NLP Homework 1

Joe Malatesta

Dataset

The dataset I chose to analyze was between fiction and nonfiction books from Project Gutenberg. I used three full books for each category but in retrospect I should have used a few more. Many of the topics of the books ended up showing up more than the categories that I was hoping for. For example, one of the books I chose for nonfiction was the autobiography of Napoleon Bonaparte. This book was 5-10x larger than the other books I chose and drowned out a lot of the extra information provided by the other texts. A lot of the most weighted words in the nonfiction corpus, therefore, have more to do with Napoleon than the whole of the rest of the corpus. This, I believe, could be simply avoided by adding more books of the same genre. This would allow the main themes of these kinds of books to come out instead of the theme of a single book outweighing everything else.

The books I chose for fantasy / fiction were Alice in Wonderland, The Wizard of Oz, and Peter Pan. These provided a strong base for a fictional children's story that followed common tropes and themes of magic and unreal happenings. The books I chose for nonfiction were biographies of Napoleon Bonaparte, Ben Franklin, and Hellen Keller. Like I said prior, I believe my selections were strong and would paint common themes but I did not pick enough text to correctly and strongly determine categories.

I knew what the results would entail but I was interested in how much the topics would reflect the distinctive narrative styles and storytelling elements specific to each genre. Specifically, I was curious to see if the topic models could effectively discern the imaginative, and fantasy elements of fiction from the factual, historical content of non-fiction.

Methodology

1. I began by choosing the text, separating them to their own text files, and appending them back together programmatically. I should have only had two text files to make it simpler to append new text and make the corpus larger
2. I normalized the text just as I did in the previous homework, lowercasing, removing capitalization, and removing stop words. I previously used stemming but it was not as useful in this scenario and I found that having the option of lemmatization was far more useful for extracting meaning from the corpus.
3. After preprocessing the large appended texts for each category, I calculated the probabilities of each word in the vocabulary and added them all up with laplace

smoothing to make sure that my calculations would not be zeroed out in later steps.

4. Afterwards, I passed the probabilities from the previous function to a log likelihood ratio to find out the joint LLR for the total of the subjects.
5. I sorted the sides of the LLR to show the top words of each category, which the top 10 were mainly referring to the specific texts that I chose and not the overarching themes. After the first 10, though, you begin to see themes as I will talk about later
6. Finally, for topic modeling, I used the Gensim package to create two LDA topics that I was quite pleased with. While they had some of the same issues that could be fixed with more text, the topics Gensim pulled out seemed fairly accurate and well grouped. From Gensim I used corpora and LdaModel packages.

Results and analysis

I added the top results with and without lemmatization for both LLR and LDA from Gensim. There is a link here to view the results
https://silk-impatiens-c4e.notion.site/NLP-Homework-1-Diagrams-ef18f475bcba4043b38feea483de0c87?pvs=4

Discussion

- In this NLP project, I analyzed fiction and non-fiction books from Project Gutenberg. The results highlighted the impact of dataset selection, with Napoleon's extensive autobiography skewing the non-fiction analysis. This underscored the importance of balanced datasets in NLP, as the fiction texts like 'Alice in Wonderland' clearly reflected their fantastical themes. Overall, the texts I chose were meant to display the overarching ideas of fiction vs nonfiction (fake conflict/topics vs real conflict/topics) but I chose my selections poorly and found that it was difficult to tell the groups well.
- Overall, I could have easily extended the corpus' with more alike text but I thought that showing my thoughts about what I expect and what I got would help tell a better story of what I learned. This project helped me understand, above all else, was how much a model can be skewed by an overpowering text. In my case, Napoleon overwhelmed the nonfiction category. This ties back to how models can be really good at one thing and poor at others. If my model was meant to guess whether or not some text came from Napoleon's biography it would do a great job but otherwise it was a bit underwhelming.