NLP Homework 1 - Wednesday

Due Date: Tuesday, 9/26, 11:59 pm.

Corpus Statistics and Python Programming

For this assignment, please read Chapter 1 and 2 of NLTK book carefully.

Fake news detection has become an increasingly important NLP tasks in social media. In this class, you will analyze a dataset that contains fake news and real news articles applying the techniques we learn from the lectures. In this assignment, you will perform the following tasks.

1. **Data**

The dataset used in our assignments is ISOT Fake News Dataset. It contains two types of articles fake and real news. The real news articles were collected from Reuters.com and the fake ones were from various sources that were flagged by Politifact (a fact-checking organization in the USA) and Wikipedia. There are about 12, 600 articles in either type and the information of these articles is saved in two CSV files, "True.csv" and "Fake.csv". Specifically, each article contains the following information: article title, text, type and the date the article was published on. You can read more about this dataset and download it from Kaggle web site: https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets

2. **Data Pre-processing**

You will write a Python code that extracts data from the field "text" for both CSV files. You will apply pre-processing tasks to the texts. Please process them separately.

Note: you will decide how to pre-process the words, i.e. decide on tokenization and whether to use all lower case, use or modify the stop word list, or lemmatization. Briefly state why you chose the processing options that you did in your Jupyter Notebook file (in a text block). Please keep in mind that you will practice using word frequencies to analyze the text in this assignment.

3. **Data Analysis**

To get a rough understanding of what these texts were about, you will perform the following tasks on the pre-processed data:

- list the top 50 stop words by frequency in fake news articles and those in real news articles
- list the top 50 content words by frequency in fake news articles and those in real news articles
- list the top 50 bigrams by frequencies in fake news articles and those in real news articles

- list the top 50 bigrams by their Mutual Information scores (using min frequency 5) in fake news articles and those in real news articles
- list top 50 adjective words in fake news articles and those in real news articles

For each article, you will also obtain total number of words, total number of content words, total number of words that are all capitalized (excluding "I"), total number of exclamation marks, and total number of punctuation marks. You will save this information in CSV files by creating new columns and saving the information in these columns.

4. Interpretation of the Results

Please compare the analysis results from the true news articles and from the fake news articles and explain what you have discovered (e.g., do they tend to have similar ratio of capitalized words?). Please feel free to conduct additional analysis that helps you understand the above results.

How to Submit Homework: <u>Due Date: Sept. 26, 11:59 pm.</u>

Go to the Blackboard system and the Assignment for Homework 1 and submit your notebook file. The file needs ***to be run already*** and ***results are displayed***. Please provide ***enough comments*** in the code and ***name the variables mindfully*** to make it more intuitive for people to understand the logic behind your code.

Additionally, please provide text blocks that:

1. ***describe the data pre-processing tasks*** and why you chose to do such pre-processing
2. Interpret the results of the analysis

Please also submit the ***two revised CSV files*** that contain the analysis results as described above.

NOTE: Please name each of your submission files such that it contains your first and last name.