Software Design
Text Analysis
Emily Kohler

Part 1: Project Overview

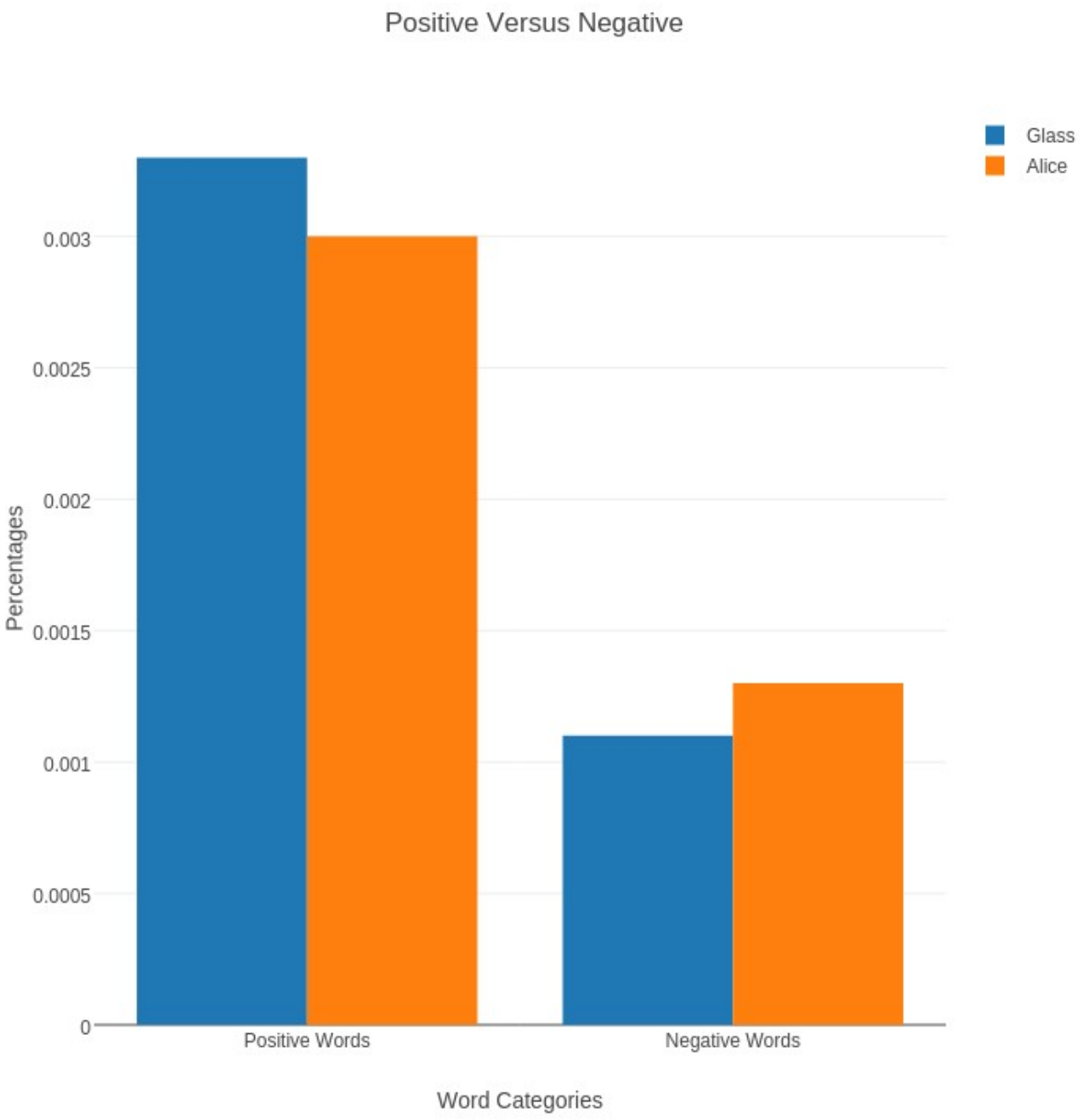
For my Text Analysis mini project, I used the Gutenberg project website as my data source and I specifically downloaded the texts of *Alice's Adventures in Wonderland* by Lewis Carroll (Charles Lutwidge Dodgson) and *Through the Looking-Glass, and What Alice Found There*, also by Lewis Carroll. The techniques I used for this project include: characterization by word frequencies, pickling files and text similarity, specifically cosine similarity. I hoped to learn how to use these techniques and how to use the data gathered from these techniques to compare the two books in terms of the author's mental and emotional status (similar to a physiological analysis).

Part 2: Implementations

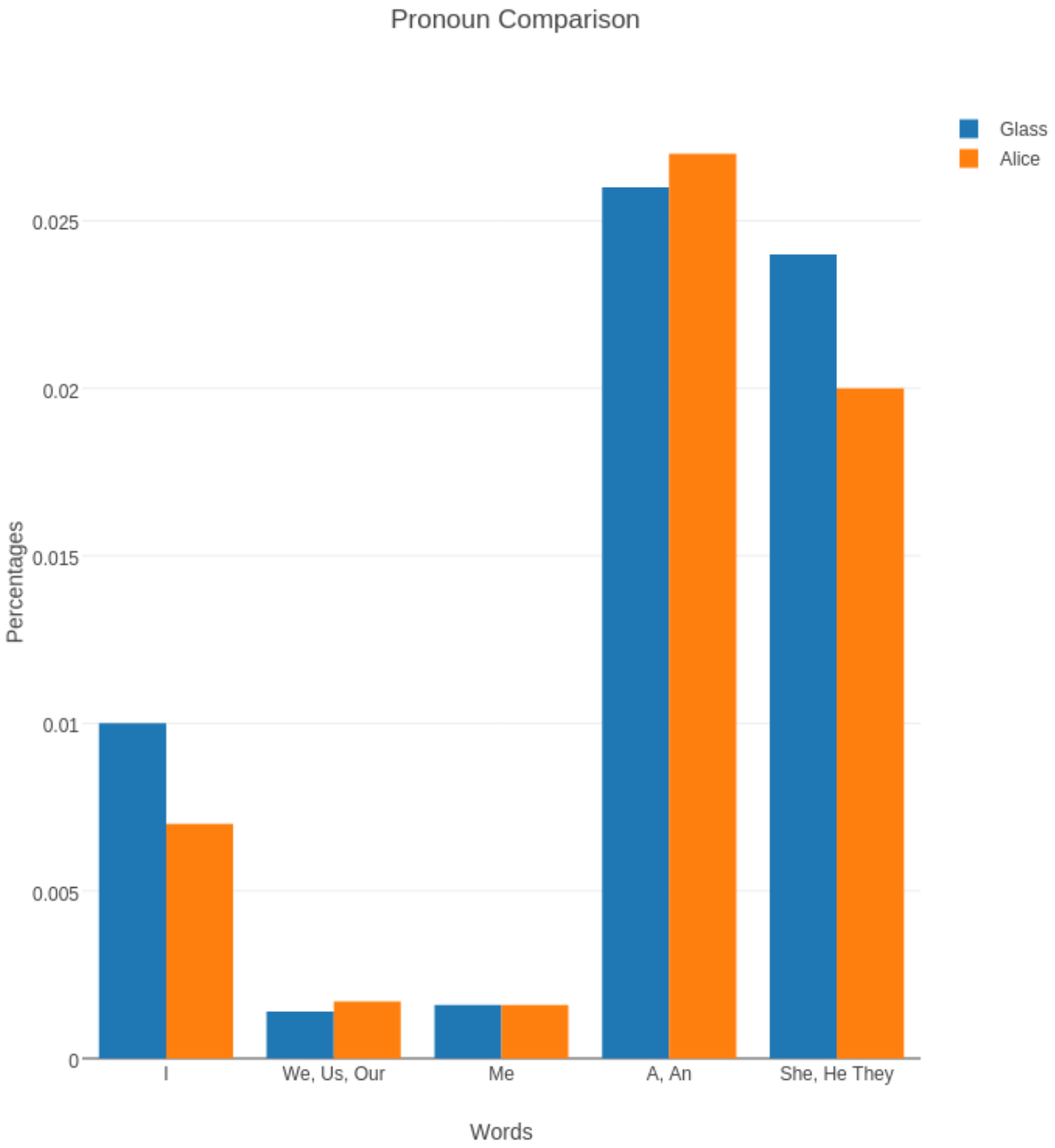
In fairly general terms, I converted a plain text file to a pickle file, so that it would be in a more agreeable format with the computer. Then I used this pickle file to extract data and certain properties of the text. For example I made a series of functions that computed the number of times certain words appeared, a technique called characterization by word frequencies, and then a program to compute how many total words appeared in the text. I also created a function to compute the cosine similarity between two texts. This function worked by converted the text files into vectors and then doing a mathematical analysis to see how similar the two vectors are (getting the dot product multiplied by the cosine of theta).

One design choice where I had to choose between multiple alternatives was when I had to choose what type of data structure to store the word frequency information in. I ended up choosing a list of tuples, over a list of lists, because tuples are immutable and the number of word frequencies wouldn't be changing once a text has been analyzed. Tuples were also able to show the word and the frequency of that word, such as: ('we', 24).

Part 3: Results



The two stories Alice's Adventures in Wonderland and Through the Looking-Glass are very personal to Lewis Carroll, in that he was describing a syndrome he dealt with called Todd's Syndrome, lilliputian Hallucinations or Alice in Wonderland Syndrome. According to physiological analysis of text, the percentage of positive and negative words reflects the author's feelings (which also makes logical sense). I thought Carroll would feel more at peace in his life after having written Alice's Adventures in Wonderland, since it would be an outlet for his emotions. Instead the data shows Carroll started using fewer positive words and more negative words, pointing toward a negative emotional turn. This is interesting and probably reflects what was going on in his life at the time. He was also said to have migraines and epilepsy, which could have contributed to his worsening emotional state.



The percentage of the letter “I” between the two stories had the second most drastic increase from Alice in Wonderland to Through the Looking Glass. The use of the letter “I” is associated with more self-focus, people of lower status, or someone who is depressed. Since there are multiple different meanings, the use of the letter depends on the situation. For Lewis Carroll, it's possible he got more depressed between the first and second book (which would also make sense in terms of the positive and negative word analysis). The “We, Us, Our” category usually signifies that the person is lying if it is higher than the “I” category. In this case it makes sense for Lewis Carroll not to be lying because the book was indirectly about a syndrome he had. The “A, An” category has to do with the number of nouns used and the “She, He, They” category references the fact that the author is more about other people and personal relationships—considering this it seems like Lewis Carroll's personal relationship got better between his first and second book.

The cosine similarity function between the two books was .98, which means the books were extremely similar in terms of the vocabulary used. This is interesting, because you would expect two different books to be very different, even though they have the same author.

Part 4: Reflection

I think the pickling, word frequency analysis, and cosine similarity functions went well and gave accurate values. The use of the “Try Except” function also went well and it allowed me to learn another part of python that could help me in later assignments. I think I could improve the way I did the word frequency code in the future. As an example, I only did a random list of positive and negative words instead of finding every positive and negative word possible. I would also want to include cognitive words in the future, which would require more research.

The doc string testing also went fairly well because they gave me more confidence that my code was right. For the doc string testing I made my own short text file where I am able to count each word easily and therefore verify the results of my functions. This was very helpful and probably the only doc string testing I could do. It also helped me to find potential problems in my code that I fixed.

Going forward, I learned how to pickle text files, how to use pickled text files, how to calculate word frequencies, how to be creative in writing doc tests, the “try except” function and how to code cosine similarity. I also learned how to apply coding to yet another academic field.