Pa1 Report (b07801004 陳佳雯)

Environment:

Visual studio code

Langage:

Python3

Execute:

1. Install nltk.(if needed)
2. Require “dictionary”, “Doc”, “dictionary” empty files and “NLTK_stopWord.txt” before running the program.
3. Require “IRTM” files with 1095 documents inside.
4. Run pa2.py.
5. Output will be:
   i. “dictionary.txt”.
   ii. “Doc” file with each document’s tf-idf unit vector.
   iii. “dictionary” file with each document’s term frequencies.
   iv. Cosine similarity of two documents that user inputs.

Discription:

1. Use pa1 function to tokenize, remove stopwords and stem each document.
2. For each document, record the term frequencies. Save as DOCID.txt in “dictionary” file.
3. After traverse all the documents, we use dict_df (type: dictionary) to record the idf in “dictionary.txt”. Besides, we use dict_term (type: dictionary) to record the term index for the dictionary.

$$idf_t = \log_{10} \frac{N}{df_t}$$

4. For each document, we compute the tf-idf by document frequencies from “dictionary.txt” and term frequencies from DOCID.txt. Save the result into DocID.txt in Doc file.
5. Ask user to input two documents’ ID and compute their unit vectors separately.
6. Find terms which appears both in doc1.txt and doc2.txt. Compute their cosine

similarity.

Output:

Please input the ID number of document1:1

Please input the ID number of document2:2

Cosine Similarity: 0.1806011369047524