

Classificação de Tweets

Como Identificar Discurso de Ódio em Tweets

Ronald Albert

<https://ronaldalbertjr.github.io/>





UFRJAnalytics



Conteúdo

1. O que é Análise de Sentimento?
2. Ambiente de Desenvolvimento: Google Colab
3. Tratamento do texto e Lemmatização
4. Tokenização
5. Feature Extraction
 - Bag of Words
 - Tf-idf
6. Uma Breve Introdução ao Aprendizado de Máquina
 - Aprendizado de Máquina Supervisionado - Classificação
 - Árvores de Decisão e Florestas Aleatórias
 - Regressão Logística
 - Divisão da Base de Dados em Treino e Teste
 - Underfitting e Overfitting
 - K-Fold Cross Validation
7. Resultados



O que é Análise de Sentimento?

- A análise de sentimentos é a interpretação e classificação de emoções (positivas, negativas e neutras) nos dados de texto usando técnicas de análise de texto.
- Podemos usar a análise de sentimentos para detectar qualquer tipo de sentimento em um texto
 - Raiva
 - Alegria
 - Tristeza
 - ...
- É possível também detectar sentimentos mais gerais, muitas empresas automatizam o processo de detecção de satisfação do cliente, avaliando o sentimento do texto submetido pelo cliente como positivo ou negativo.




Análise de Sentimento: Exemplo Prático

Koi Sushi Lounge


Estr. do Bananal, 38 - Freguesia de Jacarepaguá, Rio de Janeiro - RJ


4,6 ★★★★★ 1.106 comentários ?

 **André S**
Local Guide · 257 comentários · 52 fotos

★★★★★ uma semana atrás **NOVA**


Recepção na chegada muito boa, atendimento garçons razoáveis, tive que fazer o pedido pelo menos 3 vezes para ser atendido optei pelo rodízio, porém a variedade nao e das melhores. ... [Mais](#)

 Gostei

 **camila vaz de melo**
9 comentários · 2 fotos

★★★★★ um mês atrás

Infelizmente não foi uma boa experiência. Fomos ontem, domingo, na hora do almoço e optamos pelo rodízio. As frituras estavam bem engorduradas, a lula empanada não é em anel, é uma bola estranha e borrachuda, vc fica mastigando por um bom ... [Mais](#)

 1

Resposta do proprietário um mês atrás



Detecção de Discurso Ofensivo em Tweets de Políticos

- Nesse minicurso iremos abordar o problema de Detecção de ofensas contra minorias no twitter, mais especificamente, detectaremos essas ofensas em tweets realizados por políticos.





Passo a Passo

1. Criar uma conta Google
2. Acessar <https://colab.research.google.com/>
3. Baixar <https://gist.github.com/ronaldalbertjr/8b1c8706a8ea2af97b392fc0633bf56f>
4. Extrair arquivos, importar `AnaliseDeTweetsComPython.ipynb` no Colab



Tratamento do Texto

- Na etapa de tratamento do texto, retiramos todas as palavras do texto que teoricamente não são relevantes para a análise.
- É evidente que essa etapa é diferente dependendo do tipo de sentimento e do contexto o qual se está avaliando, no entanto, alguns passos são comuns a todos os problemas.



Tratamento do Texto

- Remoção de url's
- Remoção de caracteres especiais e acentos (?).
- Remoção de stopwords

As stopwords são um conjunto de palavras de determinada língua que não "teoricamente" não oferecem muita informação ao nosso texto, geralmente são palavras muito constantes em determinada língua e por esse motivo não oferecem muita discriminação sobre o texto.

Exemplos no português: [eu, não, de, a, o, que]



Lematização

- Tem como objetivo reduzir uma palavra à sua forma base e agrupar diferentes formas da mesma palavra. Por exemplo, os verbos no tempo passado são alterados para alguma forma comum, geralmente o infinitivo, (por exemplo, “foi” é alterado para “ir”) e palavras com a mesma raiz (por exemplo, “melhor” é alterado para “bom”), padronizando palavras com significado semelhante à sua raiz.
- Existem bibliotecas no python que já possuem esse mapeamento entre palavras pré-estabelecido e realizam esse trabalho para nós.

Exemplos: Spacy, nltk



Tokenização

- A tokenização ou segmentação de palavras, é basicamente transformar o texto em uma mera lista de termos.

Exemplo: O texto **"Oi, eu sou o Ronald"** se transforma em ["Oi", ",", "eu", "sou", "o", "Ronald"]

- Alguns problemas surgem no processo de tokenização, o mais clássico deles são termos compostas, que apesar de serem formados por duas ou mais palavras devem ser um único token, além disso endereços também devem ser um único token.

- Exemplos de possíveis termos problemáticos na tokenização:
Pé de moleque, Rua Voluntários da Pátria 59



Tokenização no Twitter

- Realizar a tokenização com textos da internet é um desafio um pouco mais complicado do que com textos formais.
- O caráter desafiador da tokenização é agravado ainda mais, quando consideramos que a rede social objetivo é o Twitter.
- Para resolver esse problema usaremos um tokenizador específico da biblioteca nltk, chamado TweetTokenizer.



Feature Extraction

- Features são informações mensuráveis acerca de algum fenômeno, em outras palavras, são uma maneira estruturada de organizar dados.
- Texto não possuem uma estrutura organizada o suficiente para ser entendida por um algoritmo de Aprendizado Estatístico.
- Por isso realizamos a ‘feature extraction’ do texto criando ‘features’ que teoricamente o caracterizam.



Bag of Words

- No meio do caminho tinha uma pedra tinha uma pedra no meio do caminho...
- João amava Teresa que amava Raimundo que amava Maria que amava Joaquim que amava Lili..

Abaixo podemos ver a base de dados gerada usando o Bag of Words como método de feature extraction para cada um dos poemas.

	acontecimento	amava	caminho	casou	com	convento	de	desastre	desse	do	...	retinas	se	suicidou	teresa	tia	tinha	tão	uma	unidos	vida	
0		1	0	6	0	0	0	1	0	1	6	...	1	0	0	0	0	7	1	7	0	1
1		0	6	0	1	1	1	1	1	0	0	...	0	1	1	2	1	1	0	0	1	0

2 rows × 47 columns



Tf-idf

- Tf-idf significa Term Frequency - Inverse Document Frequency
- Cada 'token' ou termo do texto gerado pela tokenização se torna uma feature, e a cada um dos tokens é associado uma 'pontuação', dados pela fórmula do Tf-idf

$$w_{x,y} = tf_{x,y} \cdot \log \left(\frac{N}{df_i} \right)$$

- $w_{x,y}$ é a "pontuação" do token x no texto y .
- $tf_{x,y}$ é a frequência de ocorrências do token x no texto y .
- N é o número de textos.
- df_x é o número de textos que contém o token x .



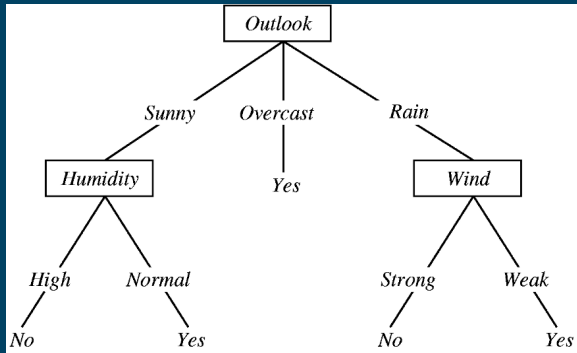
Uma Breve Introdução ao Aprendizado de Máquina

- No aprendizado de máquina podemos ter aprendizado Supervisionado, Não-Supervisionado e por Reforço.
- Neste minicurso focaremos em técnicas de Aprendizado Supervisionado, quando conhecemos a classe que queremos prever e fazemos a máquina aprender a prever tal classe.
- Mais especificamente, introduziremos métodos de Aprendizado Supervisionado para classificação, que é quando as nossas amostras estão classificadas.



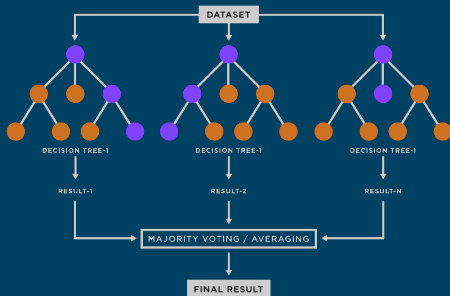
Árvores de Decisão

O algoritmo de construção da Árvore de Decisão, constrói uma árvore considerando as características mais importantes para a classificação.



Florestas Aleatórias

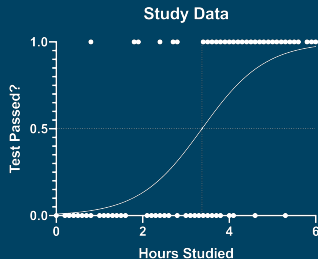
- Árvores de Decisão são extremamente interpretáveis, no entanto, possuem pouco poder preditivo.
- A ideia por trás de Florestas Aleatórias é o de gerar diferentes conjuntos de dados a partir do conjunto original de maneira aleatória, e assim, usar cada um desses conjuntos gerados para construir uma Árvore de Decisão diferente, construindo uma espécie de "comitê" de Árvores de Decisão.



Regressão Logística

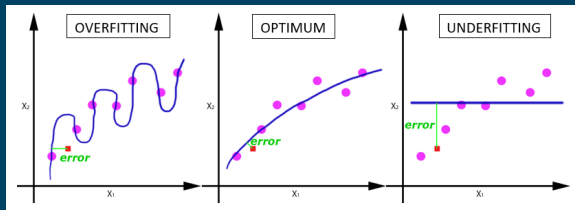
$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}$$

- Onde p_i é a probabilidade da amostra x pertencer à classe i .
- Maximizamos a função de verossimilhança para encontrar o vetor β



Underfitting e Overfitting

- Dois problemas muito comuns quando trabalhamos com aprendizado de máquina são o Overfitting e o Underfitting.
- Underfitting acontece quando o modelo que escolhemos para representar o fenômeno que queremos descrever não é representativo o suficiente.
- Overfitting acontece quando o modelo é representativo demais e por isso aprende muito bem, mas não possui tanta capacidade de generalização



Todo o modelo de Aprendizado de máquina está propenso ao overfitting, e por isso é perigoso usar o mesmo conjunto de dados para treinar e testar o modelo. Dessa forma, procuramos sempre dividir o nosso conjunto de dados entre treino e teste.



K-Fold Cross Validation

- Para cada um dos k experimentos: $k - 1$ folds para treinar e 1 para testar.
- Todos os elementos são usados para treinamento e teste.



Resultados

