Information Retrieval Homework, Report II

R08725002 資管碩二 黃文鴻

1. Environment

```
VSCode
```

2. Programming Language

```
Python 3.8
```

3. Usage

```
pip3 install nltk
python3 ./pa1.py
```

4. Introduction

`In line 18-24` , get the directory file list and sort by number

`In line 26-38` , tokenize document with porter stemmer and remove punctuation exclude hyphens

`In line 40-65` , filter empty strings and tokens started with number or punctuation, then calculate tf and df in each document

`In line 67-77` , generate collection-wide df dictionary

`In line 79-96` , calculate idf and tf-idf and save as files

`In line 98-124` , get tf-idf vector of document and calculate cosine similarity

Document 1 and 2 cosine similarity = 0.5237166991853831