Twitter Sentiment Analysis: Web Interface to Scrape Tweets and Perform Sentiment Analysis

Prepared by-

Harsh Jain(180102013)
Sachin Verma(180101038)
Uday Pratap Singh(180101046)

Overview :

The project intends to create a web interface which can be used to extract tweets in the form of a required format which can be further used for various analyses. We used Tweepy which is an open source Python package that gives you a very convenient way to access the Twitter API with Python.

The Twitter API gives developers access to most of Twitter’s functionality. One can use the API to read and write information related to Twitter entities such as tweets, users, and trends.

Data that can be extracted

Technically, the API exposes dozens of HTTP endpoints related to:

- Tweets
- Retweets
- Likes
- Direct messages
- Favourites
- Trends
- Media

Tweepy provides a way to invoke those HTTP endpoints without dealing with low-level details.

Motivation :

Twitter is one of the most widely used social networks. For many organizations and people, having a great Twitter presence is a key factor to keeping their audience engaged. Twitter can as over 500 million tweets are exchanged on its platform daily – a huge percent of these being text, then followed by images, then videos. For most researchers, tweets made of text are quite important for their social research, which could be used for sentimental analysis, text classification, and for some kinds of predictive analysis. For these researchers collecting and formatting required data which are Tweets in this case can be a tedious task. This Tool which creates pdf as well as csv file of Tweets for given hashtag, can be very helpful and handy to use for these researchers, so they can focus more on more important aspects of their research.

Getting Started:

The first thing to do is get the consumer key, consumer secret, access key and access secret from twitter developer available easily for each user. These keys will help the API for authentication.

Steps to obtain keys:

– Login to twitter developer section

– Go to “Create an App”

– Fill in the details of the application.

– Click on Create your Twitter Application

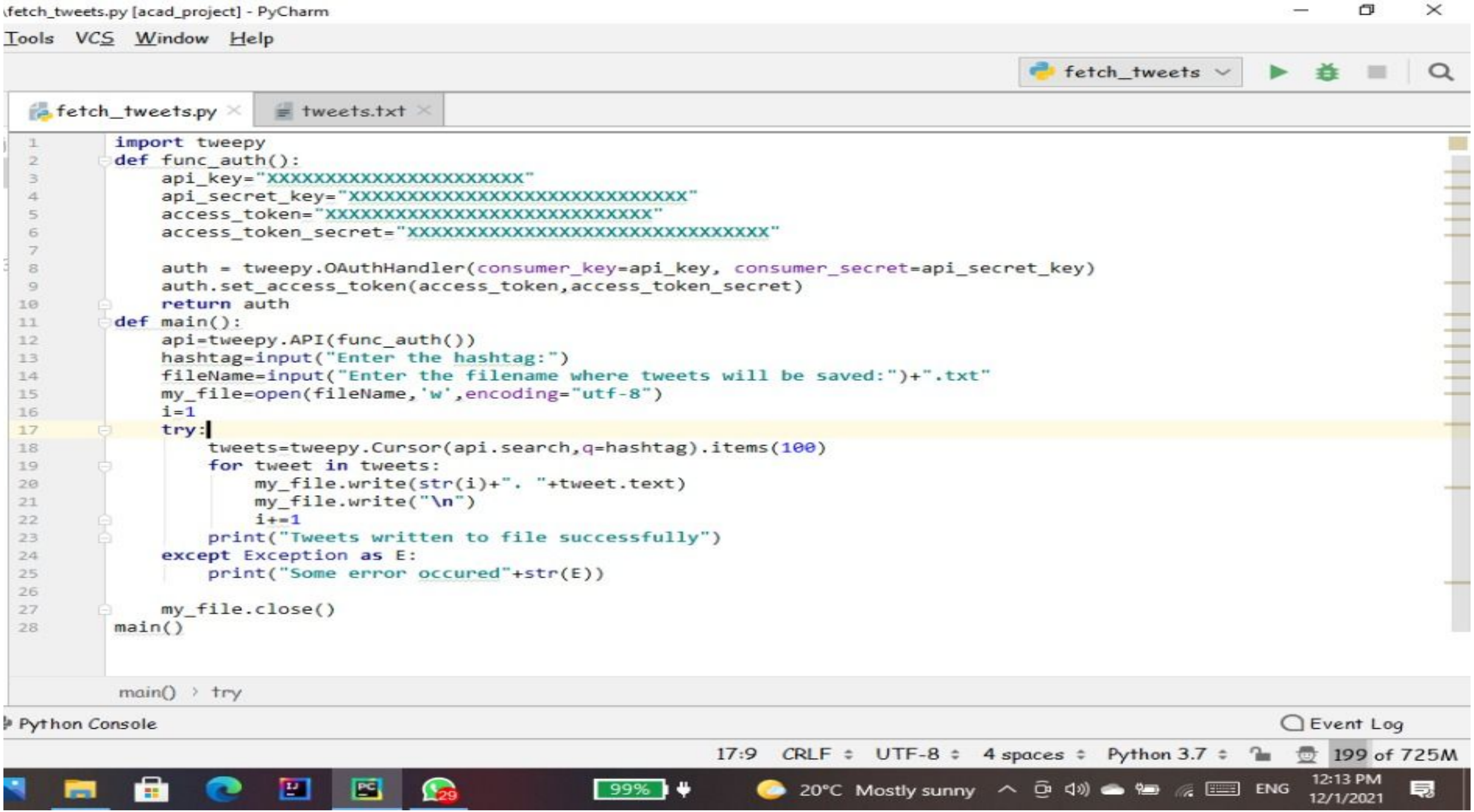
– Details of your new app will be shown along with consumer key and consumer secret.

– For access token, click ” Create my access token”. The page will refresh and generate access tokens.

Tweepy

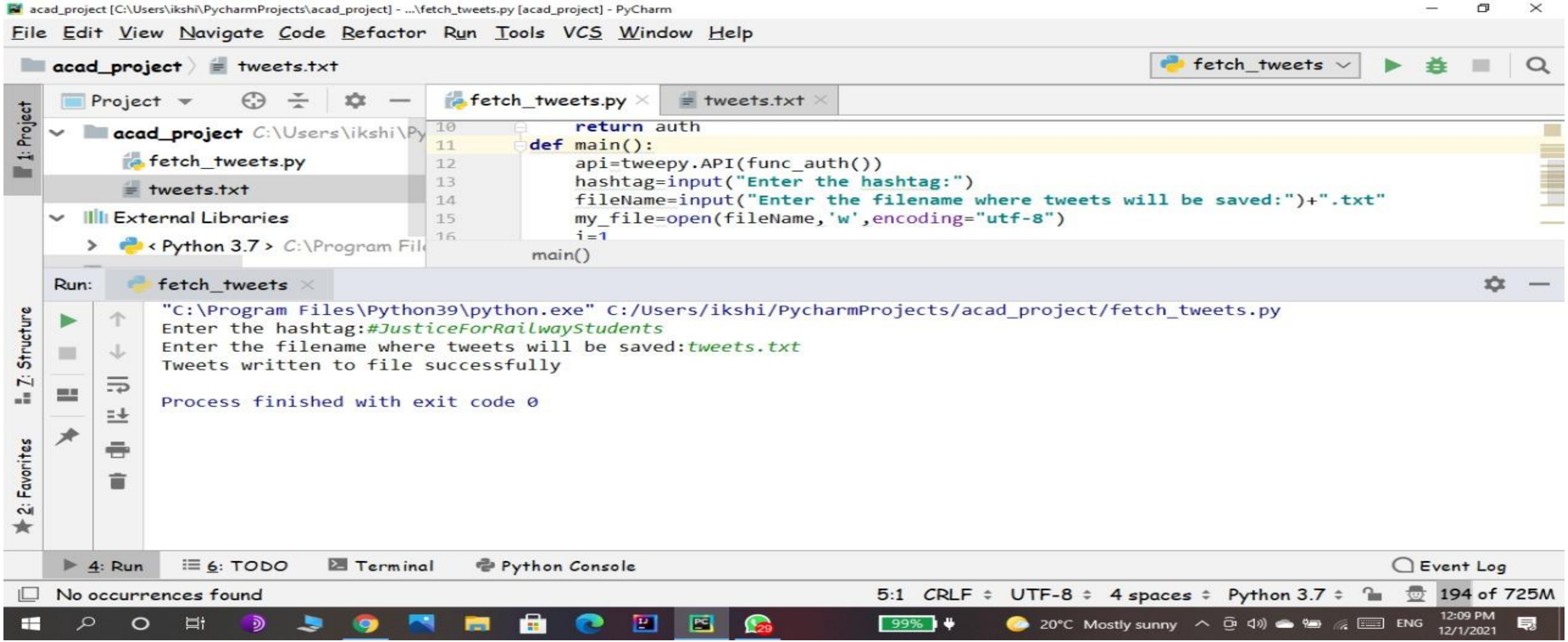
Tweepy is one of the libraries that should be installed using pip. Now in order to authorize our app to access Twitter on our behalf, we need to use the OAuth Interface. Tweepy provides the convenient Cursor interface to iterate through different types of objects. Twitter allows a maximum of 3200 tweets for extraction.

Code



```python
import tweepy
def func_auth():
    api_key="XXXXXXXXXXXXXXXXXXXXXX"
    api_secret_key="XXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
    access_token="XXXXXXXXXXXXXXXXXXXXXXXXXXXX"
    access_token_secret="XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX"

    auth = tweepy.OAuthHandler(consumer_key=api_key, consumer_secret=api_secret_key)
    auth.set_access_token(access_token,access_token_secret)
    return auth
def main():
    api=tweepy.API(func_auth())
    hashtag=input("Enter the hashtag:")
    fileName=input("Enter the filename where tweets will be saved:")+".txt"
    my_file=open(fileName,'w',encoding="utf-8")
    i=1
    try:
        tweets=tweepy.Cursor(api.search,q=hashtag).items(100)
        for tweet in tweets:
            my_file.write(str(i)+". "+tweet.text)
            my_file.write("\n")
            i+=1
        print("Tweets written to file successfully")
    except Exception as E:
        print("Some error occured"+str(E))

    my_file.close()
main()
```

Output

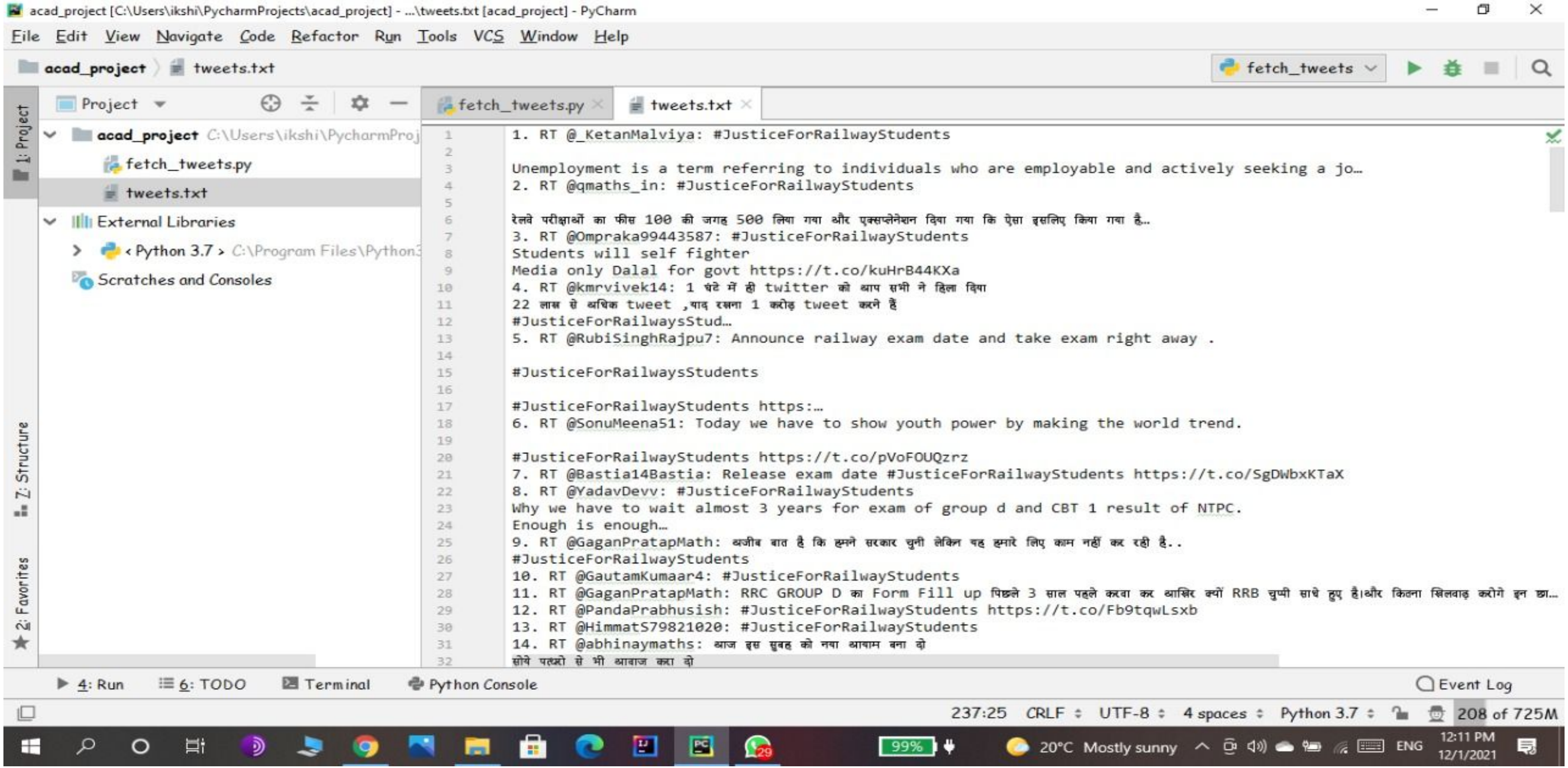


acad_project [C:\Users\ikshi\PycharmProjects\acad_project] - ...\fetch_tweets.py [acad_project] - PyCharm
File Edit View Navigate Code Refactor Run Tools VCS Window Help
acad_project
tweets.txt
fetch_tweets
1: Project
Project
acad_project C:\Users\ikshi\Py
fetch_tweets.py
tweets.txt
External Libraries
< Python 3.7 > C:\Program Fil
fetch_tweets.py
tweets.txt
main()
7: Structure
2: Favorites
Run:
fetch_tweets
4: Run
6: TODO
Terminal
Python Console
Event Log
No occurrences found
5:1 CRLF UTF-8 4 spaces Python 3.7 194 of 725M
99%
20°C Mostly sunny
ENG
12:09 PM
12/1/2021

```
    return auth
def main():
    api=tweepy.API(func_auth())
    hashtag=input("Enter the hashtag:")
    fileName=input("Enter the filename where tweets will be saved:")+".txt"
    my_file=open(fileName,'w',encoding="utf-8")
    i=1
```

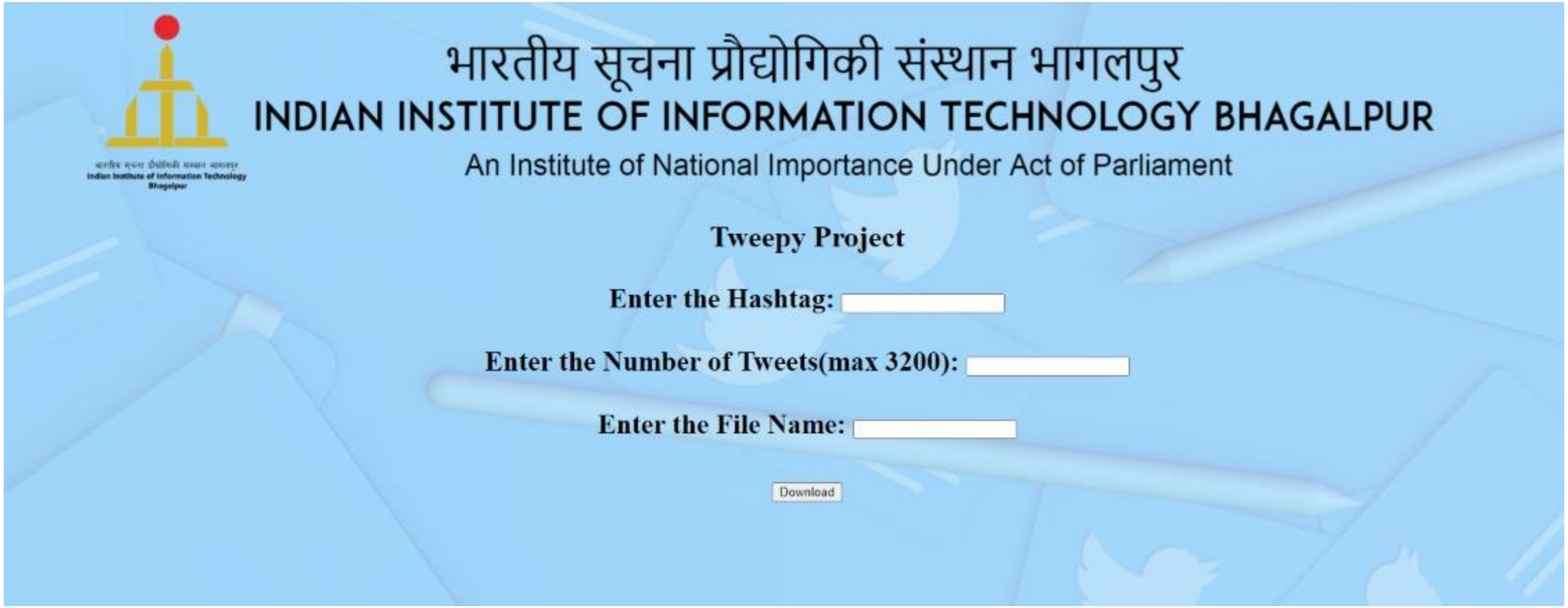
```
"C:\Program Files\Python39\python.exe" C:/Users/ikshi/PycharmProjects/acad_project/fetch_tweets.py
Enter the hashtag:#JusticeForRailwayStudents
Enter the filename where tweets will be saved:tweets.txt
Tweets written to file successfully

Process finished with exit code 0
```

Scrapped Tweets


acad_project [C:\Users\ikshi\PycharmProjects\acad_project] - ...\tweets.txt [acad_project] - PyCharm
File Edit View Navigate Code Refactor Run Tools VCS Window Help
acad_project
tweets.txt
fetch_tweets
Project
acad_project C:\Users\ikshi\PycharmProj
fetch_tweets.py
tweets.txt
External Libraries
< Python 3.7 > C:\Program Files\Python3
Scratches and Consoles
fetch_tweets.py
tweets.txt
1: Project
7: Structure
2: Favorites
4: Run
6: TODO
Terminal
Python Console
Event Log
237:25 CRLF UTF-8 4 spaces Python 3.7 208 of 725M
99%
20°C Mostly sunny
ENG
12:11 PM
12/1/2021

```
1. RT @_KetanMalviya: #JusticeForRailwayStudents

Unemployment is a term referring to individuals who are employable and actively seeking a jo…
2. RT @qmaths_in: #JusticeForRailwayStudents

रेलवे परीक्षाओं का फीस 100 की जगह 500 लिया गया और एक्सप्लेनेशन दिया गया कि ऐसा इसलिए किया गया है…
3. RT @Ompraka99443587: #JusticeForRailwayStudents
Students will self fighter
Media only Dalal for govt https://t.co/kuHrB44KXa
4. RT @kmrvivek14: 1 घंटे में ही twitter को आप सभी ने हिला दिया
22 लाख से अधिक tweet ,याद रखना 1 करोड़ tweet करने हैं
#JusticeForRailwaysStud…
5. RT @RubiSinghRajpu7: Announce railway exam date and take exam right away .

#JusticeForRailwaysStudents

#JusticeForRailwayStudents https:…
6. RT @SonuMeena51: Today we have to show youth power by making the world trend.

#JusticeForRailwayStudents https://t.co/pVoFOUQzrz
7. RT @Bastia14Bastia: Release exam date #JusticeForRailwayStudents https://t.co/SgDWbxKTaX
8. RT @YadavDevv: #JusticeForRailwayStudents
Why we have to wait almost 3 years for exam of group d and CBT 1 result of NTPC.
Enough is enough…
9. RT @GaganPratapMath: अजीब बात है कि हमने सरकार चुनी लेकिन वह हमारे लिए काम नहीं कर रही है..
#JusticeForRailwayStudents
10. RT @GautamKumaar4: #JusticeForRailwayStudents
11. RT @GaganPratapMath: RRC GROUP D का Form Fill up पिछले 3 साल पहले करवा कर आखिर क्यों RRB चुप्पी साधे हुए है।और कितना खिलवाड़ करोगे इन छा…
12. RT @PandaPrabhusish: #JusticeForRailwayStudents https://t.co/Fb9tqwLsxb
13. RT @HimmatS79821020: #JusticeForRailwayStudents
14. RT @abhinaymaths: आज इस सुबह को नया आयाम बना दो
सोये पत्थरो से भी आवाज करा दो
```

Web Interface



भारतीय सूचना प्रौद्योगिकी संस्थान भागलपुर

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHAGALPUR

An Institute of National Importance Under Act of Parliament

Tweepy Project

Enter the Hashtag:

Enter the Number of Tweets(max 3200):

Enter the File Name:

Conclusion:

The above script would generate all the tweets of the particular hashtag and would be appended to the empty array tmp which is then saved in a txt file in the above code. The file format can be changed to any other such as csv,etc. Here Tweepy is introduced as a tool to access Twitter data in a fairly easy way with Python. There are different types of data we can collect, with the obvious focus on the “tweet” object. Once we have collected some data, the possibilities in terms of analytics applications are endless.

One such application of extracting tweets is sentiment or emotion analysis. The emotion of the user can be obtained from the tweets by tokenizing each word and applying machine learning algorithms on that data. Such emotion or sentiment detection is used worldwide and will be broadly used in the future.

Sentiment Analysis

“Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in the text source”, in other words we can say that “Sentiment analysis is the process of automatic extraction of writer’s opinions and their characterization in terms of polarity: positive, negative and neutral”.

BERT stands for Bidirectional Encoder Representations from Transformers and it is a state-of-the-art machine learning model used for NLP tasks. Jacob Devlin and his colleagues developed BERT at Google in 2018.
BERT is a transformers model pre-trained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts.

Setup

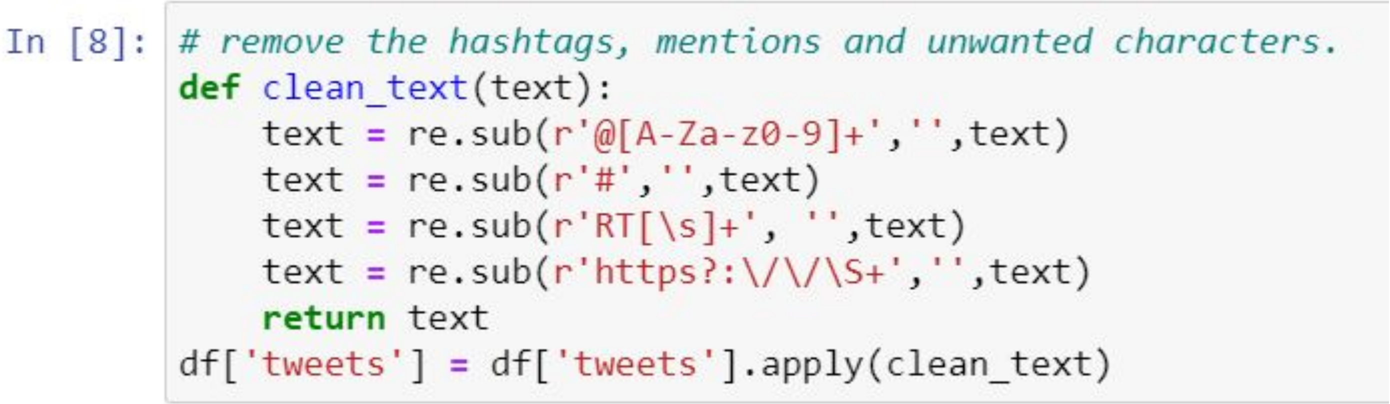
We’ll need the Transformers library by Hugging Face:

```
import transformers
```

```
from transformers import BertModel, BertTokenizer, AdamW,
get_linear_schedule_with_warmup
```

PreProcessing of Tweets

Tweets contains unnecessary objects like hashtags, mentions, links and punctuation that can affect the performance of an algorithm thus they have to be rid off. All the texts are converted to lower case to avoid algorithms interpreting same words with different cases as different

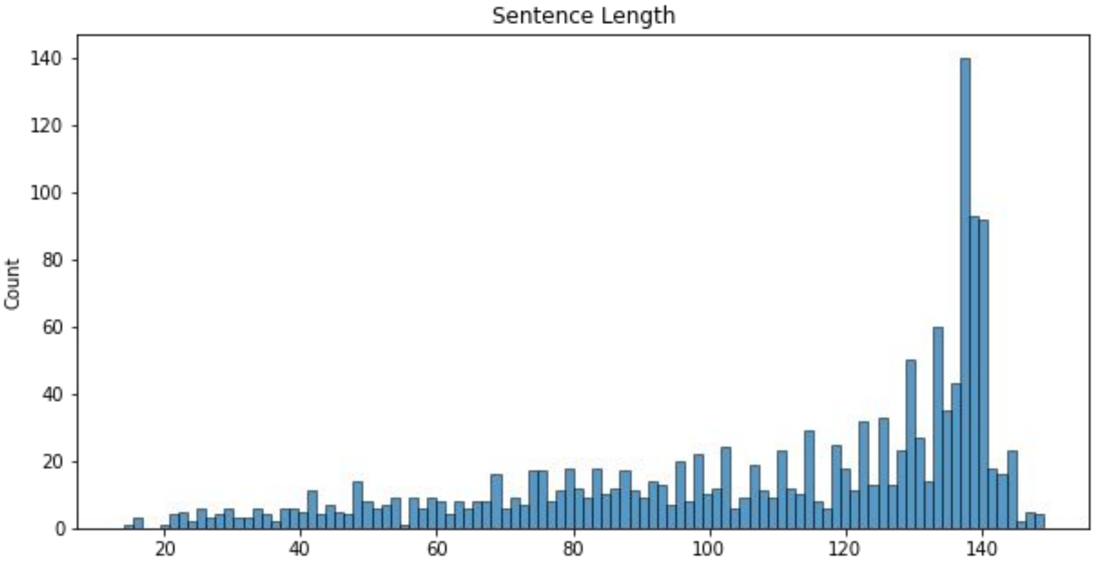


```
# remove the hashtags, mentions and unwanted characters.
def clean_text(text):
    text = re.sub(r'@[A-Za-z0-9]+','',text)
    text = re.sub(r'#','',text)
    text = re.sub(r'RT[\s]+', '',text)
    text = re.sub(r'https?:\/\/\S+','',text)
    return text
df['tweets'] = df['tweets'].apply(clean_text)
```

Choosing sequence length

BERT works with fixed-length sequences.

We'll use a simple strategy to choose the max length.
We can also see the token length of each tweet by plotting the distribution.



Training

We will use the SMILE Twitter dataset. This dataset is collected and annotated for the SMILE project. This collection of tweets mentioning 13 Twitter handles associated with British museums was gathered between May 2013 and June 2015. It was created for the purpose of classifying emotions, expressed on Twitter towards arts and cultural experiences in museums.

It contains 3,085 tweets, We need to split our dataset into a train and val set using train_test_split. We’ll use 85% of the dataset as the training data and evaluate the performance on the remaining 15%

To reproduce the training procedure from the BERT paper, we'll use the AdamW optimizer provided by Hugging Face.The BERT authors have some recommendations for fine-tuning:

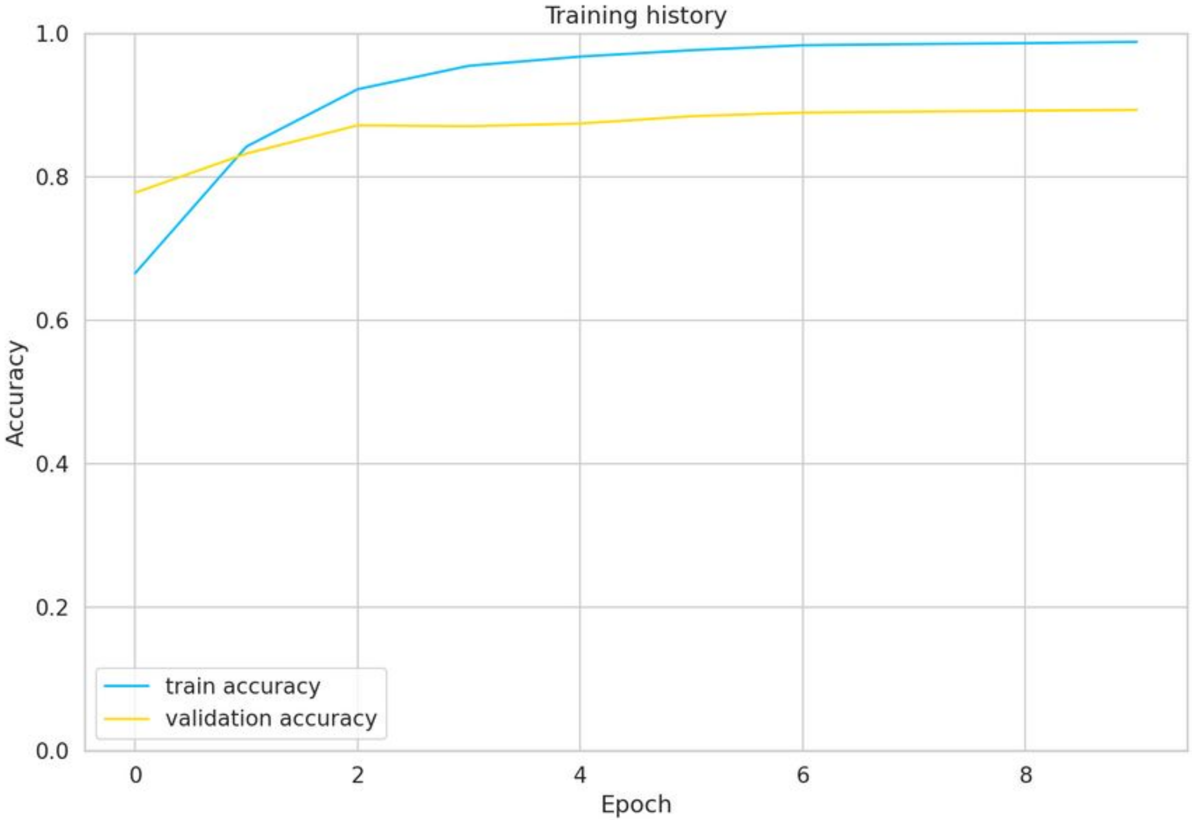
Batch size: 16, 32

Learning rate (Adam): 5e-5, 3e-5, 2e-5

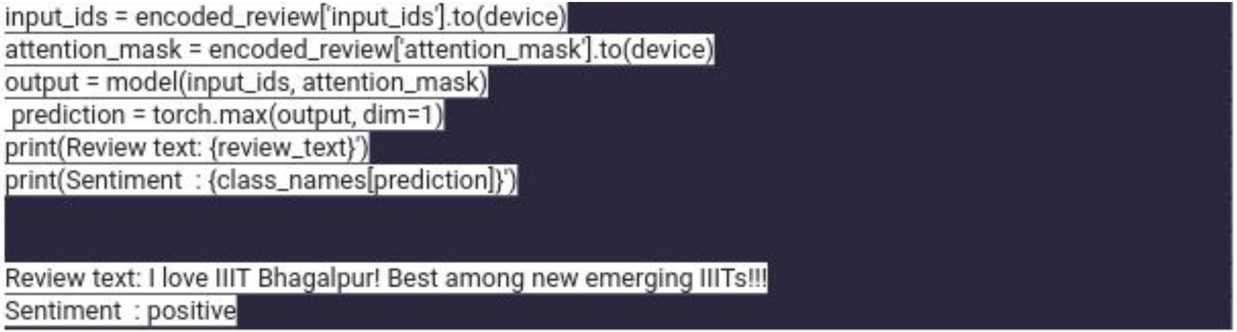
Number of epochs: 2, 3, 4

Note that increasing the batch size reduces the training time significantly, but gives you lower accuracy.

The training accuracy starts to approach 100% after 10 epochs or so.


Training history
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
2
4
6
8
Epoch
train accuracy
validation accuracy

Result



```
input_ids = encoded_review['input_ids'].to(device)
attention_mask = encoded_review['attention_mask'].to(device)
output = model(input_ids, attention_mask)
_prediction = torch.max(output, dim=1)
print(Review text: {review_text}')
print(Sentiment  : {class_names[prediction]}')


Review text: I love IIIT Bhagalpur! Best among new emerging IIITs!!!
Sentiment  : positive
```

Evaluation

Since the classes were imbalanced, the resulted predictions are also imbalance. We achieved a 87% accuracy score overall.

The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important. Especially with sentiment analysis where we want to know the honest and constructive feedbacks from the community to fix mistakes and make improvements on our product and service,

Future Work

- Correct class imbalance.
- Deploy our trained model behind a REST API and build a simple web app to access it.

THANK YOU