**PROJ**ECT **3 REPORT**

Theodore Block, 1003573377

# Part (a): Experimenting on a large corpus of text

The accuracy of the cosine similarity test, based on the descriptors built from the proper files, was 67.5%. The runtime on my laptop with pyzo, Word, and Chrome open, was 106.4 seconds for building the descriptor dictionary and 108.8 seconds total to compute the accuracy for each of the tests.

# Part (b): Experimenting with alternative similarity measures

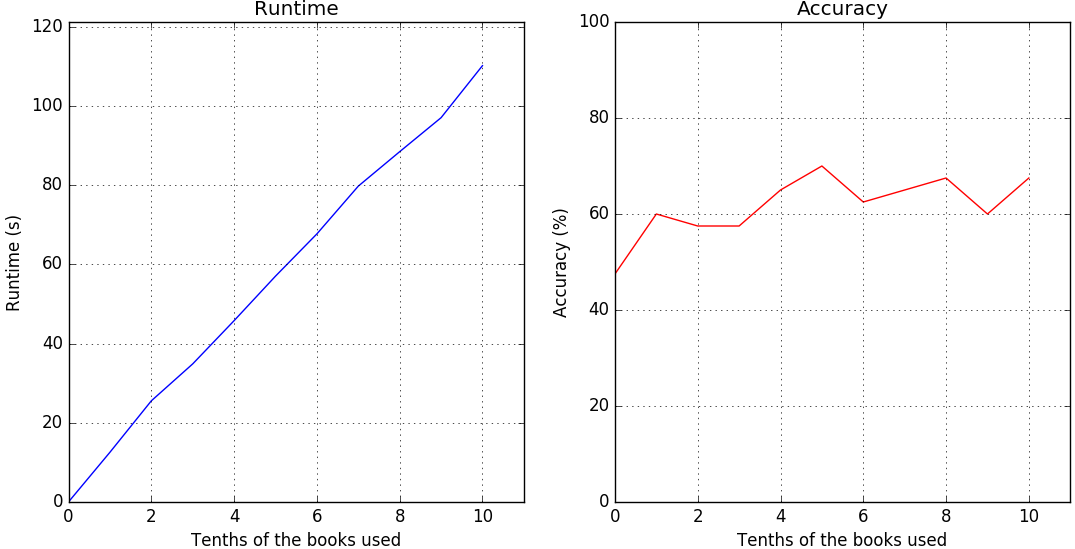
The accuracy of the Euclidean space similarity test was 30.0%.

The accuracy of the normalized Euclidean space similarity test was 67.5%.

The inaccuracy of the Euclidean similarity test comes from the fact that words which have exactly the same descriptor vectors, with one multiplied by some scalar constant, will give a dramatic distance. Normalizing each vector takes care of this issue.

# Part (c): Experimenting with smaller corpora of text: efficiency and performance

Below are the graphs of runtime and accuracy over tenths of the books used to create the descriptor dictionary:



The code used to produce these results is shown below.