Dear All,

This assignment builds on the text similarity and classification material. (Please refer to the notebook on Naive Bayes for the latter.)

You will be working with senator speeches located in the '105-extracted-date' folder within the 'Inputs' directory. Please complete the following tasks:

i) Use cosine similarity to determine which senator's speech is closest to Senator Biden's. Describe your findings. Make sure to describe your text preprocessing and justify your choices. Validate your findings using 'sen105kh_fix.csv' and/or Wikipedia to see if the most similar speeches belong to senators from the same state and/or party.

ii) How do your results change if you apply stemming or lemmatization? In your opinion, which is better to apply: stemming or lemmatization? Why?

iii) Create at least two visualizations to support your analysis.

iv) Use 'sen105kh_fix.csv' as the target variable for your predictions. Can you predict the party of the senator from their speech? Should you use the same text preprocessing as above? Justify your choices.

v) Write a Medium article summarizing your findings. This article should be written with a general audience in mind. The best articles may be invited for submission to CEU Economic Threads.

vi) Bonus 10 points: Read one of the articles that I have linked in this Medium post. Contrast your findings in i) with theirs and discuss any discrepancies. What are possible solutions?

Feel free to collaborate and use resources such as StackOverflow and ChatGPT. However, you must submit individual solutions titled 'name-surname.ipynb'. Include the link to your Medium article in the notebook.

Good luck!