DS2500: Lab for week 6

Prof. Rachlin

Fall 2021

Oct 13-15, 2021

Submit on Gradescope everything you've completed by 11:59pm eastern on Friday of this week.

Labs are graded on a 0–4 scale. Full credit is given for demonstration of participation and effort on the lab assignment.

Lab Problem:

In class this week we learned about cosine similarity as a way of measuring similarity in text documents. Today, we're focused on the similarity of fictional presidents instead of real ones, but working through the lab will still be helpful as you get deeper into HW3.

The text files for this lab are speeches given by President Bartlet (from the West Wing) and President Shepard (from the American President), and we're going to see how similar they are, using (1) visualization, and (2) cosine similarity. Both characters are written by Aaron Sorkin, but the speeches are in totally different contexts, so it's kind of hard to predict what we'll find.

You may use the code we did together in class to compute the cosine similarity of two texts based on the dot product of two vectors.

Here's the approach today:

- Read in each text file as a string.
- Visualize both Presidents' speeches using the wordcloud library (also discussed in Tuesday's lecture).
- Find the most common words used by each President. You might find it convenient to use the `Counter` module to do this. Try creating vectors using the unique words among the most frequent k words of each speech, where k=10-20 words (your choice).

- Use the `cosine_similarity` function provided in class to measure the similarity of the two speeches. Is it what you expected? Would you have guessed that these characters were written by the same person?
- As an optional enhancement, try cleaning up the text to remove small words and punctuation. This might produce a more accurate result.