Fake News Detection.

Nesse projeto utilizei Jupyter, qualquer outra forma de "IDE" funcionaria.

Introdução.

Classificação.

No campo do machine learning, classificação refere-se a um tipo de modelo preditivo, o qual tenta prever a classe de um conjunto de dados. Um exemplo é o modelo de spam, o algoritmo analisa o conteúdo do email e decide o que ou não spam

---

CountVectorizer.

Vamos utilizar uma biblioteca do Sklearn chamada CountVectorizer. Vamos converter nosso texto em um dicionário que mapeia cada palavra única para o número de vezes que ela aparece nos dados.

Naive Baynesr.

A probabilidade de o evento A acontecer, dado o evento B acontecer. O Teorema de Bayes calcula a probabilidade de que A seja verdadeiro dado o evento B com base na probabilidade inversa de B dado A. O método Naive Bayes para classificação de texto é muito popular porque pode ser escalado com muita facilidade. Naive Bayes assume independência condicional entre cada par de recursos, isso significa que estamos presumindo que as palavras em um artigo de notícias não têm impacto umas sobre as outras; estamos apenas examinando a probabilidade de ver cada palavra receber uma notícia falsa ou real

In [1]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score
```

In [2]:
```
import pandas as pd
# Carregando os dois datasets
fake = pd.read_csv("Fake.csv")
true = pd.read_csv("True.csv")
```

In [3]:
```
#Carregando os 'heads' dos datasets
fake.head()
```

Out[3]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017

	title	text	subject	date
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

In [4]:
```
#Carregando os 'heads' dos datasets
true.head()
```

Out[4]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

In [5]:
```
# No head temos title, text, a subject, and date published
# Vamos flagear os conjuntos real = 0 fake = 1, coloquei uma coluna com 0 e 1 para a
```

In [6]:
```
true["fake_news"] = 0
fake["fake_news"] = 1
```

In [7]:
```
true.head(0)
fake.head(0)
#Criação de uma coluna
```

Out[7]:

title	text	subject	date	fake_news

In [8]:
```
true.head()
```

Out[8]:

	title	text	subject	date	fake_news
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	0
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	0

	title	text	subject	date	fake_news
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	0

Overfitting

Todos os artigos começam com Reuters, se o modelo for treinado desta forma o modelo aprenderia que todos os inputs com Reuters seriam verdadeiros e os que nao tivesse essa marca seria falso, isso é chamado Overfitting

In [ ]:

In [9]:
```
just_text = true["text"]
just_text.head()
```

Out[9]:
```
0    WASHINGTON (Reuters) - The head of a conservat...
1    WASHINGTON (Reuters) - Transgender people will...
2    WASHINGTON (Reuters) - The special counsel inv...
3    WASHINGTON (Reuters) - Trump campaign adviser ...
4    SEATTLE/WASHINGTON (Reuters) - President Donal...
Name: text, dtype: object
```

In [10]:
```
#Vou corrigir isso removendo todas as infos iniciais para tornar os textos dos dataf
#Para fazer isso, o pandas tem uma função útil chamada extractall () que aceita um p
#Regex é uma sequência especial de caracteres que define um padrão de pesquisa.
#Vamos extrair todo o texto que vem após o hífen que vem depois da Reuters.
```

In [11]:
```
just_text = just_text.str.extractall(r"^.*? - (?P<text>.*)")
#https://pandas.pydata.org/docs/reference/api/pandas.Series.str.extractall.html
```

In [12]:
```
just_text = just_text.droplevel(1)
#https://www.w3resource.com/pandas/series/series-droplevel.php
```

In [49]:
```
true = true.assign(text=just_text["text"])
```

In [50]:
```
#Juntando os dois dataset

df = pd.concat([fake, true], axis = 0)
```

In [51]:
```
#Não vou utilizar essas colunas no treinamento
df = df.drop(["subject", "date", "title"], axis = 1)
```

In [52]:
```
df.info()
df = df.dropna(axis = 0)
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44898 entries, 0 to 21416
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
```

```
 0   text       44869 non-null  object
 1   fake_news  44898 non-null  int64
dtypes: int64(1), object(1)
memory usage: 1.0+ MB
```

In [53]:

```
#Criaçao de um novo dataframe

clean_text = df.to_csv("cleaned_news.csv", index = False)
```

In [18]:

```
import pandas as pd
df = pd.read_csv("cleaned_news.csv")
df.head()
```

Out[18]:

	text	fake_news
0	Donald Trump just couldn t wish all Americans ...	1
1	House Intelligence Committee Chairman Devin Nu...	1
2	On Friday, it was revealed that former Milwauk...	1
3	On Christmas day, Donald Trump announced that ...	1
4	Pope Francis used his annual Christmas Day mes...	1