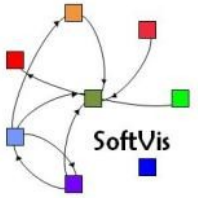
PROGRAMA DE
PÓS-GRADUAÇÃO EM
ENGENHARIA DE SISTEMAS E PRODUTOS



Análise de Sentimentos

Gláucya Boechat
Doutorando em Ciência da Computação
LES-DCC-UFBA
glaucya.boechat@ufba.br

Agenda

- Definição de Análise de Sentimentos
- Exemplos de Aplicações
- Exemplos práticos
- Exercícios
- Referências

Análise de sentimentos

- Análise de sentimentos (também conhecida como mineração de opinião) é um campo de estudo que analisa fragmentos textuais e determina a emoção, opinião, apreciação, avaliação ou sentimento do escritor com relação a algum tópico, como produtos, indivíduos, eventos e seus atributos (Pang & Lee 2008, Liu 2012).

Análise de sentimentos

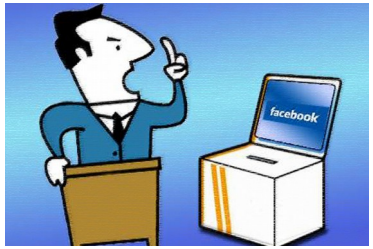
- Os sentimentos podem ser classificados como positivo, negativo e neutro e, além disso, estes podem estar associados a emoções, como felicidade, raiva, tristeza, dentre outras.

Fato x opinião

- Informação textual pode ser classificada em dois tipos principais:
  - Fatos - Expressões objetivas que traz informações concretas sobre entidades, eventos ou suas propriedades.
    - Ex: “Eu comprei um iPhone.”
  - Opiniões - Expressões subjetivas que descrevem os sentimentos pessoais, opiniões, avaliações e emoções.
    - Ex: “A câmera do meu iPhone é boa.”

Opiniões na Web

- São encontrados em:
  - Redes Sociais
    - Ex. Facebook, Instagram, Twitter, Youtube.
  - Avaliações de Sites de Comércio Eletrônico
    - Submarino, Ponto Frio, Amazon, Ebay.
  - Fóruns e Grupos de Discussão
    - Fórum Ubuntu Linux.
  - Comentários de Blogs e sites
    - Reclame Aqui, Google shopping.



Tipos de opiniões

- Regular
  - Direta – expressa um sentimento para um determinado alvo.
    - “A tapioca está ótima.”
  - Indireta – expressa indiretamente um sentimento para um alvo.
    - “Logo após comer a tapioca, passei mal.”

Tipos de opiniões

- Comparativa
  - Expressa um sentimento de comparação entre dois ou mais alvos.
  - Exemplos:
    - “Eu acho a série House of Cards melhor que a The Rain.”
    - “Game of Thrones é a melhor série dos últimos 20 anos.”

Aplicações Práticas de SA

- Aceitação ou rejeição de um determinado político durante a época de eleições.
- Analisar a aceitação de um novo produto ou serviço lançado no mercado.
- Tomada de decisão sobre compras e vendas de açõe.
- Buscar opinião de outros consumidores antes de efetuar uma compra.

Desafios

- Textos curtos (140 caracteres)
- Linguagem informal, gírias e emoticons
- Caracteres repetidos !!!
- Presença de URLs
- Textos irônicos

Modelo da Análise de Sentimentos

Etapa: Coleta de Dados

- A coleta de dados está relacionada com a extração de textos de alguma fonte (e.g. redes sociais, blogs, fóruns, etc) através de busca de palavras-chave e hashtags sobre um determinado alvo.
  - #bigdata
  - #analisedesentimentos

Etapa: Pré-processamento

- Etapa responsável pela preparação dos dados para serem processados na próxima etapa.
- Nessa etapa checagem é feita uma limpeza dos dados na qual, pode ser eliminado:
  - Qualquer tipo de acentuação, pontuação, caractere especial, números, URLs.
  - Erros ortográficos
  - Normalização (UFBA vs ufba)

Processamento de Linguagem Natural

- Processamento de Linguagem Natual (PLN) um conjunto de técnicas computacionais para analisar e representar dados textuais com o propósito de realizar um processamento de uma linguagem similar ao humano para diversas tarefas e aplicações (Liddy, 2001)

Remoção de stop-words

- Consiste em remove todas as palavras do texto que estão presentes nas lista de Stop-words.
- Stop-words é uma lista de palavras conhecidas e que não agrega muito no que diz respeito ao sentimento de uma sentença.
  - Exemplos de stop-words da língua inglesa:
    - The, a, about, etc.

Stemming (Stemização)

- Processo de reduzir palavras flexionadas (ou às vezes derivadas) ao seu tronco (stem), base ou raiz, geralmente uma forma da palavra escrita.
- O tronco não precisa ser idêntico à raiz morfológica da palavra.
- Exemplos
  - "avião", "aviões", "aviação", "viação", "aves", "balão", "balões"
  - **Stem** : "avião" , "avião" , "avião" , "viação" ,"aves" , "balão" , "balão"

Lemmatization (Lematização)

- Reduzir inflexões das palavras ou formas variantes para a forma base (forma canónica).
  - Lema : palavras com mesmo tronco (forma canônica)
  - Exemplo
    - “Tenho”, ”tens”, ”temos”
    - **Lema**: ”ter” (verbo)

Tokenização

- A tokenização consiste na segmentação de um texto em unidades linguísticas como palavras, pontuações, números, alfanuméricos, etc (TRIM, 2013).
- Exemplo
  - “Salvador é uma cidade do estado da Bahia.”
  - **Tokenização:** [Salvador] [é] [uma] [cidade] [do] [estado] [da] [Bahia][.]

Bag-of-Words (BoW)
Bolsa de Palavras

- Chamada de unigrama, indica que cada palavra representa uma feature.
  - Ex. “João gosta de assistir filmes. Maria também gosta de filmes.”
  - **Representação em BoW** : [“João” , “gosta” , “de” , “assistir” , “filmes”, “Maria”, “também”]

Bag-of-Words (BoW)

- Exemplo: “The restaurant is not good!”

  Representação em Bow: “The”, “restaurant” , “is” , “not” , “good”

  good → palavra de conotação positiva.

  - *No entanto a frase good possui conotação negativa pois ela é precedida da palavra not.*

N-gram

- Variações de Bag-of-words que permitem que cada feature possa ser representada por um grupo de palavras (bigrams, trigrams, 4-grams, ...).
  - Ex. *The restaurant is not good!*
  - **Representação em bigram:** “the restaurant”, ”restaurant is” , ”is not”, ”not good”
  - ”not good” → conotação negativa

Etapa: Classificador

- Principais Abordagens para Análise de Sentimentos
  - Aprendizagem de máquina
  - Baseado em Léxicos

Classificação de texto

- Definição
  - Entrada:
    - Um documento d
    - Um grupo fixo de classes $C = \{c_1, c_2, \ldots, c_J\}$
  - Saída:
    - Uma classe prevista $c \in C$

Exemplo de documento

- (1)Eu comprei um iPhone alguns dias atrás. (2)Ele parecia ser um ótimo celular. (3) O touch screen era realmente bom. (4)A qualidade de voz era clara também. (5)Porém, minha mãe ficou furiosa comigo por não ter avisado a ela antes de ter feito a compra dele. (6)Ela também achava que o celular era muito caro e queria que eu o devolvesse para a loja.

Níveis de Classificação dos Sentimentos

- No nível de documento
  - Analisa o documento inteiro.
- No nível de frase
  - Analisa cada frase separada
- No nível do aspecto/característica.
  - Analisa os sentimentos de cada aspecto/característica
    - Classificação mais refinada

Exemplo de documento

- (1)Eu comprei um iPhone alguns dias atrás. (2)Ele parecia ser um ótimo celular. (3) O touch screen era realmente bom. (4)A qualidade de voz era clara também. (5)Porém, minha mãe ficou furiosa comigo por não ter avisado a ela antes de ter feito a compra dele. (6)Ela também achava que o celular era muito caro e queria que eu o devolvesse para a loja.

- Nível de classificação de documento
  - Documento **– neutro**

Exemplo de documento

- (1)Eu comprei um iPhone alguns dias atrás. (2)Ele parecia ser um ótimo celular. (3) O touch screen era realmente bom. (4)A qualidade de voz era clara também. (5)Porém, minha mãe ficou furiosa comigo por não ter avisado a ela antes de ter feito a compra dele. (6)Ela também achava que o celular era muito caro e queria que eu o devolvesse para a loja.

- Nível de classificação de frase
  - Frases (2),(3),(4) – **Positivas**
  - Frases (5),(6),(7) – **Negativas**

Exemplo de documento

- (1)Eu comprei um iPhone alguns dias atrás. (2)Ele parecia ser um ótimo celular. (3) O touch screen era realmente bom. (4)A qualidade de voz era clara também. (5)Porém, minha mãe ficou furiosa comigo por não ter avisado a ela antes de ter feito a compra dele. (6)Ela também achava que o celular era muito caro e queria que eu o devolvesse para a loja.

- Nível de classificação de aspecto/característica
  - iPhone – **Positiva**
  - ....

Aprendizagem de máquina

- Aprendizado supervisionado
  - Composta por técnicas que emprega o termo supervisionado, na etapa de treinamento é utilizado amostras previamente classificadas.
- Aprendizado não-supervisionado
  - Diferente do aprendizado supervisionado, não utiliza amostras previamente classificadas para criação do modelo.

Aprendizado supervisionado

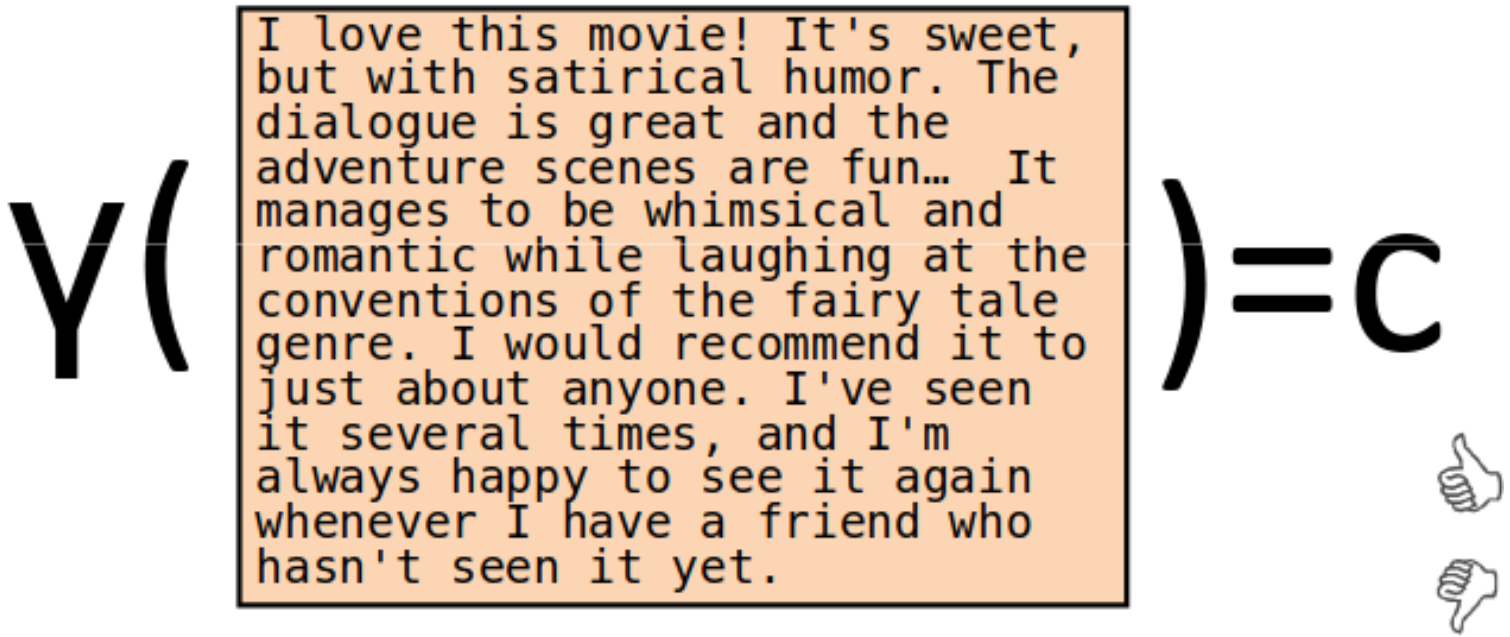
- Entrada:
  - Um documento d.
  - Um grupo fixo de classes $C = \{c_1, c_2,..., c_j\}$.
  - Um conjunto de treinamento de mdocumentos rotulados manualmente $(d_1,c_1),....,(d_m,c_m)$.

- Saída:
  - Um classificador treinado $y:d \rightarrow c$.

Aprendizado supervisionado

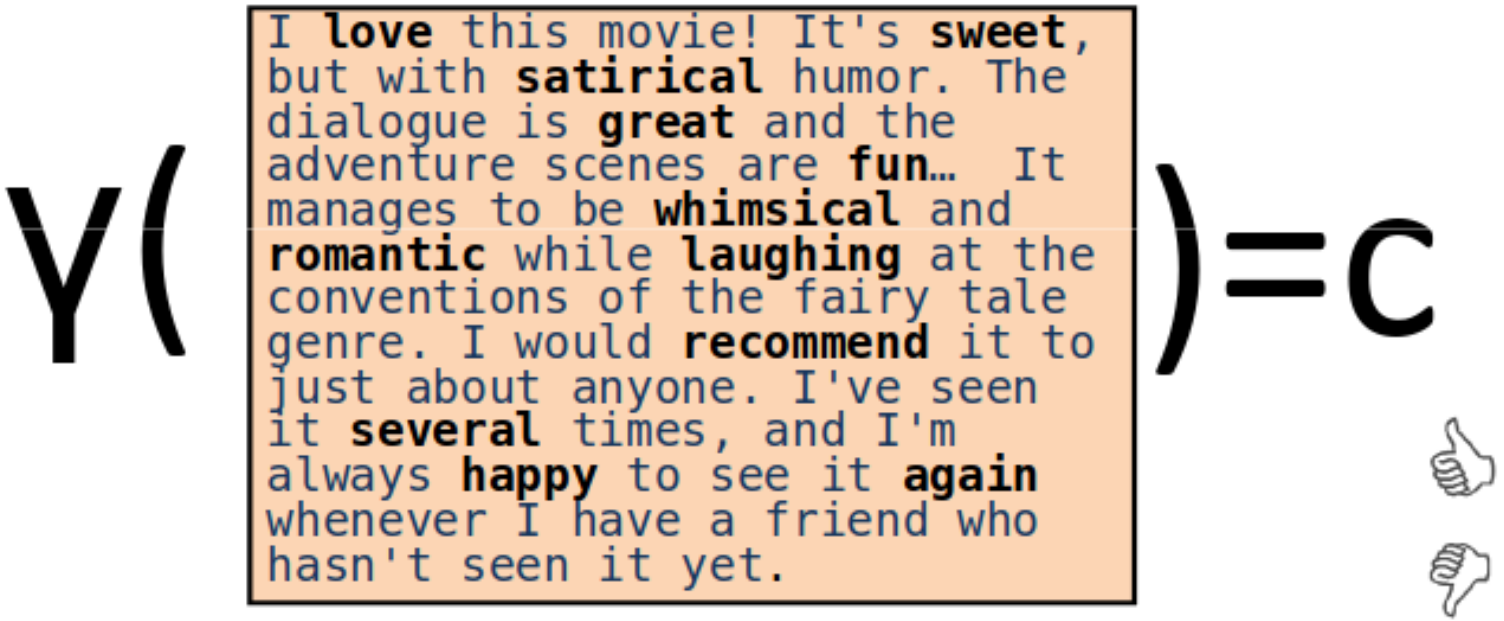
- Tipos de classificadores:
  - Naïve Bayes
  - Regressão Logística
  - Máquinas de Vetores de Suporte
  - Redes Neurais Artificiais
  - ...

Análise de Sentimentos



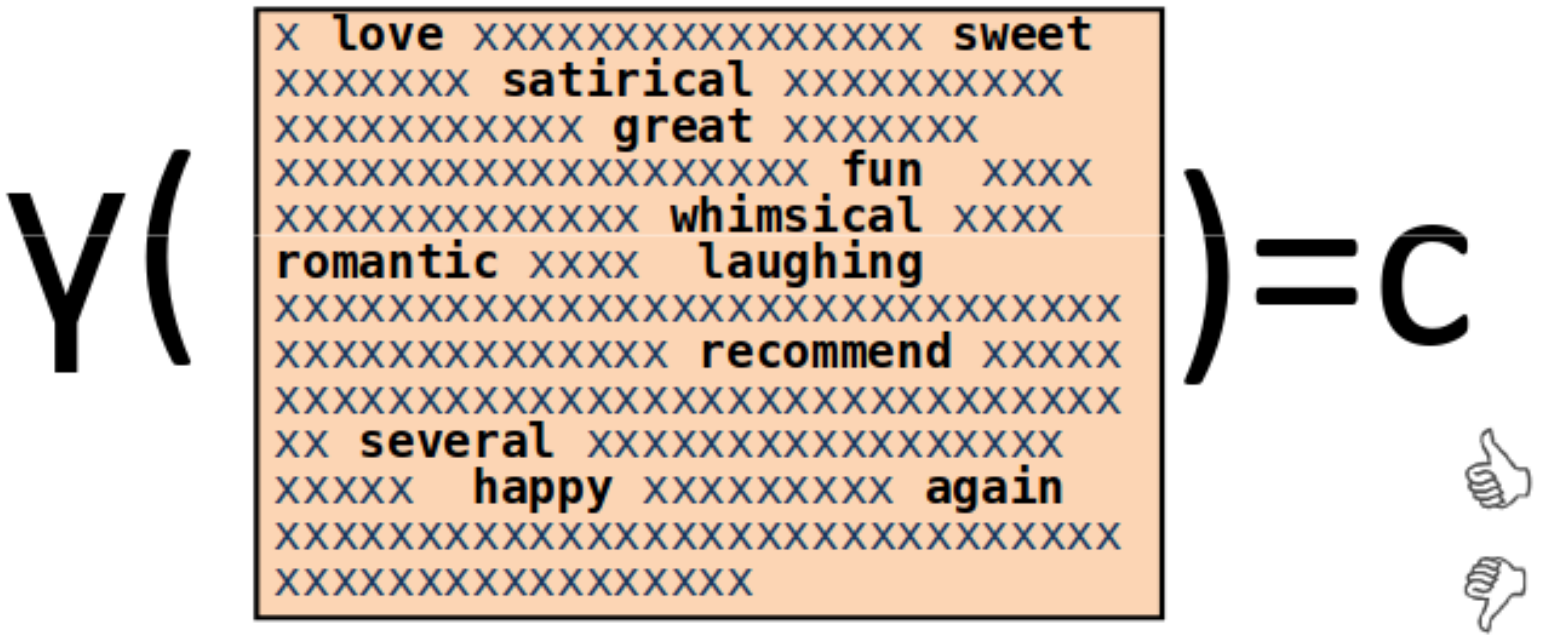
Análise de Sentimentos

- Identificando as opiniões do documento



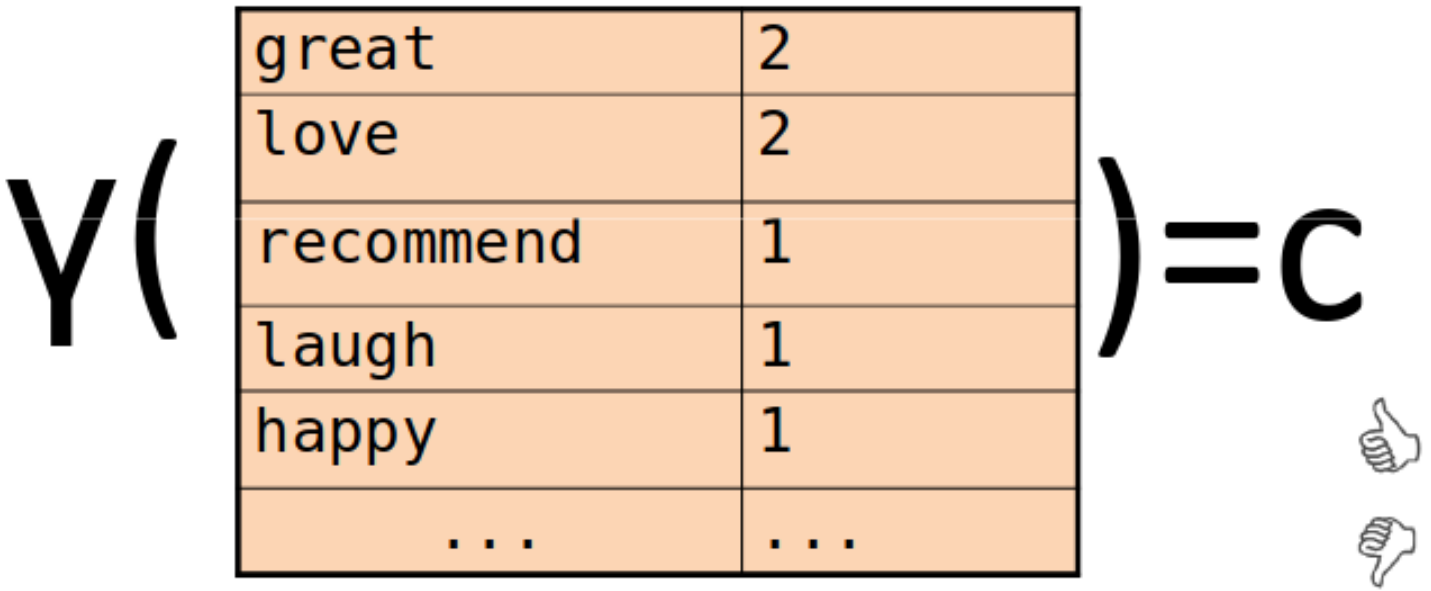
Análise de Sentimentos

- Representação BoW das opiniões



Análise de Sentimentos

- Representação BoW das opiniões



Classificador Naïve Bayes

- Utiliza como base o Teorema de Bayes.
- Conta com uma representação Bag-of-Words (BoW).
- Para analisar se um documento d e uma classe c.

$$P(c \mid d) = \frac{P(d \mid c)\,P(c)}{P(d)}$$

Abordagem baseada em Léxicos

- A abordagem Léxica utiliza um Dicionário de palavras de sentimento (Chamado de léxico) para auxiliar na tarefa de classificação dos sentimentos.
- O léxico contém as palavras de sentimento que geralmente foram coletadas e classificada manualmente.
- Classificação das sentimentos :
  - negativa (-1), neutra (0) e positiva (1).

Exemplo de arquivo léxico

```
## # A tibble: 23,165 × 4
##           word sentiment
##          <chr>     <chr>
## 1       abacus     trust
## 2      abandon      fear
## 3      abandon  negative
## 4      abandon   sadness
## 5    abandoned     anger
## 6    abandoned      fear
## 7    abandoned  negative
## 8    abandoned   sadness
## 9  abandonment     anger
## 10 abandonment      fear
## # ... with 23,155 more rows
```

Exemplos de léxicos

- LexiconPT
  - https://github.com/sillasgonzaga/lexiconPT
  - Léxico em Português
  - Disponível na Linguagem R
- Bing Liu Opinion Lexicon
  - Léxico em Inglês
  - Classificação: Positivo e Negativo
  - https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

```
## # A tibble: 23,165 × 4
##            word sentiment
##           <chr>     <chr>
## 1         abacus     trust
## 2       abandon      fear
## 3       abandon  negative
## 4       abandon   sadness
## 5     abandoned     anger
## 6     abandoned      fear
## 7     abandoned  negative
## 8     abandoned   sadness
## 9   abandonment     anger
## 10 abandonment      fear
## # ... with 23,155 more rows
```

Etapa: Análise

- Responsável por avaliar e interpretar os resultados do classificador
- Matriz de Confusão

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Métrica

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right)$$

$$Recall = \left(\frac{TP}{TP + FN}\right)$$

$$Precision = \left(\frac{TN}{TN + FP}\right)$$

$$F - measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right)$$

Exemplos de Aplicações

SentiStrength

- Realiza detecção automática de sentimento de textos curtos da web.
- http://sentistrength.wlv.ac.uk/

The text 'Eu te amo, mas odeio o clima político atual.'
has positive strength **1** and negative strength **-4**

Approximate classification rationale: Eu te amo ,mas odeio[-4] o clima político atual .[sentence: 1,-4]
[result: max + and - of any sentence][overall result = -1 as pos<-neg] (Portuguese)

Globo Esporte: Humor das torcidas

- Mostra os sentidos dos torcedores no Twitter

HUMOR DAS TORCIDAS

Veja quais são as torcidas mais otimistas e pessimistas no Twitter. Os dados referem-se à análise de sentimento das últimas 3 horas. Participe e compartilhe a sua opinião.

Clique no seu time e veja os dados completos.

	TIME	AVALIAÇÃO DO TORCEDOR		TWEETS	Como está o seu humor?
1	Internacional	85%	15%	1,341	
2	Atlético-PR	84%	16%	64	
3	São Paulo	82%	18%	348	
4	Atlético-MG	82%	18%	1,292	
5	Flamengo	71%	29%	2,383	
6	Cruzeiro	70%	30%	1,466	
7	Santos	68%	32%	178	
8	Figueirense	67%	33%	432	
9	Fluminense	64%	36%	414	
10	Grêmio	63%	37%	1,441	

SentiWordNet

SentiWordNet

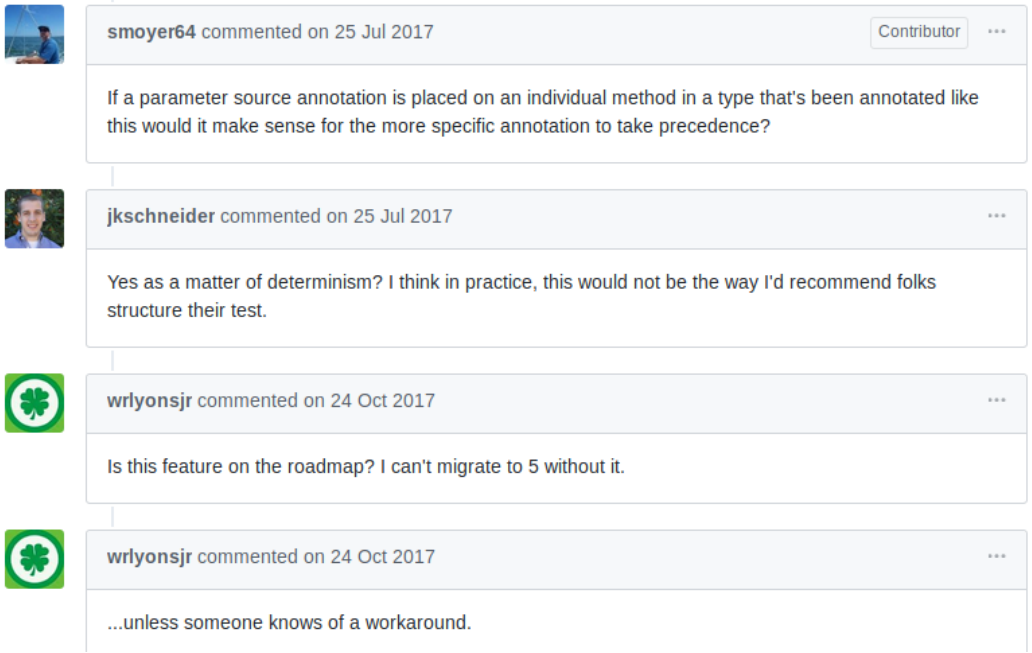
- O método é baseado no dicionário léxico WordNet, onde as palavras em inglês são agrupadas em conjuntos de sinônimos.
- E para cada conjunto é atribuito três valores de pontuação que indicam o sentimento de um texto, positividade, negatividade, objetividade (neutro).
- http://sentiwordnet.isti.cnr.it/

Ifeel - Sentiment Analysis Framework

- Usa 18 métodos de análise de sentimento
  - SentiStrength, Sentiment140, SentiWordNet, VADER, etc.
  - http://blackbird.dcc.ufmg.br:1210

AS e Engenharia de Software

- Analisar os sentimentos dos desenvolvedores de software através das mensagens de Issue



Exemplos Práticos em Python

- Configurações necessárias
  - python --version
  - sudo apt-get install python
  - sudo pip install numpy
  - pip install nltk
  - pip install pandas
  - pip install sklearn
  - pip install scipy

Base de Dados do Twitter

- Exemplos de Tweets sobre Minas Gerais
  - https://github.com/minerandodados/mdrepo/blob/master/Tweets_Mg.csv

	Created At	Text	Classificacao
0	Sat Jan 07 13:47:55 +0000 2017	" bom é bandido morto" Deputado Cabo Júlio é condenado e fica inelegível por 10 anos - Politica - Estado d	Neutro
1	Wed Jan 04 23:00:53 +0000 2017	"..E 25% dos mineiros dizem não torcer para time nenhum,mesmo dentro de um es	Neutro
2	Sun Jan 08 18:34:22 +0000 2017	"A gigantesca barba do mal" em destaque no caderno Cultura do Estado de Minas.	Neutro
3	Wed Jan 04 22:55:08 +0000 2017	"BB e governo de Minas travam disputa sobre depósitos judiciais" https://t.co/CnM	Negativo
4	Sat Jan 07 01:37:10 +0000 2017	"com vcs bh fica pequena!" Belo Horizonte (pron. [bɛloriˈzõntʃi][10]) é a capital	Neutro
5	Thu Jan 05 00:41:09 +0000 2017	"É bonita e é bonita..." #latergram #ibituruna #home @ Governador Valadares, Minas Gerais https://t.co/yI	Neutro
6	Mon Jan 09 16:06:23 +0000 2017	"erro desconhecido" é mato! Aliás, é da secretaria estadual de fazenda que tá assi	Negativo
7	Mon Jan 09 09:45:07 +0000 2017	"La La Land: Cantando Estações" arrasa no Globo de Ouro - Estado de Minas https	Neutro
8	Wed Jan 04 21:17:07 +0000 2017	#DefesaAgropecuária "Governo de Minas Gerais aposta nos Arranjos Produtivos L	Positivo
9	Mon Jan 09 22:24:33 +0000 2017	#EBC Governo de Minas investiga casos suspeitos de febre amarela e malária no e	Positivo

Analisando mensagens do Twitter

```
# -*- coding: utf-8 -*-
import nltk
import re
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics
from sklearn.model_selection import cross_val_predict
```

Analisando mensagens do Twitter

- Função de Pré-processamento

```
def Preprocessamento(instancia):
    #remove links, pontos, virgulas,ponto e virgulas dos tweets
    #coloca tudo em minusculo
    instancia = re.sub(r"http\S+", "", instancia).lower().replace(',','').
replace('.','').replace(';','').replace('-','').replace(':','')
    return (instancia)
```

Analisando mensagens do Twitter

- Coleta dos dados

```
#Ler arquivo de dados e conta a quantidade de linhas
dataset = pd.read_csv('Tweets_Mg.csv',encoding='utf-8')

#Separando tweets(Entrada x) e seus sentimentos (Classe y)
tweets = dataset['Text'].values
classes = dataset['Classificacao'].values
```

Analisando mensagens do Twitter

- Extração de Características

```
#Pre-processamento -> Extracao de caracteristicas
vectorizer = CountVectorizer(analyzer="word")   #bag of words
freq_tweets = vectorizer.fit_transform(tweets)
```

Analisando mensagens do Twitter

- Treinamento do modelo

```
#Gerando o modelo (Treinamento Supervisionado)

# MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
modelo = MultinomialNB()
modelo.fit(freq_tweets,classes)
```

Analisando mensagens do Twitter

- Testando o modelo

```
#### TESTE
#Instancias de teste dentro de uma lista
testes = ['Esse governo está no início, vamos ver o que vai dar',
'Estou muito feliz com o governo de Minas esse ano',
'O estado de Minas Gerais decretou calamidade financeira!!!',
'A segurança desse país está deixando a desejar',
'O governador de Minas é do PT']

classes_testes = ['Neutro', 'Positivo', 'Negativo', 'Negativo', 'Neutro']

freq_testes = vectorizer.transform(testes)
resultados_testes = modelo.predict(freq_testes)
```

Analisando mensagens do Twitter

- Avaliando o modelo

```
# Acurácia média
print (metrics.accuracy_score(classes_testes,resultados_testes))

# Medidas de validação (Precision, Recall, F1-Score)
sentimento=['Positivo','Negativo','Neutro']
print (metrics.classification_report(classes_testes,
resultados_testes,sentimento))
```

Analisando mensagens do Twitter

- Testando o modelo
  - Acurácia = 0.883156482497865 6

	precision	recall	f1-score	support
Positivo	0.00	0.00	0.00	1
Negativo	1.00	1.00	1.00	2
Neutro	0.67	1.00	0.80	2
avg / total	0.67	0.80	0.72	5

Analisando mensagens do Twitter

- Cross validation k-fold (10)

```
#Cross validation k-fold (10)
resultados = cross_val_predict(modelo, freq_tweets, classes, cv=10)

# Acurácia média
metrics.accuracy_score(classes,resultados)

# Medidas de validação (Precision, Recall, F1-Score)
sentimento=['Positivo','Negativo','Neutro']
print (metrics.classification_report(classes,resultados,sentimento))
```

Analisando mensagens do Twitter

- Matriz de confusão

```
# Matriz de confusão
print ("Matriz de confusão")
print (pd.crosstab(classes, resultados, rownames=['Real'],
colnames=['Predito'], margins=True))
```

Analisando mensagens do Twitter

- Resultado

	precision	recall	f1-score	support
Positivo	0.95	0.88	0.91	3300
Negativo	0.89	0.93	0.91	2446
Neutro	0.80	0.84	0.82	2453
avg / total	0.89	0.88	0.88	8199

Matriz de confusão

Predito / Real	Negativo	Neutro	Positivo	All
Negativo	2275	162	9	2446
Neutro	240	2067	146	2453
Positivo	45	356	2899	3300
All	2560	2585	3054	8199

Exercícios

1) Adaptar os códigos dos exemplos para aceitar o novo dataset, utilizar as colunas Sentiment e TweetText
   - https://github.com/zfz/twitter_corpus/blob/master/full-corpus.csv
2) Escolher no dataset pelo menos 5 instâncias de cada sentimento(positive,negative e neutral) e testar na ferramenta Ifeel. http://blackbird.dcc.ufmg.br:1210
3) Preparar um vídeo de 5 minutos mostrando como foi feito o trabalho

Referências

- Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, v.2, n.1-2, 2008
- Bin Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data- Centric Systems and Applications). Springer, 2008
- Bing Liu. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, Morgan & Claypool. 2012
- E.D. Liddy. Natural Language Processing. In Encyclopedia of Library and Information Science.2 ed. Nova York: Marcel Decker. 2001
- Notas de Aula da Prof Flavia Barros, Gabriela C. Sampaio, Roberto S. Maior, Emanuel Ferreira, Paulo R. Soares, Nelson G. Silva, Francisco Ricarte Neto, Gleibson R. S. Oliveira. Mineração de Opinião/Análise de Sentimento. Cin – UFPE. Out. 2016.
- Notas de Aula do Prof. Hansenclever Bassani. Processamento de Linguagem Natural - IF704