Programming Assignment 2 (1/2)

- Write a program to convert a set of documents into tf-idf vectors.
 - Text collection:
 - □ 1095 news documents (https://ceiba.ntu.edu.tw/course/b079e8/content/IRTM_news_files.zip_)
 - 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
...
ascending order, by term
dictionary.txt
```

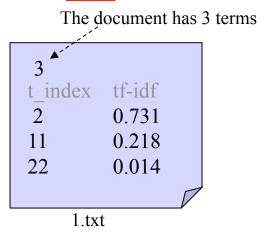
Programming Assignment 2 (2/2)

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).





- 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, 2012/10/23.