Assignment 1: Text Pre-processing and Analysis in Twitter

Assignment Description

The main objective of this assignment is to understand the importance of text pre-processing in NLP tasks. Students will work with Twitter data and perform a series of pre-processing steps including lowercasing, removal of stop words, punctuation, URLs, lemmatization, stemming etc. They will also analyze the frequency of tokens after pre-processing and characterize the tokens based on their commonalities and differences between **positive** and **negative** sentiment tweets.

Dataset

The students will work with the *tweet_eval* [1] sentiment dataset that contains tweets with **positive** and **negative** sentiments.

* *Keep only the positive and negative tweets, filtering out the neutral class.*

Tools

Python, Jupyter Notebook, and the *twokenize* library will be used to tokenize the tweets.

Instructions

Apply a series of text pre-processing and normalization steps in order to gradually decrease the dimensionality of the dataset vocabulary. At each step characterize the percentage of the vocabulary size that was redacted. Note that you can keep the value of the vocabulary size to later visualize it in a figure.

In each step, students must justify their approach, considering both what we learned in class and any step suggested by them. Considering the text normalization steps of Lemmatization or Stemming, one must be selected and justified.

Analysis

1. Frequency of Tokens: Analyze the frequency of tokens in the pre-processed tweets.
2. Characterization of Positive and Negative Tokens: Characterize the tokens based on their commonalities and differences between positive and negative sentiment tweets.

Deliverables

1. Jupyter Notebook: Students should write the code in the Jupyter Notebook provided.
2. Results and Findings: The results of the pre-processing steps and analysis should be included in the Jupyter Notebook.
3. A .csv file with the fields of:

- **Step**: The step of the pre-processing / normalization phase.
- **#Tokens**: The total number of tokens at the specific step.
- **#|V|**: The cardinality of your vocabulary (unique tokens).
- **#Tokens_Negative**: The total number of tokens in the negative tweets at the specific step.
- **#|V|_Negative**: The cardinality of your vocabulary (unique tokens) of the negative tweets.
- **#Tokens_Positive**: The total number of tokens in the positive tweets at the specific step.
- **#|V|_Positive**: The cardinality of your vocabulary (unique tokens) of the positive tweets.

Grading

The grading will be based on the completion of each pre-processing step, the justification of the approach, the results of the analysis, and the quality of the code.

Deadline

The deadline for this assignment will be communicated by the instructor.

[1] Barbieri, Francesco, et al. 2020 "*TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.*" Findings of the Association for Computational Linguistics: EMNLP 2020.