Text Ranking report

Intro:

The script is made in python and don’t use external library.

I use the cos method with vector space model to extract a result from my file.

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

d_2 is my vector of document1 combines with the vector of document2

q the vector of my query combines with vectors of document1 and document2

Example :

Document 1: toto
Document 2: titi
query: tata
vector document1: [1, 0, 0]
vector document2: [0, 1, 0]
vector query: [0, 0, 1]

In output, the script displays the result in output of the calculation.

How to use it:

For use this script, you must install python3 and to do this:

```
./text_ranking file_name.txt file_name2.txt query_string
```

Output example:

```
Result for the query "jar obi anakin":
0.56142690767283 --> test/star_wars_phantom_menace.txt
0.5589049792726646 --> test/star_wars_attack_of_the_clones.txt

Result for the query "frodo sam":
0.4312718063731454 --> test/lord_of_the_rings_return_of_the_king.txt
0.524777068927326 --> test/lord_of_the_rings_fellowship_of_the_ring.txt

Result for the query "frodo":
0.3294332296962889 --> test/lord_of_the_rings_return_of_the_king.txt
0.5508378828115733 --> test/lord_of_the_rings_fellowship_of_the_ring.txt

Result for the query "dad":
0.006967374111920669 --> test/forrest_gump.txt
0.0018247947315469706 --> test/thor_ragnarok.txt

Result for the query "dad honest":
0.004926677481602704 --> test/forrest_gump.txt
0.0038709741868510447 --> test/thor_ragnarok.txt
```

Problem occurred:

This method is a good way to know how exactly the words are important in the texts, but it have some limits. For example, the order of the words in the query have no importance or the words can be from different sentences in the texts and they are counted as the same way. So, result is to take with gloves and could be improved.