Project Overview:

I used Wikipedia articles and encyclopedia Britannica articles to determine how similar go-to information sources from the past and present are. After isolating the texts, I took the cosine similarities of articles from each source of the same subject.

Implementation:

The first step after pickling the section of Encyclopedia Britannica titled "Encyclopaedia Britannica, 11th Edition, "Latin Language" to "Lefebvre, François" was to create a sort of table of contents and isolate each section. After looking at the texts entries, I determined to only take the first paragraph of each even if the actual entry spanned much longer because most people do not read that much anyway. Isolating the text entries proved harder than expected because many of the text entry titles were different than the titles listed in the table of contents.

After I had a list of text entries and titles of entries, I used the Wikipedia library to find articles of the same title as the Britannica articles and used the summaries of these articles. I used the summaries because I used a short excerpt from the Britannica entries as well. Unfortunately, some of these articles did not exist on Wikipedia so I had to figure out how to deal with exceptions that Wikipedia was throwing. After getting an article with the same title in both Wikipedia and Britannica I would make an alphabetically sorted list of word frequencies.

Finally, I took the lists of word frequencies from each source and found their cosine similarity. I repeated this process with each article in the Britannica text. So my final output was a list of cosine similarities whose indexes corresponded to the title of their respective article.

Results

Cosine similarity determines how similar two texts are based on their word frequencies as vectors. If the cosine similarity is 1 that means that that the the articles are basically the same (because the angle between them is 0 and cos(0) = 1). If the cosine similarity is 0 that means that the two texts are as different from each other as possible (because cos(90) = 0 and 90 is as far apart the angles can be).

With my list of cosines similarities I determined the maximum (most similar), minimum (least similar), average similarity and median similarity. The results are as follows:

Quality	Value	Entry
Maximum (most similar)	0.9779905930176569	Charles Le Beau
Minimum (least similar)	0	Latin Language

Average	0.8356969104400732	
Median	0.8793378442281854	LAUNCESTON (Tasmania)

These results make some sense because Charles Le Beau's entry, is about a person who existed before the Britannica edition was published so their history probably has not changed much. However, the value of Latin Language is probably a mistake, because even though the language has evolved it is difficult to believe that the entries were that different, considering that the average cosine similarity was 0.8356969104400732. The entry for Launceston makes sense because their must be new information about the city but because it existed a long time ago it has not changed that much.

Reflection

This project took much longer than expected because when testing I discovered a lot of exceptions. If I were to improve this project I would find Wikipedia articles that have the most similar name to the entry in Britannica so that I did not throw out a lot of data. Another one of my issues was that I always tested with the entire Britannica text, which often took a long time. I should have figured out a better way to test with samples. There are also definitely ways that I could have made my program run faster.