**Introduction:**

This Read Me file contains the various approaches of the NLP project done by me.

It also explains the reasons behind making a particular decision and the method of approach.

**Difficulties and Approaches:**

The Question and its understanding was already a bit confusing as the education level did not guide through the exposures of the project requirement.

***The question had a statement:***

***“Share a CSV with extracted entities and the frequency of the extracted entity from all the tweets in the following format”***

***Data:***

***Tweet1 by Author1: Pink Pearl Apples are tasty but Empire Apples are not.***

***Tweet2 by Author2: Empire Apples are very tasty.***

***Tweet3 by Author3: Pink Pearl Apples are not tasty.***

***Tweet4 by Author1: Pink Pearl Apples smells really good.***

***Output:***

***Pink Pearl Apples : 2***

***Empire Apples : 2***

If we see in the first line apples are of frequency two and my guess was as such and stuck in the second line of the data.

In output the second states empire apples are 2 so I had an idea to check the entire data set but still with a confusion that the pink pearl apples were 3 in freq.

The Second problem had the biggest problem were we have to set the author names and also the overall sentiment which is not for the extracted entities but the overall sentiment.

In terms of approach I have done the 2nd question 1st and the 1st question 2nd.

I have used the spacy library to find the entities and also before extracting the entities I have basically done the sentiment analysis.

There was also a difficulty in removing the URLs from the data I have anyway used a ‘Reg-ex’ to set the rules and conditions to resolve the problem.

**Techniques and Reason Behind it:**

The first library was the “gdown”, it is a library to import the files of a google drive without having to connect to the drive. If we have the id that is more than enough for the collab to download and use the file.

The regular libraries of NLTK and WordNet to tag the parts of speech and also the stop words to remove the un-necessary words. The word net is for lemmatizing to the real form of the words with their POS as the sentiment is easily accessible.

The sentiment Analysis was done through Vader’s sentiment analysis as it provides 4 factors analysing the text (i.e.) Negative, Neutral, Positive and Compound scores.

So based on the best scores we set it to the data. This analysis was actually done on the text after lemmatizing rather than doing on the extracted entities as they have no good meaning or result.

The entity was removed using the Named Entity Recognition (NER) which is a part of the library Spacy.

Additionally what I have learned in this project is the chunk library in Python. I have not used it but learned its purpose and how it works like ‘Reg-ex’, where we set grammatical rules to select the entities.

**Shortcomings and Mistakes:**

The data had a lot of other languages too but still they might have meaning full insights in the sentiment analysis part and we also need not lose some info from the data. But I think we are supposed to separate the different languages to the different data frames and do the analysis individually so the insights are much clearer.

The ‘en’ library could have done the requirement but I wanted to leave the other language info which is still a rectifiable case.