Foundations of AI
Instructor: Rasika Bhalerao
Assignment 7
Due November 18

This assignment is to analyze the results of LDA on a set of documents, and then use the LDA vectors for classification and clustering.

Learning goals:

- Using Pandas to handle data
- Using Scikit-learn for LDA, classification, and clustering
- Analyzing the results of LDA
- Using LDA vectors as feature vectors for classification
- Using LDA vectors to cluster documents

What to do:

1. Download this dataset and read it into a Pandas dataframe.
    a. **It is important to note that this assignment is for learning only; in reality, distinguishing real and fake news is more nuanced.**
    b. I recommend converting the documents to lowercase and filtering out stopwords.
2. Start with k = 10 topics. Fit an LDA object to the set of all news text. Then, examine the top *n* words from each topic (choose a reasonable *n* such as 10 or 20). How well do the topics represent real-world topics? (One sentence)
3. Randomly select 5 real news examples and 5 fake news examples, and examine the topic distributions for each document. Which topics are prevalent in the real news documents? (One sentence) Which topics are prevalent in the fake news documents? (One sentence)
4. Use the LDA vectors for the documents as features in a Logistic Regression classifier to predict whether each document is real news or fake news. According to the resulting coefficients from the regression, which topics are most useful in determining whether something is real news or fake news? (One sentence)
5. Pick real news or fake news, whichever is more interesting to you. Then, use the LDA vectors for those news documents to cluster them. You can use KMeans clustering with a reasonable value for K (if you don't have strong feelings for a particular K, I recommend 10). Then, select 5 news documents from each resulting cluster. Do the clusters correspond to anything? (One sentence)
    a. If you don't like KMeans, you can use a different clustering method.

What to turn in:

Please submit these via Gradescope:

- For this assignment, the questions specific to this assignment are integrated into the coding assignment. You may turn it in as one PDF printout of your Colab notebook, or as

a regular Python file with the answers to the questions as comments. (Or, you may turn in your code and your answers to the questions in two separate files.)
    - The “(One sentence)” suggestions are suggestions. You will be graded based on evidence that you did the tasks described and understood the intentions behind the questions.
- Your answers to the usual questions:
    - How long did this assignment take you? (1 sentence)
    - Whom did you work with, and how? (1 sentence each)
        - Discussing the assignment with others is encouraged, as long as you don’t share the code.
    - Which resources did you use? (1 sentence each)
        - For each, please list the URL and a brief description of how it was useful.
    - A few sentences about:
        - What was the most difficult part of the assignment?
        - What was the most rewarding part of the assignment?
        - What did you learn doing the assignment?
        - Constructive and actionable suggestions for improving assignments, office hours, and class time are always welcome.