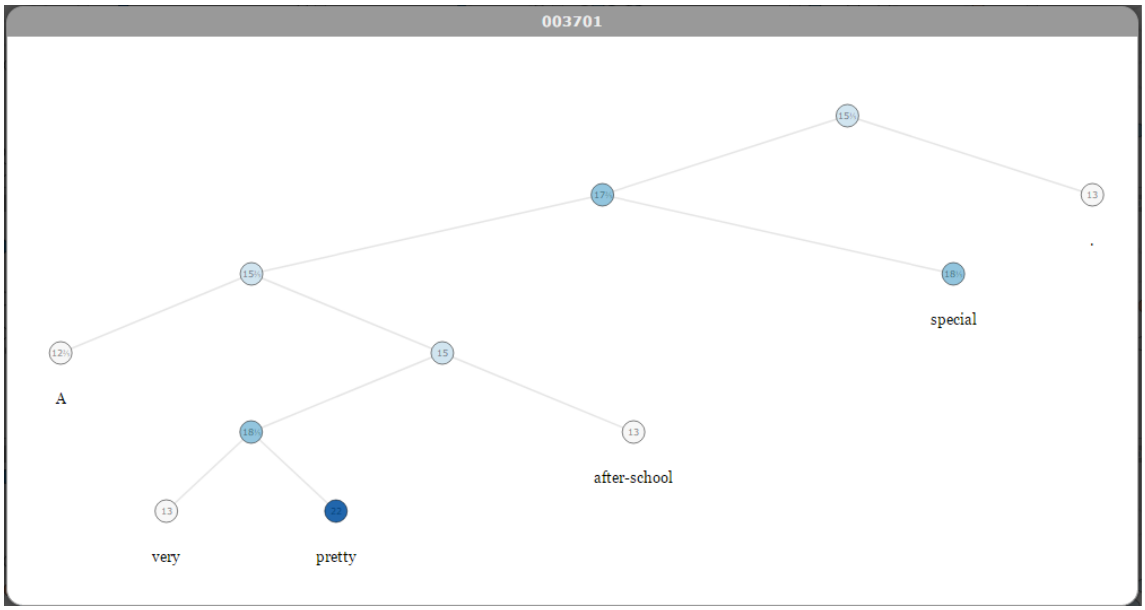


Figura 4.1: Frase já classificada disponível no Sentiment Treebank

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

```
1  public static void main(String[] args) throws IOException {
2          String text = "This World is an amazing place";
3          Properties props = new Properties();
4          props.setProperty("annotators", "sentiment");
5          StanfordCoreNLP pipeline = new StanfordCoreNLP(props);
6
7          Annotation annotation = pipeline.process(text);
8          List<CoreMap> sentences = annotation.get(CoreAnnotations.SentencesAnnotation.class);
9          for (CoreMap sentence : sentences) {
10             String sentiment = sentence.get(SentimentCoreAnnotations.SentimentClass.class);
11             System.out.println(sentiment + "\t" + sentence);
12         }
13 }
```

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.

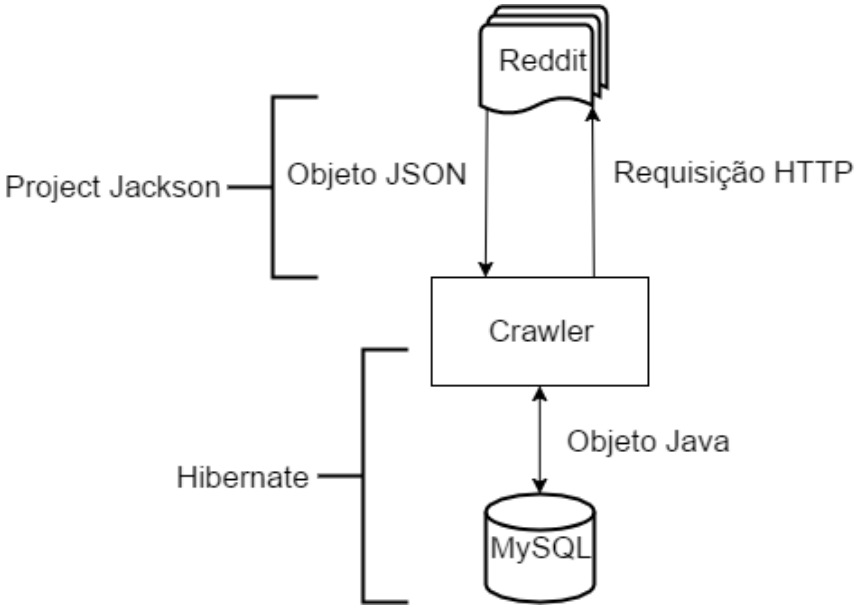


Figura 6.1: Arquitetura do *Crawler*

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"*[2].
- *Brazil descends into chaos as Olympics looms*[3].
- *Plane carrying Brazil Supreme Court judge crashes into sea* [4].
- *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.*[5].
- *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*

[2] `https://www.reddit.com/r/worldnews/comments/36ny58/brazil_blogger_known_for_reporting_on_corruption/`

[3] `https://www.reddit.com/r/worldnews/comments/4bqcc3/brazil_descends_into_chaos_as_olympics_looms/`

[4] `https://www.reddit.com/r/worldnews/comments/5oyz3b/plane_carrying_brazil_supreme_court_judge_crashes/`

[5] `https://www.reddit.com/r/worldnews/comments/21f3as/brazil_passes_internet_governance_bill_brazil_has/`

World Cup[6].

- *Brazil Seeks To Copy U.S. Gun Culture "to allow embattled citizens the right to defend themselves from criminals"*[7].

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.".*[8].
- *Fidel Castro is dead at 90.*[9].
- *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*[10].
- *Manchester Arena 'explosions': Two loud bangs heard at MEN Arena*[11].
- *Sweden asks the U.S. to explain Trump comment on Sweden*[12]
- *"Canada will welcome you," Trudeau invites refugees as Trump bans them*[13]

Tópicos diversos.

- *I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation. Ask Me Anything.*[14].
- *Hey, it's Lars from Metallica. AMA*[15].
- *I'm the CEO of Renault and Nissan and we're making autonomous driving vehicles happen by 2020. Ask me anything!*[16].

---

[6] https://www.reddit.com/r/worldnews/comments/2t65ql/fifa_generated_more_than_4_billion_in_sales_from/

[7] https://www.reddit.com/r/worldnews/comments/3skpe7/brazil_seeks_to_copy_us_gun_culture_to_allow/

[8] https://www.reddit.com/r/worldnews/comments/4d75i7/26_terabyte_leak_of_panamanian_shell_company_data/

[9] https://www.reddit.com/r/worldnews/comments/5exz2e/fidel_castro_is_dead_at_90/

[10] https://www.reddit.com/r/worldnews/comments/67ivae/donald_trump_to_strip_all_funding_from_state_dept/

[11] https://www.reddit.com/r/worldnews/comments/6cqdye/manchester_arena_explosions_two_loud_bangs_heard/

[12] https://www.reddit.com/r/worldnews/comments/5uzetf/sweden_asks_the_us_to_explain_trump_comment_on/

[13] https://www.reddit.com/r/worldnews/comments/5qqa51/canada_will_welcome_you_trudeau_invites_refugees/

[14] https://www.reddit.com/r/IAmA/comments/5whpqs/im_bill_gates_cochair_of_the_bill_melinda_gates/

[15] https://www.reddit.com/r/IAmA/comments/1wl9ic/hey_its_lars_from_metallica_ama/

[16] 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*[17].

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.

[17] `https://www.reddit.com/r/IAmA/comments/5n58sm/i_am_julian_assange_founder_of_wikileaks_ask_me/`

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).

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.; 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.

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.