Documentation

Architecture:

- We create a TCP socket between Twitter's API and Spark, which waits for the call of the Spark Structured Streaming and then sends the Twitter data.
- We receive the data from the TCP socket and preprocess it with the pyspark library, which is Python's API for Spark.
- We pass the data to RESTful API where the model is located.
- We apply sentiment analysis using fine-tuned BERT-Base-Uncased and return Positive, Negative, or Neutral.
- Finally, we save the tweet and the sentiment analysis polarity in a parquet file.

Steps to Run The Application:

- Run pip 'install -r requirements.py' command.
- Download the saved model from this link into the same path of *.py files ('03_Scripts' directory).
- Go to '03_Scripts' directory.
- Run 'chmod u+x run_server' command.
- Run './run_server' command.
- Type the keyword you want in the 'keyword.txt' file.
- In another terminal, run ' python twitter_connection.py'.
- In another terminal, run ' python sentiment_analysis.py'.
- After a reasonable amount of time, terminate the running of 'twitter_connection.py' and 'sentiment_analysis.py'.
- In the end run, 'get_insights.py' to get some insights about the keyword like how many tweets included that keyword and if the public is positive, negative, or neutral about it.

Evaluation:

- Test set (unseen data): the model achieved 86.8% accuracy on the test set and 86.3% on the validation set.
- Streaming data (keyword: Harry Potter):

```
There are 64 tweets streamed.
62.50% of the tweets are Positive
25.00% of the tweets are Negative
12.50% of the tweets are Neutral
```

- Streaming data (keyword: Interstellar):

```
There are 13 tweets streamed.
92.31% of the tweets are Positive
7.69% of the tweets are Neutral
```

Language Extention:

- We can fine-tune BERT-BASE-MultiLingual instead of BERT-BASE as the former is trained on 102 languages.
- We can also still use BERT-Base but we should add a neural machine translation model in the pipeline to translate different languages to English, then we can use BERT.
- To do sentiment analysis for the Arabic language only we can fine-tune AraBERT model instead of BERT.