## Programming Assignment 2 (1/2)

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

```
(<a href="https://cool.ntu.edu.tw/files/1239294/download?download_frd=1">https://cool.ntu.edu.tw/files/1239294/download?download_frd=1</a> ) zip code: IRTM2021
```

- Construct a dictionary based on the terms extracted from the given documents.
  - Record the <u>document frequency</u> of each term.
  - □ Save your dictionary as a txt file (dictionary.txt).

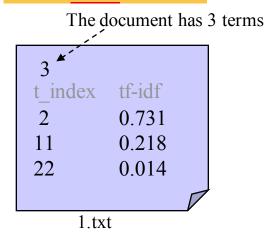
```
t_index term df
1 Apple 3
2 Basketball 12
...
ascending order, by term
dictionary.txt
```

## 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 (<u>DocID.txt</u>).



- Write a function  $\frac{cosine(Doc_x, Doc_y)}{cosine(Doc_x, Doc_y)}$  which loads the tf-idf vectors of documents x and y and returns their cosine similarity.
- Please zip and submit <sup>1</sup> your dictionary, <sup>2</sup> the vector file of document 1, <sup>3</sup> source code, and <sup>4</sup> a report to TA.
  - Also mention the <u>cosine similarity between document 1 and 2</u> in your report.
  - 3 weeks to complete, that is, 2021/11/16.