UNIVERSITY OF SOUTHERN MAINE

Natural Language Processing, Spring 2023, Assignment 1

Instructor: Behrooz Mansouri (behrooz.mansouri@maine.edu)

Due: February 6, 2023

Notes for submission:

1. Submit your file(s) with the correct naming as: NLP_Assignment1_StudentName

2. Two files should be uploaded, one .zip file having all the codes (directory is zipped, and it is named codes) and one .pdf file. There would be a penalty for uploading wrong formatted files. Any other formatting will be ignored

3. Codes should be well-structured with comments to run

4. Codes should be available on your GitHub Repo. Failure to have codes publicly available, results in a 20% reduction in your grade. Make sure to include the GitHub link in your PDF file.

5. Any assumptions made by students should be explicitly mentioned in the submitted document

6. Answers to the questions should be easy to detect. For your codes, name the .ipynb files according to the question numbers (e.g., Question1.ipynb). In your document, use the question number and just write your answer. You should not have the question itself in your document. (e.g., Question 2. Generate)

Consider the Posts.xml file from the Coffee stack exchange. In the next two questions, we only focus on the questions' titles.

Question 1 (30%)**:** Creat a Zip's law plot (word vs. probability) on the words in the questions' titles. Does this plot follow your expectations of Zip's law? Discuss this in a paragraph in your solution. Include your plot and refer to it in your discussion.

Question 2 (30%)**:** Generating word clouds. Provide two word clouds on the top-20 frequent tokens, once without removing the stop words and another after removing them. You should use the NLTK library for tokenization and stop word removal. In your solution, put the two figures next to each other and discuss your observations.

Question 3 (20%)**:** In an NLP project, we are interested in designing a model to generate text from all the presidents. To do this, we need to train our model with the speech given by the presidents. You are not going to provide this data, but what would you suggest as the solution? Provide the steps and commands that you would recommend for gathering this data.

Question 4 (20%)**:** How WordPiece is different from BPE? Use the following example to show how WordPiece tokenization work:

low low low low low lowest lowest newer newer newer newer newer newer wider wider wider new new