Homework 01: Document Distance Problem

Due Tuesday, September 26th, 2017

Instruction

Submit your answer to this question via PC^2 under your account by the posted due time. No late submissions will be accepted. Note that homework is opened-book, but no outside assistance is permitted.

Problem

Document similarities are measured based on the content overlap between documents. With the large number of text documents in our life, there is a need to automatically process those documents for information extraction, similarity clustering, and search applications.

There exist a vast number of complex algorithms to solve this problem. One of such algorithms is a cosine similarity - a vector based similarity measure. The cosine distance of two documents is defined by the angle between their feature vectors which are, in our case, word frequency vectors. The word frequency distribution of a document is a mapping from words to their frequency count.

Write a program that ask the user to enter two documents' name, obtains the two documents' name from user and opens the documents. Finally, calculates the word frequency and cosine distance of two documents.

Sample input

```
t9.bacon.txt
t1.verne.txt
```

Sample output

```
File t9.bacon.txt : 7815 distinct words
File t1.verne.txt : 2150 distinct words
The distance between the documents is: 0.476474 (radians)
```