Shruti Iyer

Project overview:

The data source used in this project are books from Project Gutenberg. There are two sub-projects within this project. In both the projects, I analyzed the text and found the most frequently used words. The code then assumes the frequency of words as vectors and find the cosine similarity of the two texts. The first subproject analyzes how the cosine similarity of two texts changes as the input words (essentially vectors) increase. In the second subproject, I tried to find how similar texts are over a century.

Implementation:

The major common block of the two codes are as follows. The first block opens all the .txt files and finds the top n most frequently used words. These word- frequency pairs are then used as vectors to compute the cosine similarity.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum\limits_{i=1}^{n} A_i \times B_i}{\sqrt{\sum\limits_{i=1}^{n} (A_i)^2} \times \sqrt{\sum\limits_{i=1}^{n} (B_i)^2}}$$

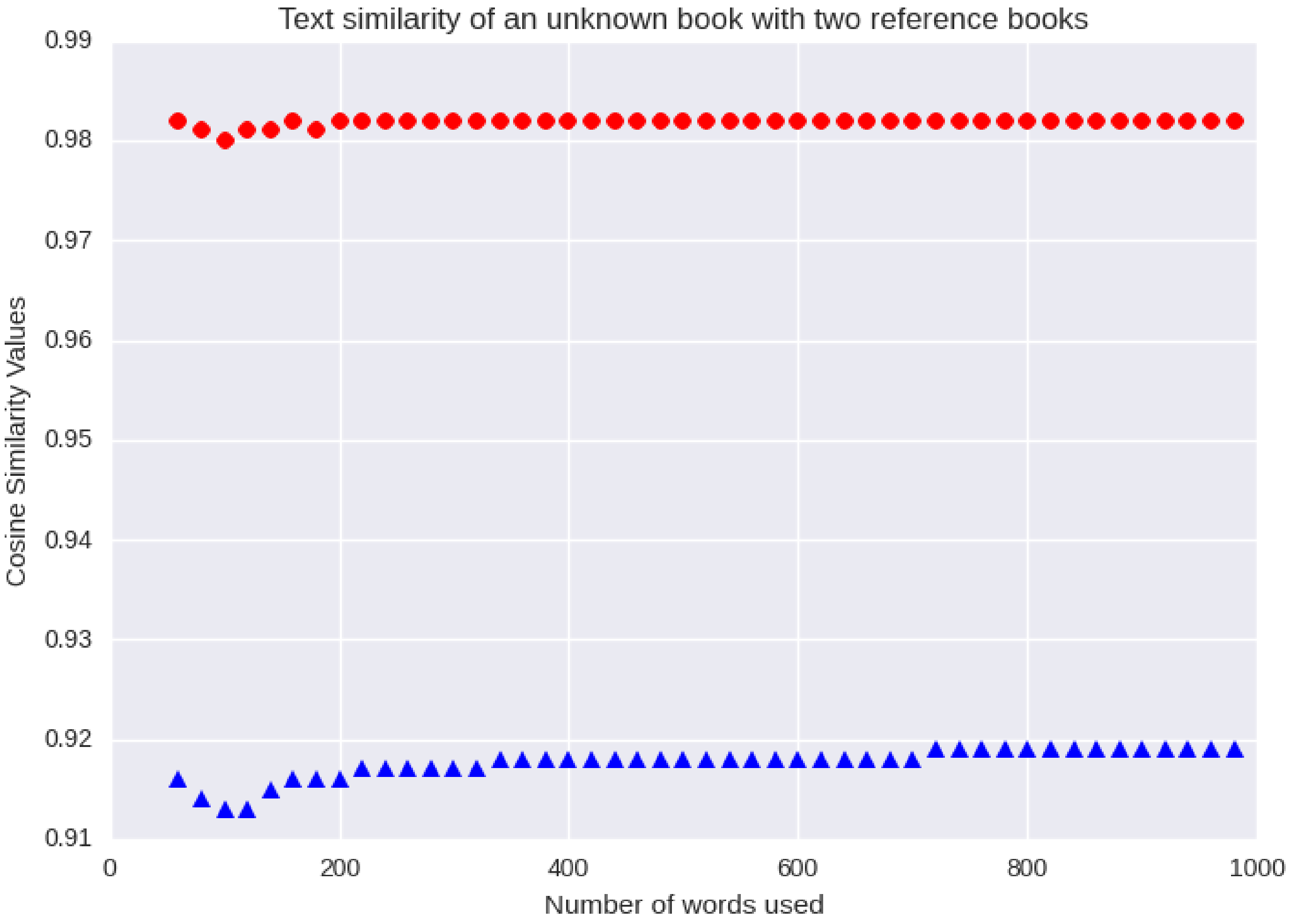
Here, Ai and Bi are the word vectors.

A major design decision in this project is assuming “Eating nutritious food is extremely important” and “Consuming nourishing food is absolutely necessary” mean two complete different things. The code does not check for synonyms or different word forms. I took the decision because computers cannot understand the information the way humans do.

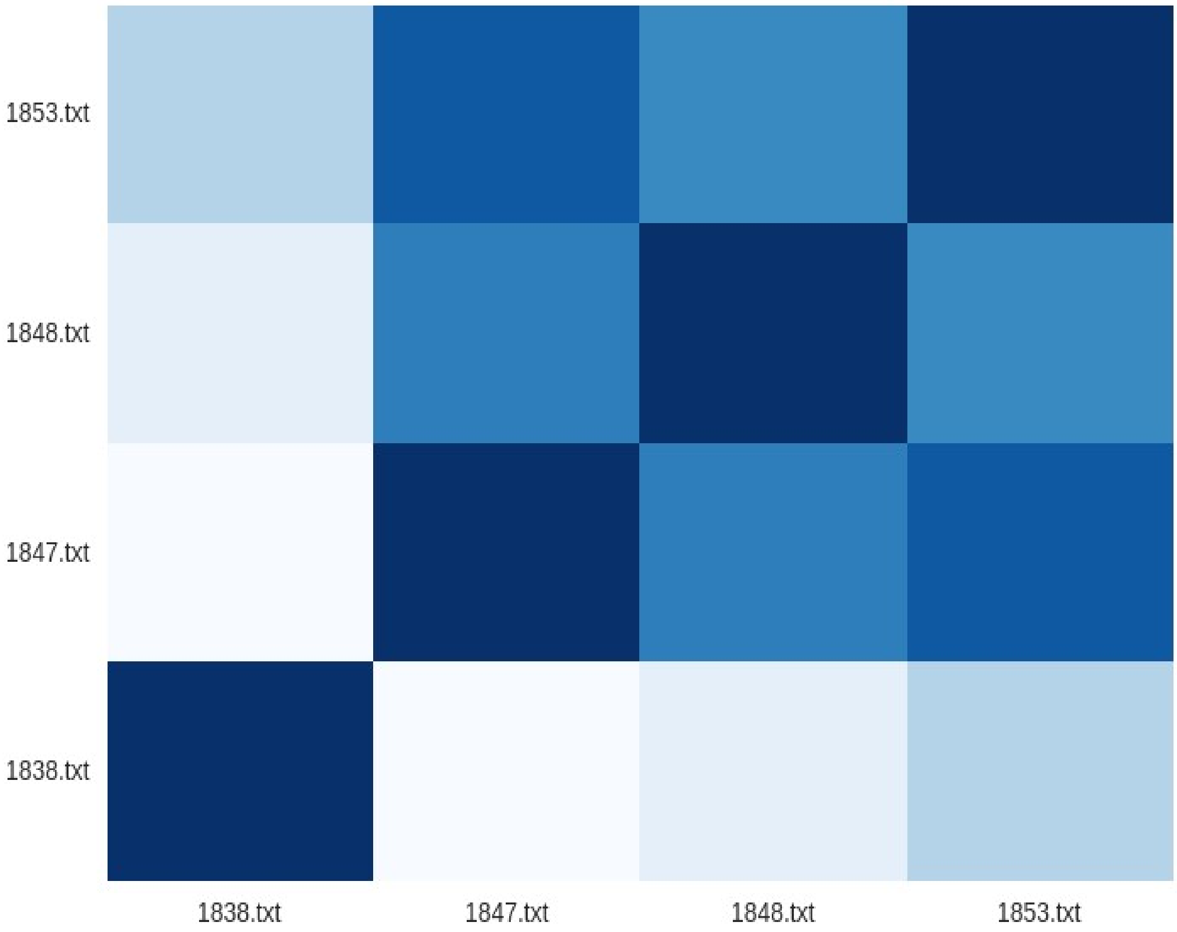
In the fist subproject, I tried to find the similarity of the text with two sample texts. The sample texts that I used were Jane Eyre by Charlotte Bronte and Oliver Twist written by Charles Dickens. The text with the unknown author was Villette by Charlotte Bronte. So, I was expecting the cosine similarity of the unknown text with Jane Eyre to be significantly higher than with Oliver Twist. In the second subproject, I tried to find how the text similarity changes over years. Hence, I created a heatmap of the texts and arranged them chronologically.

Results:

In the first program, I changed the number of vectors and found the output of the cosine similarity of the unknown text with the two reference texts. The value of the n goes from 60 to 1000. The red dots are the text similarity of the text with jane Eyre and the blue are the values from Oliver Twist. It can be seen from the graph that above a certain threshold (which is 60 in this case), the cosine similarity does not change by a lot. In all the values of n, the program correctly identified the correct author of the unknown text (as the red dots are significantly above the blue ones)



For the second subproject, the heatmap for the books arranged chronologically is:



As we read the heatmap, the more closer the two blocks' years are, the darker the cell is. But, we cannot make well-supported thesis because of insufficient data.

Reflection:

One major discovery during this project was the vast and large number of libraries that Python has. I realized that there is always a library to help you reduce the length of the code. A lot of my explorations revolved around numpy, arrays and graphs. Before this project, I wish I knew how to efficiently read documentation. One aspect of the project that did not go well was that I couldn't download more books from Project Gutenberg to test my second subproject (It gave me Error 403, my IP address was blocked). Overall, this was the first mini-project that had absolutely no scaffolding which taught me how to plan the code, write pseudo code and do incremental development.