- ## Introdução

Bitcoins e outras criptomoedas estão sendo muito exploradas no mundo atualmente e sua variação tem um valor elevado.

Juntamente com o mundo conectado que vivemos, podemos acompanhar em redes sócias informações on-line sobre bitcoins, informações essas que nos ajuda tomar decisões se esta na hora de entrar nesse mercado ou não.

Por esse motivo, me senti motivado a minerar o Tweeter em busca uma ajuda para identificar se o que estão falando no momento sobre bitcoin é bom ou ruim e possivelmente utilizar essa informação para acompanhar o mercado de Bitcoins para poder tomar uma decisão de quando será o melhor momento para a compra dessa criptomoeda.

- ## Resumo da proposta

Primeiramente irei capturar em dias diferentes uma quantidade razoável de Tweets que contenham a palavra bitcoin. Após a captura desses Tweets irei analisa-los e classifica-los manualmente como Positivo, Negativo ou Neutro.

Feita essa classificação irei testar algoritmos de classificação que tragam o melhor F1-Score para prever a classificação. Para isso irei executar 2 ou 3 modelos de classificação para treinamento em cima de 70-80% dos Tweets, usar um Cross-Validation e também Grid Search para melhor ajustar meu modelo.

Tweets com a palavra Bitcoin Classificados previamente

Treinamento e Teste

Tweets Novos

Previsão

Correlação entre a classificação prevista e o valor do Bitcoin

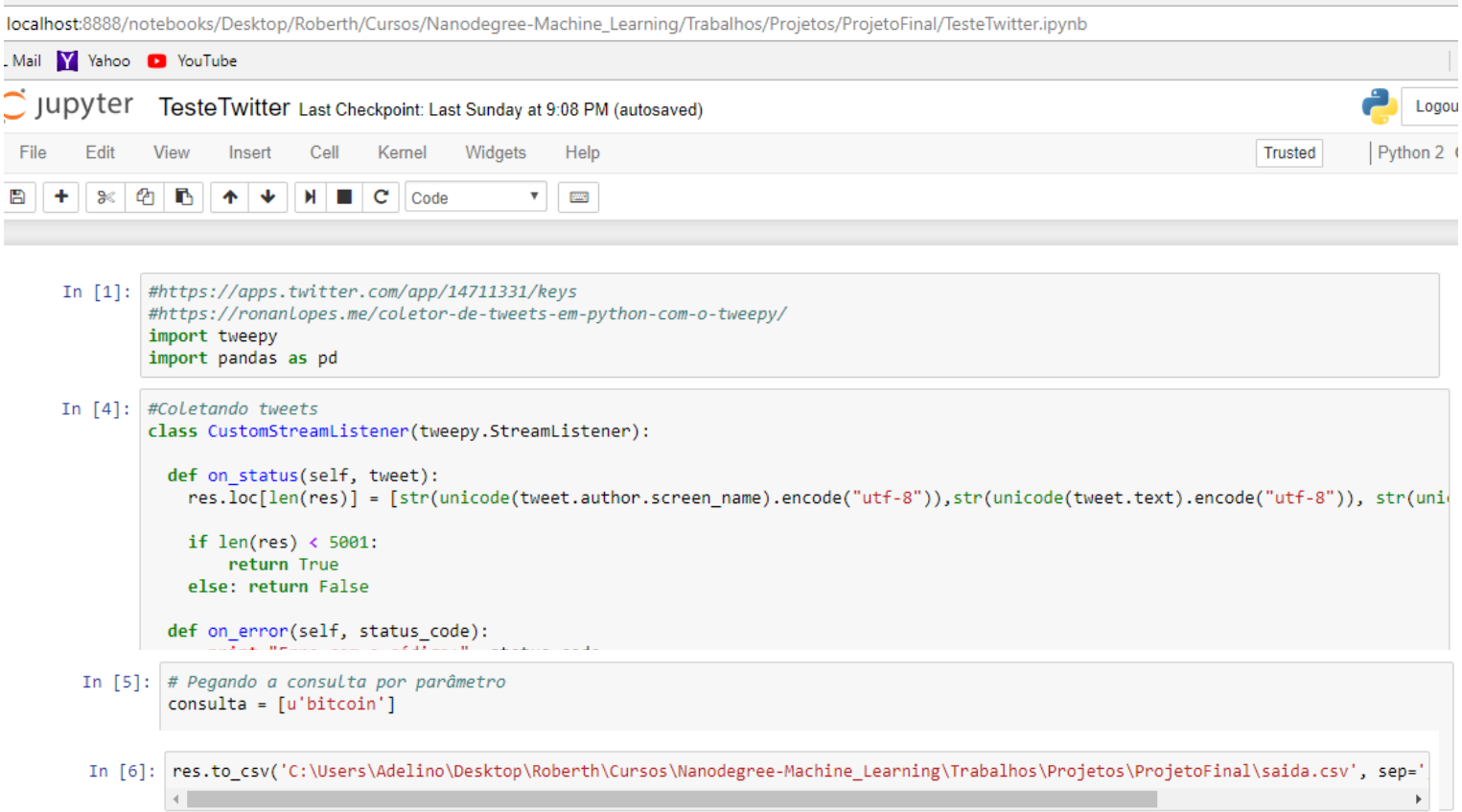
- **Conjunto de dados**

Para realização desse projeto usarei os dados coletados do Tweeter.
As informações necessárias para isso são:

- **USR**
  - Nome do usuário do Tweeter que publicou esse Tweet
- **TWEET**
  - O texto publicado pelo usuário, o qual sempre terá a palavra bitcoin.
- **LANG**
  - Linguagem em que o Tweet foi publicado
- **DATE**
  - Data de publicação do Tweet
- **CLAS**
  - Classificação que eu darei para esse Tweet, sendo: Positiva, para as publicações que me influenciariam a comprar bitcoin: Negativa, para as publicações que me influenciariam a não comprar bitcoin: Neutra, para as publicações que não tem influencia na compra de bitcoin.

Para captura dos dados utilizei o código abaixo criado por mim, mas usando como referência:

https://apps.twitter.com/app/14711331/keys
https://ronanlopes.me/coletor-de-tweets-em-python-com-o-tweepy/



```
In [2]: data = pd.read_csv('C:\Users\Adelino\Desktop\Roberth\Cursos\Nanodegree-Machine_Learning\Trabalhos\Projetos\ProjetoFinal\saida.csv
```

```
In [3]: data.count()

Out[3]: USR      5003
        TWEET    5002
        LANG     5000
        dtype: int64
```

```
In [4]: data['LANG'].count()

Out[4]: 5000
```

```
In [5]: data.loc[data['LANG']=='en'].count()

Out[5]: USR      4856
        TWEET    4856
        LANG     4856
        dtype: int64
```

```
In [6]: data.loc[data['LANG']=='pt'].count()

Out[6]: USR      144
        TWEET    144
        LANG     144
        dtype: int64
```

```
In [7]: pt = data.loc[data['LANG']=='pt']
        en = data.loc[data['LANG']=='en']
```

```
In [8]: print pt['TWEET'].head()
        print en['TWEET'].head()

        4      RT @FUTEMENTALES: Boto o pal\r\nCompra e venda...
        76     RT @marcurelio: Como é definido o valor do Bit...
        88     RT @FUTEMENTALES: Boto o pal\r\nCompra e venda...
        108    Não sei pq lembrei de ti @mobilon https://t.co...
        137    RT @FUTEMENTALES: Boto o pal\r\nCompra e venda...
        Name: TWEET, dtype: object
        0    RT @roomieofficial: Let me explain Bitcoin to ...
        1    RT @ruppomanrup: #cryptocurrency Rupee $RUP - ...
        2    RT @RciNext: Common Standards And Interoperabi...
        3    @jamesdpitley @bccponzi So you're not invested...
        5    RT @symmetryfund: "The only place where succes...
        Name: TWEET, dtype: object
```

- ## Treinamento do Modelo

Para execução do projeto, após a coleta dos dados do Tweeter, classificação dos Tweet em Positivo, Negativo e Neutro, irei tirar algumas métricas como estatística descritiva, verificar a necessidade da criação de um Bag of Words (com stemming e stop words por exemplo), depois executar 2 ou 3 modelos de classificação que são aconselhados para text mining, como por exemplo Guassian Naive Bayes, Multinomial Naive Bayes, SVM e Random Forest, e por fiz realizar avaliações e validações sobre esse modelo.

- ## Métricas de Avaliação e validação com um Benchmark

Por fim, mesmo após avaliar e validar meu modelo com os dados de teste, irei capturar dados novos, prever sua classificação com meu modelo, verificar a porcentagem de positivos e negativos e compara-la com o verdadeiro valor do bitcoin para aquele dia e o dia seguinte.

A ideia é verificar se quando a maior porcentagem foi de positivos essa mesma percepção também refletiu no valor do bitcoin no dia e no dia seguinte. Dessa forma terei como verificar na pratica a eficiência em auxiliar a tomada de decisão de compra do meu modelo.

Uma forma de realizar essa comparação seriam tirar uma correlação entre a Classificação do modelo e o valor do bitcoin, analisados no mesmo período. Com isso posso medir se existe uma relação entre as publicações do Tweeter e o valor do Bitcoin, seja essa relação positiva (publicações positivas e valores subindo) ou negativas (publicação positivas e valores caindo).