C++ Test Task

Document Distance Problem

Document distance tells us how different the two documents are. There are several approaches to define the document distance, one of them is cosine distance.

In cosine distance, we treat the document as a vector of words D[w]: the first part of the vector is the list of the words, the second part is how many times they are repeated. To find the similarity between them we use the angle formula between the two vectors

$$distance = \arccos(D1*D2/|D1|*|D2|)$$

The smaller the angle between them, the more similar the documents will be. Therefore, the identical documents will yield 0 radian, while absolutely different documents will yield pi/2 radian.

Please write a program to calculate the distance between two documents.

Program Interface

$ ccdist path_to_first_doc path_to_second_doc

Input: path to the first and second document.
Output: the distance

Prerequisites

- Words are only separated by spaces.
- All characters in documents are ASCII.

Requirements

- Writes program in C++ language, from C++11
- Designs in Object-Oriented Programming so that we can easily change the distance approach.
- Needs to analyze the complexity (big O) of the algorithm.
- Suggest a better distance formula if you have one.