Programming Assignment 2 (1/2)

- Write a program to convert a set of documents into tf-idf vectors.
  - Text collection:
    - 1095 news documents

      (https://cool.ntu.edu.tw/files/1239294/download?download_frd=1)

      zip code: IRTM2021
  1. Construct a dictionary based on the terms extracted from the given documents.
     - Record the document frequency of each term.
     - Save your dictionary as a txt file (dictionary.txt).

t_index	term	df
1	Apple	3
2	Basketball	12
…		

出現在幾篇文章

ascending order, by term

dictionary.txt

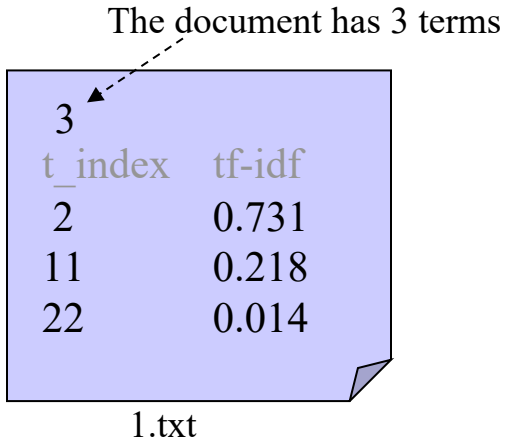
tf(t, d)=某個字 t 在單篇文章 d 中出現的次數

Programming Assignment 2 (2/2)

2. Transfer each document into a **tf-idf unit vector**.

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

- Save it as a txt file (DocID.txt).



3. Write a function `cosine(`Doc_x`,` Doc_y`)` which loads the tf-idf vectors of documents x and y and returns their cosine similarity.

- **Please zip and submit [1]your dictionary, [2]the vector file of document 1, [3]source code, and [4]a report to TA.**
  - **Also mention the cosine similarity between document 1 and 2 in your report.**
  - 3 weeks to complete, that is, **2021/11/16**.