Software Design Mini-project 3 Project Writeup – Leon Lam

Overview

I wanted to investigate similarities and differences in writing styles and word usage between different texts, and decided to rip books from Project Gutenberg. I intended to generate a matrix of cosine similarities between different texts.

Implementation

I ripped books from Project Gutenberg, split them along the triple asterisks that marked the beginnings and ends of each book's text, then stripped any formatting (I downloaded one book in html format and got super paranoid) before splitting with whitespace and then stripping punctuation. Importing string allowed me to use string.punctuation and string.whitespace, which was really helpful.

Afterward I generated word histograms (both raw and percentile) and pickled them as strings. The actual analysis was by way of dot product – I originally used pairwise sine products of each word's percentile ratios to calculate similarity, before talking to Paul and realizing the dot product method was a faster, more elegant (and probably more accurate, although I have no basis for judgement) method.

I ended up generating a table (using tabulate) with 7 ebooks: The Importance of Being Earnest, the Kama Sutra, Journey to the Center of the World, Neuromancer, Oliver Twist (dickens), Frankenstein and Count Zero.

	importance	kama_sutra	journey	neuromancer	dickens	frankenstein	count-zero
importance	(1.00000000)	(0.64964805)	(0.76723392)	(0.69969356)	(0.74380657)	(0.80038731)	(0.72921552)
kama_sutra	(0.64964805)	(1.00000000)	(0.89037758)	(0.89045500)	(0.90434253)	(0.83916402)	(0.89603787)
journey	(0.76723392)	(0.89037758)	(1.00000000)	(0.91027909)	(0.92944083)	(0.94944497)	(0.91567941)
neuromancer	(0.69969356)	(0.89045500)	(0.91027909)	(1.00000000)	(0.95001333)	(0.85562831)	(0.98202632)
dickens	(0.74380657)	(0.90434253)	(0.92944083)	(0.95001333)	(1.00000000)	(0.90706169)	(0.96261353)
frankenstein	(0.80038731)	(0.83916402)	(0.94944497)	(0.85562831)	(0.90706169)	(1.00000000)	(0.87355813)
count-zero	(0.72921552)	(0.89603787)	(0.91567941)	(0.98202632)	(0.96261353)	(0.87355813)	(1.00000000)

Count Zero and Neuromancer are both written by William Gibson – I downloaded Count Zero because I realized I needed two works by the same author to see if the similarities would come through.

Neuromancer and Count Zero have a very high cosine similarity – 0.982. Oliver Twist (dickens) was also very close to these two books, with an average similarity of 0.955 to each book. What the cosine similarity actually means in the literary sense is not something I have explored yet, though.

Reflection

I could probably have gotten more books, but felt that 7 would be roughly enough to see if the exercise worked. I unit-tested the strip method for generating word lists on smaller chunks of text. The actual similarity generation was probably the bit that gave me the most trouble, but after learning the dot product method it became very simple. Perhaps a bit too simple – the rest was just letting Python do its thing. Maybe in the future, some sort of word net could be used to associate similar words (play and playing, jump and leap etc.), which might give a more accurate representation of the similarities and differences between the texts.