Project Description: Scaping, Text Preprocessing, Classification & Sentiment Analysis

Project Summary

As a part of the curriculum of the Master 1 (M1) course entitled “Advanced Machine Learning and Text Mining”, the students will complete two project work. This is the first project called “Text pre-processing, Classification and Sentiment Analysis” which is focused on pre-processing texts, having a solid example on text classification and performing analysis to extract positive and negative sentiments.

The datasets required for this project will be delivered to the students. Besides, a Jupyter notebook will be available.

The following sections provide necessary information about the datasets and the Jupyter notebook. Also, the project tasks are explained.

For any further detail, please contact the instructor: **Khodor Hammoud**

1. About Dataset

The datasets provided for this project are about speeches of US presidents, collected from millercenter.org. 11 datasets are containing the speech of different presidents including Barack Obama, Bill Clinton, G.W. Bush. The students are to run the code from the provided notebook to complete the data collection.

2. About the provided Notebook

A notebook named “scraping-presidential-speeches” is provided which contains the code required to scrape the site for the speeches. You must put this notebook and the dataset folder “presidents-speeches” in the same directory.

Once you’ve run the code and scraping is completed, The final files structure will be as so:

presidents-speeches > [president_x_name] > speeches > [president_x_speeches]

where president_x_speeches is a list of the speeches corresponding with president_x, labeled by date and speech title.

All groups will use these scraped data to perform their required tasks.

Note:

You are highly encouraged to inspect the code and understand how it works. It is important because in the future project you would scrap data from online sources.

3. Tasks Description

This project is composed of two parts: Part A and Part B. Each group (consisting of two students) shall complete the tasks of both Parts A and Part B. See the description of the tasks in the following sections:

3.1 Part-A

You will apply text categorization techniques on the collected datasets. The input is the speeches of the presidents, the label is the name of the corresponding president.

The final goal is, given a presidential speech, to be able to identify who is the president that delivered the speech. To that end, carry out the following tasks:

Your Tasks

- Apply all the necessary pre-processing techniques that you see are necessary for the task.
- Split your data into 80/20 % groups of training/testing
- You are free to choose which model you want (SVN, KNN, Naïve Bayes…)
- Perform parameter tweeking to maximize the accuracy of your model
- Save your trained model, and be ready to load it in your presentation
- Have a ready example to run during the presentation

When presenting your work, all the code should be presented in a notebook (Jupyter, google colab…).

3.2 Part-B

You will apply sentiment analysis on the datasets provided using a pre-trained sentiment analysis model from NLTK called VADER. Here's a good article with examples on the usage of VADER:

https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

Surely you are free to do your research on VADER well.

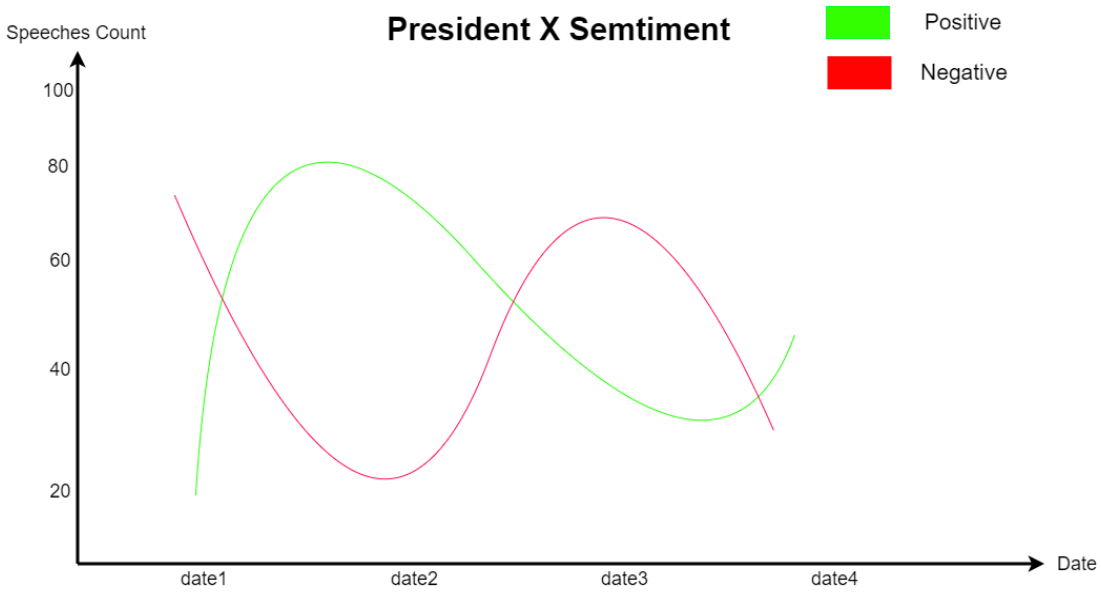
Besides, you are required to analyze the evolution of US presidential periods' sentiments over the years -- A presidential period of president X is the duration during which president X was in charge --. To do that, you must carry out the following tasks:

Your Tasks

- Order the presidents in chronological order, from oldest to newest. For reference: https://www.loc.gov/rr/print/list/057_chron.html
- Order each of their speeches in chronological order as well, from oldest to newest.
- Assuming you've already pre-processed the speeches in Part A, run the VADER sentiment analyzer on the processed speeches and extract the sentiment of every speech.
- Count the number of negative speeches vs the number of positive speeches for every president. -- The sentiment of a speech is the overall sentiment of the sentences; For

example, if 60% of the sentences are positive and 40% are negative, then the speech has a positive sentiment.

- Use a graph to visualize the sentiment of every president over his presidential period. See the example below.
- Visualize the overall sentiments of all presidential speeches in one graph, which can help us see the overall centemental change of US president speeches over the years.



While presenting your work, you should be present your codes in a notebook (Jupyter, google colab…) with the graphs pre-drawn.