ASSIGNMENT REPORT

NATURAL LANGUAGE PROCESSING WITH PYTHON

SYED ARSALAN HABIB

SEAT NO : B17101108

24-07-20

OVERVIEW

This assignment consisted of working with the tweets dataset from Zong, and using the popular NLP library 'NLTK' to clean up the data and ready it for usage in machine learning.

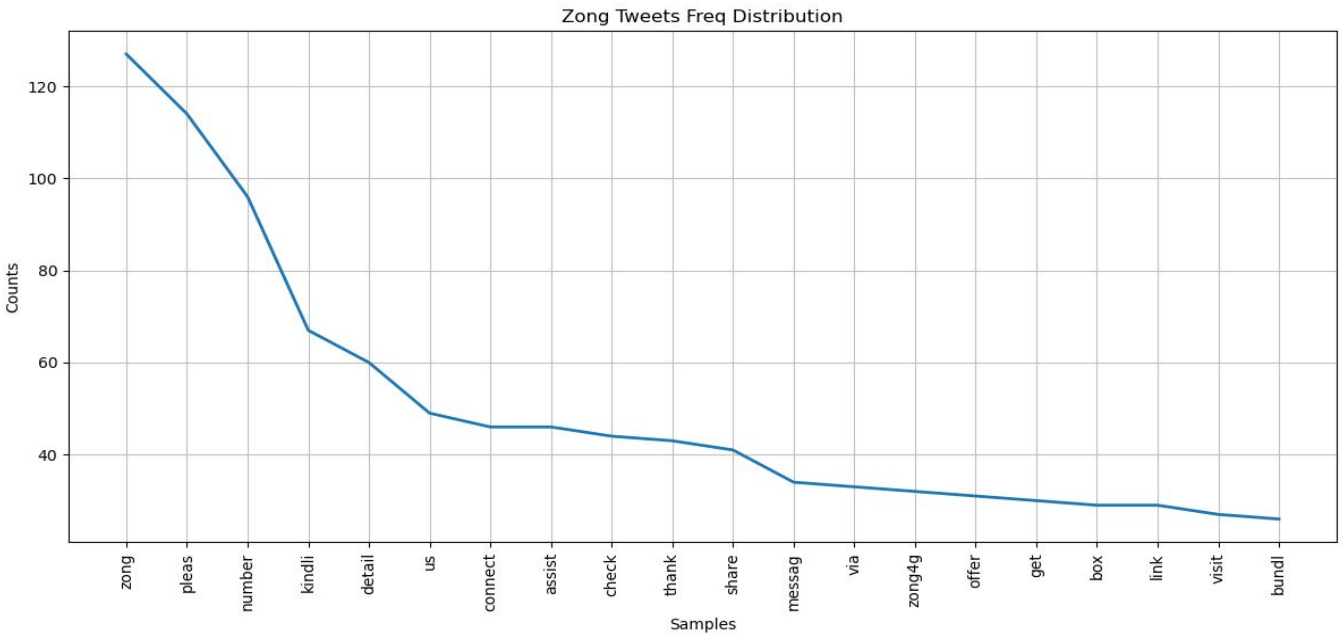
CODE ARCHITECTURE

- Firstly, we import the required libraries and packages including pandas and NLTK then we load the tweets.csv file as a pandas dataframe object into the variable df with the read_csv method.
- Then I save all of the tweets as a list in the variable tweets_list.
- Using the word_tokenize function of the NLTK library we tokenize our tweets into a list of tokens and save it in the variable tweets_words_list.
- Converted all the text in 'tweets_words_list' to lowercase.
- Created lists of stopwords and punctuations.
- Hardcoded a list of slangs.
- Created a regular expression to match the emojis in the text and then used the 're' library's findall method to find all the emojis in the text and save them to a list named emoticons.

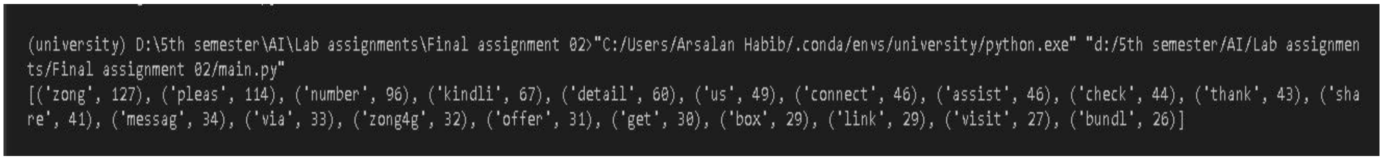
- Using list comprehension, filtered out all the stopwords, punctuations,slangs, emoticons, numbers and URLs and saved the new filtered text to a list called filtered_words_list.
- Initialized a porter stemmer object and used the stem method to stem all the words in the filtered_words_list and saved them to a new list called stemmed_words_list.
- Using NLTK's freqDist created a frequency distribution for all the words in the stemmed_words_list excluding punctuations and usernames.
- Printed and Plotted the 20 most commonly occurring words.

RESULTS

A plot showing the 20 most commonly occurring words with their counts.



The 20 most commonly occurring words after filtering and stemming.



```
(university) D:\5th semester\AI\Lab assignments\Final assignment 02>"C:/Users/Arsalan Habib/.conda/envs/university/python.exe" "d:/5th semester/AI/Lab assignmen
ts/Final assignment 02/main.py"
[('zong', 127), ('pleas', 114), ('number', 96), ('kindli', 67), ('detail', 60), ('us', 49), ('connect', 46), ('assist', 46), ('check', 44), ('thank', 43), ('sha
re', 41), ('messag', 34), ('via', 33), ('zong4g', 32), ('offer', 31), ('get', 30), ('box', 29), ('link', 29), ('visit', 27), ('bundl', 26)]
```

CHALLENGES AND DESIGN DECISIONS

No major challenges or difficulties while working on this assignment although, choosing slang words and hardcoding them into a list can be a laborious task especially when working with a large dataset.

REFERENCES

- https://stackoverflow.com/questions/53611711/compare-two-adjacent-elements-in-same-list (used this when creating the ranking in the warm-up)
- https://www.geeksforgeeks.org/python-stemming-words-with-nltk/
- https://stackoverflow.com/questions/4634787/freqdist-with-nltk
- https://stackoverflow.com/questions/5577501/how-to-tell-if-string-starts-with-a-number-with-python