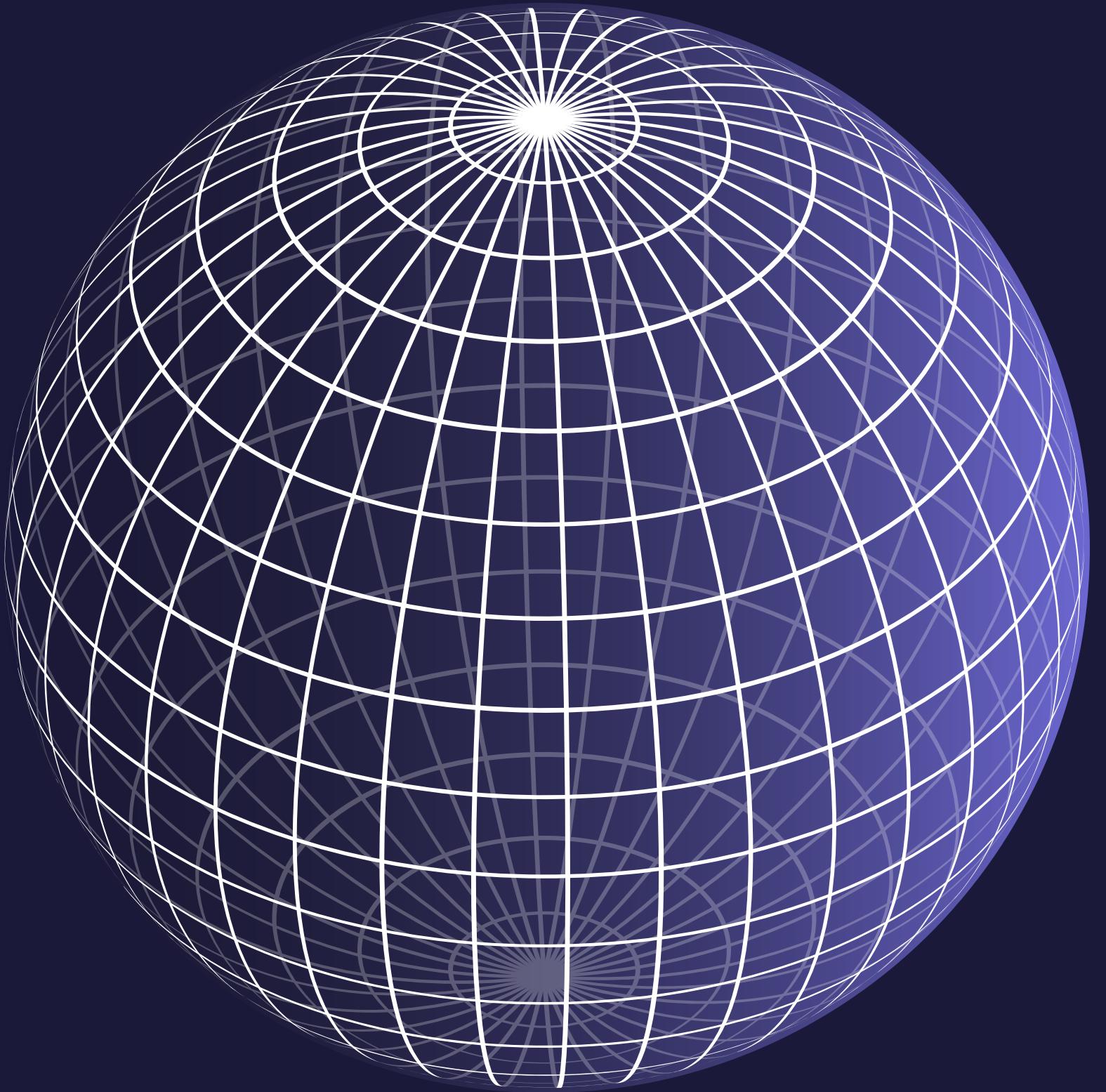


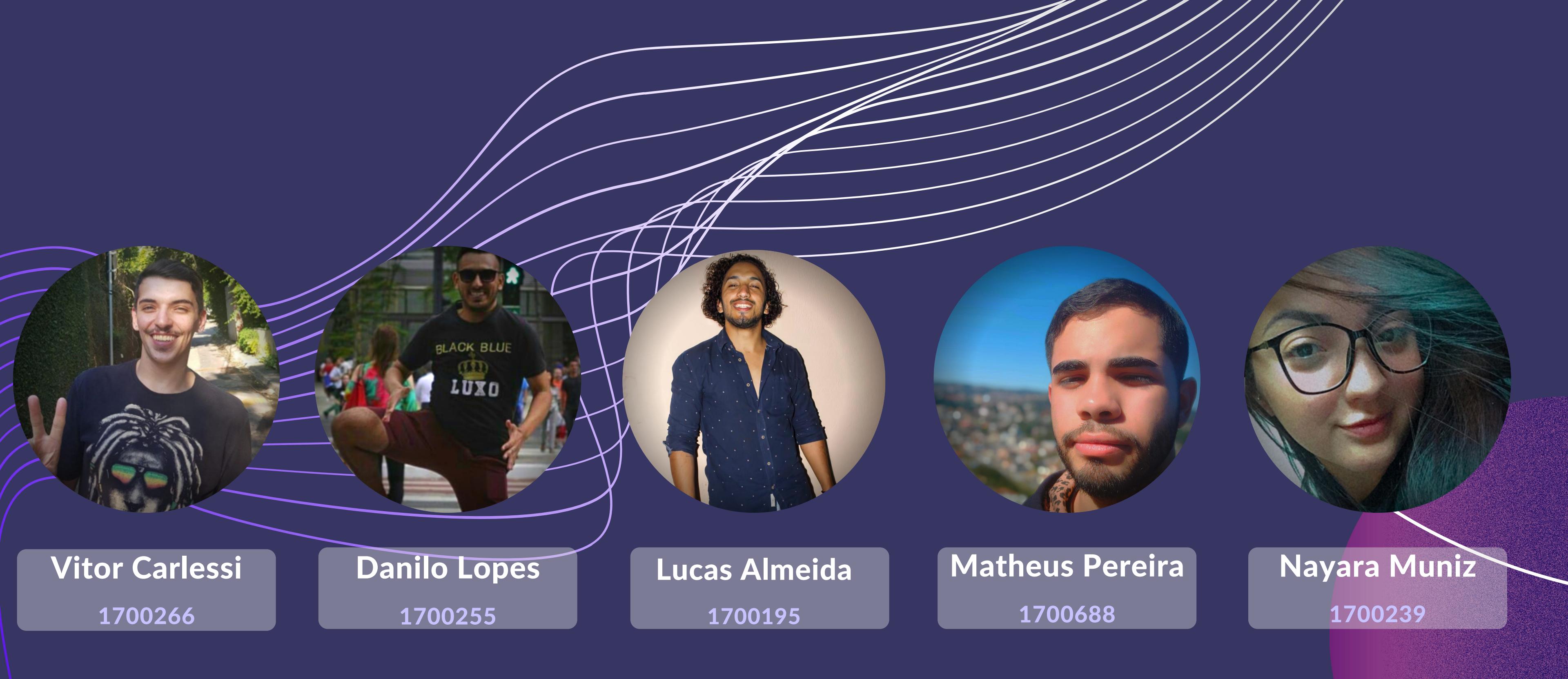
# COMPUTAÇÃO COGNITIVA

PROF. ROBERTO ÂNGELO



FIT - Faculdade Impacta de Tecnologia

Outubro de 2020



**Vitor Carlessi**

1700266

**Danilo Lopes**

1700255

**Lucas Almeida**

1700195

**Matheus Pereira**

1700688

**Nayara Muniz**

1700239

# INTEGRANTES



# Mini-Projeto da Disciplina

## TEMA

Mineração de textos e análise de sentimentos em mídias sociais.

## BASE

A base de dados será a rede social twitter. Por intermédio do developers twitter, foi criado um APP para consumo da API do twitter em código python pela utilização da biblioteca tweepy é possível realizar o consumo dessa API para trazer postagens dos usuários.

## METODOLOGIA CRISP-DM

Para esse projeto, iremos seguir a metodologia CRISP-DM que é dividida em 6 fases.

1. Entendimento do Negócio
2. Entendimento dos Dados
3. Preparação dos dados
4. Modelagem
5. Avaliação de Desempenho
6. Distribuição (Implantação)



# DESCRICAÇÃO DA SOLUÇÃO

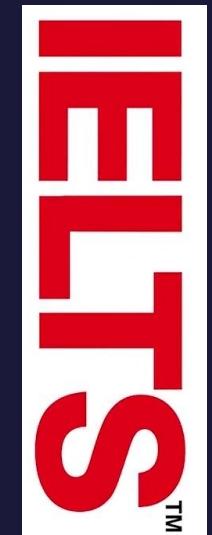
Inicialmente será criado um arquivo python para consumo da API do twitter procurando os últimos 450 tweets com a palavra #TOEFL e listando em um arquivo CSV em uma coluna o nome do usuário no twitter que fez aquele tweet e o conteúdo do tweet em outra coluna. O mesmo será feito para a prova IELTS, então também teremos um arquivo CSV com os últimos 450 tweets com a palavra #IELTS e isso também será listado em uma coluna contendo o nome do usuário no twitter que fez o post e na outra coluna o tweet que foi feito. Com os dois CSVs gerados, um segundo arquivo python irá ler esses dois datasets, criar a coluna Sentimento e classificar o tweet como positivo, neutro ou negativo com base na polaridade do mesmo. Também será criado uma coluna "Subjetividade", que irá classificar o tweet como objetivo, subjetivo ou neutro.

Ao término, o programa irá mostrar ao usuário qual a melhor opção das duas provas e os resultados obtidos.

# ENTENDIMENTO DO NEGÓCIO



Devido a pandemia do corona vírus em 2020, um casal de imigrantes ingleses que moram no Brasil tiveram seu salário reduzido pela metade, para conseguir complementar renda, resolveram lançar um curso extensivo em certificações da língua inglesa para estudantes brasileiros chamado #EasyCertification, pois são nativo no idioma e também já trabalharam em escolas de estudo da língua inglesa no Brasil. O problema encontrado pelo casal foi justamente em qual certificação focar, ficaram em dúvida em basicamente dois exames: IELTS(requisito acadêmico, corporativo e que pode ser aceito para fins de imigração. É reconhecido principalmente, por universidades no Reino Unido, Austrália, Canadá e Nova Zelândia. Inclui redação, interpretação de texto, compreensão auditiva e expressão oral) e TOEFL(permite avaliar as competências em língua inglesa principalmente em contextos acadêmicos), o casal ficou em dúvida pois viu que os dois se equivalem em vários quesitos e para critério de desempate acharam relevante consultar qual está sendo melhor falado nas mídias sociais, principalmente na plataforma do Twitter. Com isso contrataram a #GetPubli, consultoria especializada em mineração de texto e análise de sentimentos em redes sociais para conseguir ajudar a empresa #EasyCertification nessa análise.



VS



# ENTENDIMENTO DO NEGÓCIO

## QUAL CONTEXTO DA APLICAÇÃO?

- Qual área?  
Marketing Digital
- Qual empresa ?  
#EasyCertification, uma nova empresa no mercado especializada em cursos preparatórios para provas de certificação na língua inglesa.
- Qual setor ?  
Educação, cursos preparatórios para certificação.

## QUAL O PROBLEMA DE NEGÓCIO A SER RESOLVIDO?

O problema a ser resolvido é ajudar a #EasyCertification a escolher a melhor prova de certificação para lançarem seu primeiro curso. Como a quantidade de professores é curta, apenas duas pessoas, eles precisam focar naquela que tem maior popularidade/engajamento para obter um número grande de alunos.

## QUAL OU QUAIS TAREFAS DEVEMOS UTILIZAR ?

- Consumo da API do Twitter - listagem dos últimos tweets das provas
- Analise dos sentimentos desses tweets
- Comparação dos resultados entre as provas
- Exibição da melhor escolha

## QUAIS OS CRITERIOS DE SUCESSO DO PROJETO?

Os critérios de sucesso do projeto é uma análise bem feita listando o número de tweets positivos, negativos e neutros de cada base de dados e também mostrando qual a melhor opção para o cliente.

# ENTENDIMENTO DOS DADOS

Twitter

A base de dados será a rede social twitter. Por intermédio do developers twitter, foi criado um APP para consumo da API do twitter em código python e pela utilização da biblioteca tweepy é possível realizar o consumo dessa API para trazer postagens dos usuários.

 Developer



Developers

Tap into  
what's  
happening.

**Publish and analyze Tweets,  
optimize ads, and create  
unique customer  
experiences.**

# PREPARAÇÃO DOS DADOS

## Twitter API e Tweepy

Para obtenção dos dados, foi utilizado a API do Twitter.

Inicialmente foi criado um APP no developer.twitter .

Para consumo da APP criada, foi utilizada a biblioteca tweepy.

The screenshot shows the Twitter Developer Keys and tokens page. At the top, it displays the app icon (a blue gear with a white bird), the app name 'getpubli', the App ID '16387823', a 'Details' button, and a three-dot menu icon. Below this, under 'Keys and tokens', there are sections for 'Consumer API keys' and 'Access token & access token secret'. In the 'Consumer API keys' section, the API key is listed as 'wJsWaCxR2ismW20UpDQ8YObqV' and the API secret key as '8qRJTzCBqHhh5ZXBNWbODCRBLArwzCRIZqx2Lce2N9CvYfG3c'. There is a 'Regenerate' button next to the secret key. In the 'Access token & access token secret' section, the access token is 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx' and the access token secret is 'xx'. The access level is listed as 'Read and write'. A 'Revoke' button and a 'Regenerate' button are present. A note at the bottom states: 'We only show your access token and secret when you first generate it in order to make your account more secure. You can revoke or regenerate them at any time, which will invalidate your existing tokens.' At the bottom of the page, a code snippet imports the tweepy library: `import tweepy`.

# PREPARAÇÃO DOS DADOS

## Autenticação Twitter API

O código a lado tem a função de utilizar a biblioteca tweepy para autenticação na APP criada. As variáveis consumerKey, consumerSecret, accessToken e accessTokenSecret são para a autenticação e todas foram geradas pela APP no developer.twitter.

```
consumerKey      = "wJsWaCxR2ismW20UpDQ8YObqV"  
consumerSecret   = "8qRJTzCBqHhh5ZXBNWbODCRBLaArwzCR1Zqx2Lce2N9CvYfG3c"  
accessToken      = "1106181670821154817-VigfLv5XYLS3ZTR9WSFbmiwFwQBD9Q"  
accessTokenSecret = "dhdkuYTBPiElam6bVz7qpoRiHmhNFSAjxGYWsUtjH1qgL"  
  
auth = tweepy.OAuthHandler(consumer_key=consumerKey, consumer_secret=consumerSecret)  
auth.set_access_token(accessToken, accessTokenSecret)  
api = tweepy.API(auth)
```

# PREPARAÇÃO DOS DADOS

## Query de busca no twitter

Nessa parte do código, fazemos a busca no twitter pelos 500 últimos twitters com a palavra #TOEFL no idioma "en". Mais a frente repetimos a mesma busca para a palavra #IELTS

```
#Termo de busca será no twitter será #IELTS
searchTerm = "#IELTS" + "-filter:retweets" ##Filtrando os reetwets -> não pegar o mesmo reetweet mais de uma vez

#Tweepy.cursor irá buscar no twitter os últimos 450 twittes com a palavra #IELTS
tweets = tweepy.Cursor(api.search, q=searchTerm, show_user = True, lang='en').items(450)
```

```
#Termo de busca será no twitter: #TOEFL
searchTerm = "#TOEFL" + "-filter:retweets" ##Filtrando os reetwets -> não pegar o mesmo reetweet mais de uma vez

#Tweepy.cursor irá buscar no twitter os últimos 450 twittes com a palavra #TOEFL
tweets = tweepy.Cursor(api.search, q=searchTerm, show_user = True, lang='en').items(450)
```

# PREPARAÇÃO DOS DADOS

## Criação do CSV

Já nessa parte do código, rodamos todos os tweets encontrados e fazemos uma lógica para obter o usuário que fez aquele tweet e o conteúdo do tweet, colocamos em um arquivo CSV os resultados. Ao término teremos os arquivos IELTS\_tweets.csv e TOEFL\_tweets.csv

```
#Criação da lista de tweets
listTweets = []

#appendando a primeira linha que saira no .CSV, com a coluna Usuário e Tweet
row_list = [["Usuario","Tweet"]]

#percorrendo todos os tweets encontrados
for tweet in tweets:
    #appendando o nome do usuário e o tweet feito pelo mesmo na lista
    listTweets.append(list((tweet.user.screen_name, tweet.text)))

for x in listTweets:
    row_list.append(x)

#Codificações para escrever a saida no arquivo .CSV IELTS_tweets.csv
csv.register_dialect('myDialect',
                      delimiter=',',
                      quoting=csv.QUOTE_ALL)
with open('IELTS_tweets.csv', 'w', encoding='utf-8-sig', newline='') as file:
    writer = csv.writer(file, dialect='myDialect')
    writer.writerows(row_list)
```

```
#Criação da lista de tweets
listTweets = []

#appendando a primeira linha que saira no .CSV, com a coluna Usuário e Tweet
row_list = [["Usuario","Tweet"]]

#percorrendo todos os tweets encontrados
for tweet in tweets:
    #appendando o nome do usuário e o tweet feito pelo mesmo na lista
    listTweets.append(list((tweet.user.screen_name, tweet.text)))

for x in listTweets:
    row_list.append(x)

#Codificações para escrever a saida no arquivo .CSV TOEFL_tweets.csv
csv.register_dialect('myDialect',
                      delimiter=',',
                      quoting=csv.QUOTE_ALL)
with open('TOEFL_tweets.csv', 'w', encoding='utf-8-sig', newline='') as file:
    writer = csv.writer(file, dialect='myDialect')
    writer.writerows(row_list)
```

# PREPARAÇÃO DOS DADOS

## Upload dos CSVs gerados

Em outro arquivo python, agora utilizando o spyder, foi feito o upload dos dois datasets.

```
## Carregando os dados dos arquivos gerados TOEFL_tweets.csv e IELTS_tweets.csv
##Arquivo TOEFL_tweets.csv
dataset_TOEFL = pd.read_csv('TOEFL_tweets.csv')

##Arquivo IELTS_tweets.csv
datasetIELTS = pd.read_csv('IELTS_tweets.csv')
```

Nome	Tipo	Tamanho	Valor
datasetIELTS	DataFrame	(450, 2)	Column names: Usuario, Tweet
dataset_TOEFL	DataFrame	(450, 2)	Column names: Usuario, Tweet

Índice	Usuario	Tweet
0	Nexgeneduserve	#AdvancedEnglish: #CAE #CPE #IELTS #LearnEnglish #ingles #toe...
1	IELTS_T20	#IELTS Reading Practice - Job Satisfaction <a href="https://t.co/ky8C29CyXK">https://t.co/ky8C29CyXK</a>
2	iambotbatibot	#AdvancedEnglish: #CAE #CPE #IELTS #LearnEnglish #ingles #toe...
3	Ftips_Resources	#AdvancedEnglish: #CAE #CPE #IELTS #LearnEnglish #ingles #toe...
4	MobinaDiary	A. I always avoid participating in polls.
5	talkaruenglish	To Gain Confidence in English and find your...
6	MagooshEnglish	An idiom is an expression that means someth...
7	Ibrahimaliyuab5	It seems to me that... In my humble opinion
8	iaykhan786	#ielts life skills(A1-B1)-ielts- A1 spouse ... IELTS WRITING DRAFT & DISCUSS TWO STANCES LESSON...
9	IELTS_Expert_9	#ielts <a href="https://t.co/SOEVu17EoM">https://t.co/SOEVu17EoM</a> Check out this offer below in case you're i...

# MODELAGEM

## Inserção de novas colunas no Dataset

Foi inserido as coluna Sentimento no datasets, ela recebeu o valor de não classificado por enquanto e vai receber o valor de neutro, positivo ou negativo decorrente da sua classificação. Também foi inserido a coluna Subjetividade, que irá receber o valor objetivo, subjetivo ou neutro.

```
##Coluna sentimento no dataset_TOEFL
dataset_TOEFL['Sentimento'] = 'Não classificado'

##Coluna sentimento no datasetIELTS
datasetIELTS['Sentimento'] = 'Não classificado'

##Coluna subjetividade no dataset_TOEFL
dataset_TOEFL['Subjetividade'] = 'Não classificado'

##Coluna subjetividade no datasetIELTS
datasetIELTS['Subjetividade'] = 'Não classificado'
```

# MODELAGEM

## Análise de sentimento pela classificação da polaridade e subjetividade

o código a seguir tem a função de percorrer todos os tweets e pela utilização da biblioteca TextBlob verificar a polaridade de cada tweet e com base no valor que retornar dessa polaridade e classificar em neutro, positivo ou negativo. Também pela utilização da biblioteca TextBlob, classificamos os tweets como subjetivos, objetivos ou neutros.

```
## Lógica para avaliar o sentimento como Positivo, Neutro ou Negativo no arquivo TOEFL_tweets
for index, row in dataset_TOEFL.iterrows():
    analysis      = TextBlob(row['Tweet'])
    #Classificação de Polaridade -> TOEFL
    if (analysis.sentiment.polarity == 0):
        dataset_TOEFL.loc[index, 'Sentimento'] = 'Neutro'
        neutral_TOEFL = neutral_TOEFL + 1
    elif (analysis.sentiment.polarity < 0):
        dataset_TOEFL.loc[index, 'Sentimento'] = 'Negativo'
        negative_TOEFL = negative_TOEFL + 1
    elif (analysis.sentiment.polarity > 0):
        dataset_TOEFL.loc[index, 'Sentimento'] = 'Positivo'
        positive_TOEFL = positive_TOEFL + 1

## Lógica para avaliar a subjetividade como Neutro, Objetivo ou Subjetivo no arquivo TOEFL_tweets
#Classificação de Subjetividade -> TOEFL
if     (analysis.sentiment.subjectivity == 0.5):
    dataset_TOEFL.loc[index, 'Subjetividade'] = 'Neutro'
    sub_neutral_TOEFL = sub_neutral_TOEFL + 1
elif   (analysis.sentiment.subjectivity < 0.5):
    dataset_TOEFL.loc[index, 'Subjetividade'] = 'Objetivo'
    sub_objective_TOEFL = sub_objective_TOEFL + 1
elif   (analysis.sentiment.subjectivity > 0.5):
    dataset_TOEFL.loc[index, 'Subjetividade'] = 'Subjetivo'
    sub_subjectivity_TOEFL = sub_subjectivity_TOEFL + 1
```

# AVALIAÇÃO DE DESEMPENHO

Escolha do melhor exame  
IELTS ou TOEFL com base nos  
sentimentos obtivos

Com as quantidades de sentimentos neutros, positivos e negativos de cada exame armazenados em variáveis, a lógica a seguir compara e verifica qual a prova mais bem conceituada. Leva em consideração qual teve maior número de tweets positivos, caso haja empate nesse quesito, escolhe a que teve menos tweets negativos

```
##Lógica para comparação da melhor opção
if (positiveIELTS == positiveTOEFL) and (negativeIELTS < positiveTOEFL):
    melhor_opcao = 'IELTS'

if (positiveIELTS == positiveTOEFL) and (positiveTOEFL < negativeIELTS):
    melhor_opcao = 'TOEFL'

if (positiveIELTS > positiveTOEFL):
    melhor_opcao = 'IELTS'
else:
    melhor_opcao = 'TOEFL'
```

# DISTRIBUIÇÃO (IMPLEMENTAÇÃO)

Lógica para exibição dos resultados obtidos

```
-----Exame IELTS-----  
Quanto a polaridade:  
O número de tweets positivos em IELTS_tweets.csv é: 178  
O número de tweets negativos em IELTS_tweets.csv é: 35  
O número de tweets neutros em IELTS_tweets.csv é: 237  
Quanto a subjetividade:  
O número de tweets subjetivos em IELTS_tweets.csv é: 69  
O número de tweets objetivos em IELTS_tweets.csv é: 337  
O número de tweets neutros em IELTS_tweets.csv é: 44  
-----Exame IELTS-----  
-----Exame TOEFL-----  
Quanto a polaridade:  
O número de tweets positivos em TOEFL_tweets.csv é: 132  
O número de tweets negativos em TOEFL_tweets.csv é: 44  
O número de tweets neutros em TOEFL_tweets.csv é: 274  
Quanto a subjetividade:  
O número de tweets subjetivos em TOEFL_tweets.csv é: 55  
O número de tweets objetivos em TOEFL_tweets.csv é: 358  
O número de tweets neutros em TOEFL_tweets.csv é: 37  
-----Exame TOEFL-----
```

```
-----Resultado-----  
O Exame IELTS é o melhor conceituado no Twitter!!  
Tem 178 tweets positivos, 35 tweets negativos e 237 tweets neutros  
Quanto a subjetividade, tem 69 tweets subjetivos, 337 tweets objetivos e 44 tweets neutros  
O segundo colocado é o exame TOEFL  
Tem 132 tweets positivos, 44 tweets negativos e 274 tweets neutros  
Quanto a subjetividade, tem 55 tweets subjetivos, 358 tweets objetivos e 37 tweets neutros  
-----Resultado-----
```