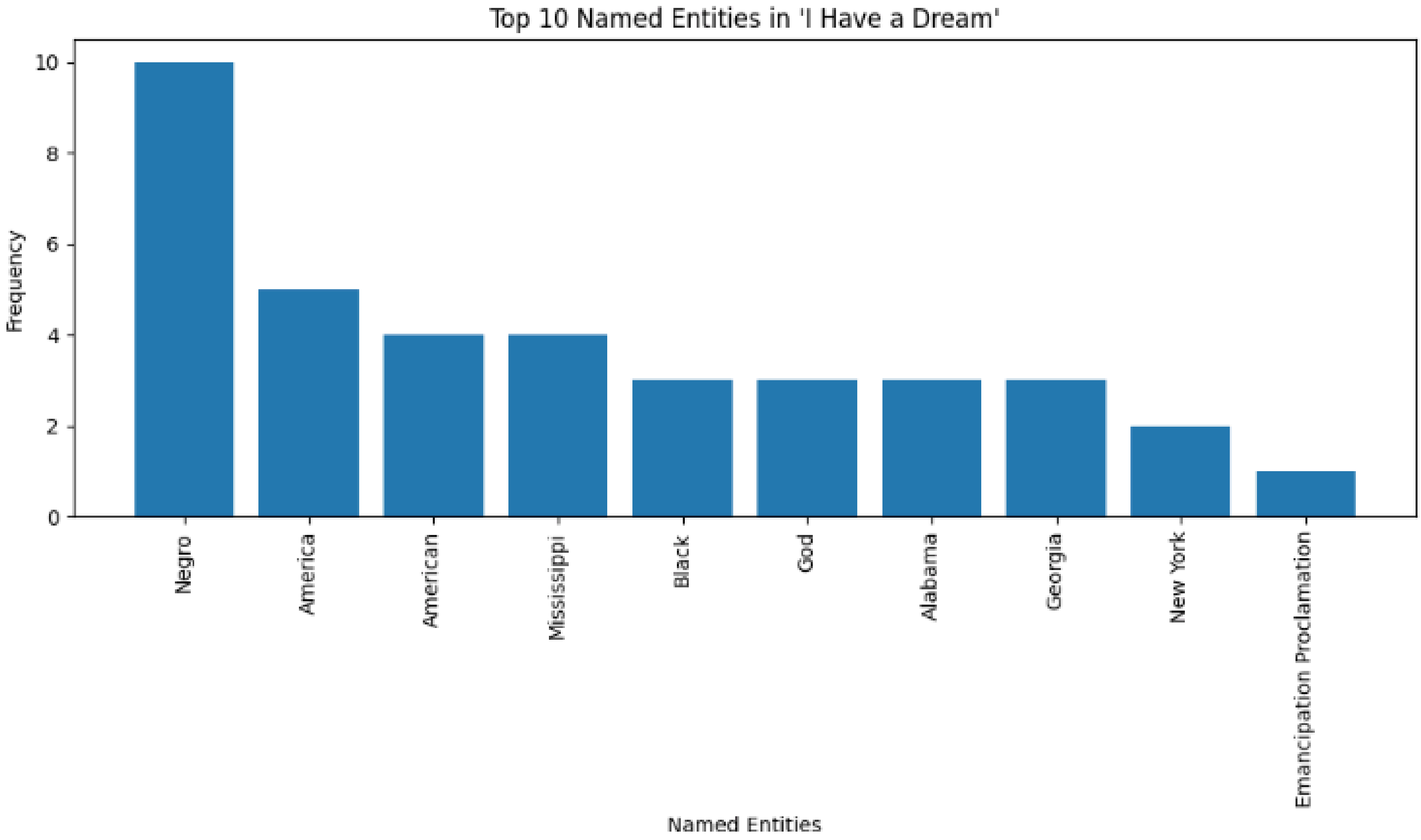
Homework #2 – NLP **CSC 4260** **Spring 2025**

Name: Tania Perdomo Flores

Word Tagging *{50 points}*

Textbook Source: Chapter 2.5 Word Tagging

1. Using the provided I-have-a-dream.txt file, do the following: {*20 points*}
   - Parse the sentences of the speech.
   - Run some parts of speech tagging to determine the top-named entities in the speech.
   -
2. Plot the "Named Entities" on the X-axis and the "Frequency" on the Y-axis (like Figure 2.5 in the book). *{10 points}*



3. What do you notice about the top 10 named entities? Does anything surprise you? **Explain your rationale**. {*20 points*}

I noticed that the top 10 named entities reflect key themes of race, nationality, geography, and religion. The most frequent entity, "Negro," reflects the speech's focus on racial injustice and civil rights, as Dr. Martin Luther King Jr. directly addressed the oppression faced by Black Americans. The presence of "America" and "American" underscores the speech’s broader appeal to national identity and unity, reinforcing the idea that the civil rights struggle is deeply tied to the nation’s foundational ideals. Geographic locations such as "Mississippi," "Alabama," "Georgia," and "New York" point to the speech’s emphasis on the widespread nature of racial inequality, as these places were central to the civil rights movement, either as sites of oppression or as beacons of change.

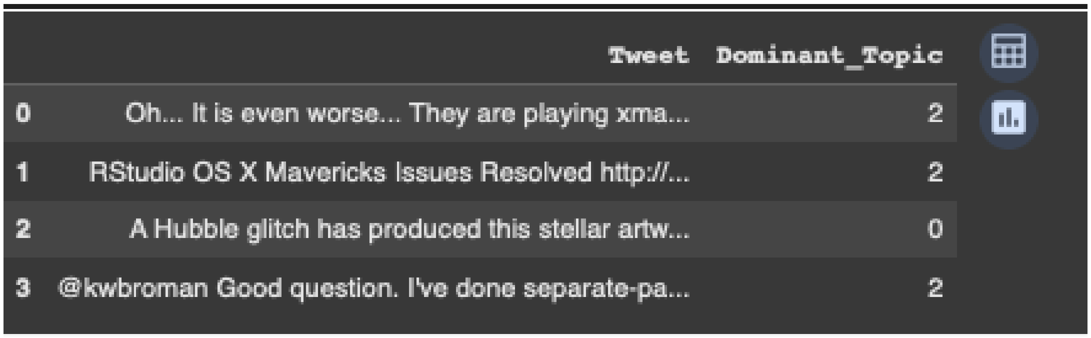
What stands out is the relatively lower frequency of terms like "Emancipation Proclamation" and "God," despite their significance in the speech. The *Emancipation Proclamation* was a key historical reference in Dr. King’s argument that freedom had been promised but not fully realized, yet it appears less frequently than expected. Similarly, "God" is present but not dominant, even though King, a pastor, rooted his call for justice in faith-based principles. This suggests that while these elements were crucial to his argument, the core focus remained on racial identity, national belonging, and the specific regions affected by racial injustice. The data aligns with the speech’s purpose: to shine a spotlight on the urgent need for racial equality in America.

Tweets *{50 points}*
Textbook Source: Chapter 2.6 LDA in Action

4. Using the provided training and testing tweets files, perform LDA. {*30 points*}

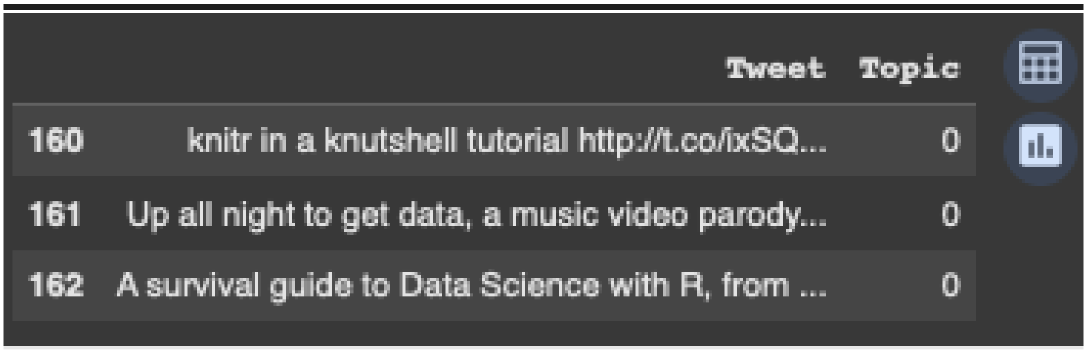
5. Show some of the predictions made on the test data. Does it look accurate to you? **Explain your rationale**. {*20 points*}

With training data (for comparison):



	Tweet	Dominant_Topic
0	Oh... It is even worse... They are playing xma...	2
1	RStudio OS X Mavericks Issues Resolved http://...	2
2	A Hubble glitch has produced this stellar artw...	0
3	@kwbroman Good question. I've done separate-pa...	2

With testing data:



	Tweet	Topic
160	knitr in a knutshell tutorial http://t.co/ixSQ...	0
161	Up all night to get data, a music video parody...	0
162	A survival guide to Data Science with R, from ...	0

The predictions made on the test data show some consistency between the training and testing data, but their accuracy depends on how well the topics were defined and whether the model effectively generalizes beyond the training set. In the training data, Topic 2 appears to capture tweets related to technical issues and software, while Topic 0 might be linked to science or data-related discussions, given the "Hubble glitch" tweet. In the testing data, all tweets are assigned to Topic 0, which suggests that the model sees them as closely related to the science or data theme. If Topic 0 was primarily about data science, programming, or research, then these assignments could be reasonable. However, without deeper insight into the topic-word distributions, it's difficult to determine if this classification is entirely accurate.

A potential concern is whether the model is overfitting to the training data and struggling to distinguish nuanced differences in new tweets. If Topic 2 was expected to be present in the testing set but wasn’t assigned to any tweets, it could indicate that the model isn’t fully capturing topic diversity. Additionally, the short and informal nature of tweets makes topic modeling challenging, as context can be ambiguous. If the model tends to lump anything technical or data-related into Topic 0, it may be missing finer distinctions that a human reader would catch. A more thorough evaluation, such as reviewing the most representative words for each topic and checking coherence scores, would help determine if the results are truly meaningful.