**NLP Class Assignment 1**

**Team Members** - Soumya Bhandari, Amanullah Shareef, Syed Araib Karim

**Github link** - <https://github.com/SoumyaBhandari/NLP>

**Coding Exercise:** Write a code to tokenize the text and grab contents from a webpage where you can find information about SpaceX. Use stopwords and use the above strategy to calculate the frequency. The graph should display the first 10 high distribution words in the webpage while ignoring others. If the frequency of the word is less the 5 times ignore those words as well.

**Questions:**

1. Why we use stopwords? Why stopwords are not necessary for NLP frequency distribution.

Ans- According to English languages, stopwords plays a vital role in forming sentences and speaking. But they do not have meaning. Stop words are the commonly used words. In NLP, it is important to eliminate stopwords. If the sentence does not contain any stopwords, the amount of memory required for storing is reduced. The fundamental goal of employing natural language processing is to extract meaningful terms from text that will aid NLP models. We may simply get rid of them by keeping a list of terms you regard to be stop words. In NLP we use Natural Language Toolkit (NLTK) for removing stop words.

For example – “This is my Class Assignment”

In the above sentence there are three stopwords like ‘this’, ‘is’, ‘my’. We can see that there is only two important words. Through this we can efficiently use memory.

1. Based on high frequency words what information you can extract from the graph?

Ans- According to the graph, the ten most common are listed there. It signifies that these ten keywords appear repeatedly on the xSpace webpage. As a result, they occur frequently.

The highest frequency words are those that describe the topic of the website.

1. Can you provide different visualization for frequency distribution? If yes, please perform. If no, why?

Ans - Yes, additional graphs/plots for the frequency distribution are possible. There are numerous plots available, including bar graphs, pie charts, histograms, and scatter plots. The most appropriate graph for this type of data, in my opinion, will be a bar graph. Due to the fact that the bar graph is

* It can be easily understood
* Mainly used for frequency distribution
* Summarize a big quantity of information in a graphic