Information Retrieval Homework, Report IIII

R08725002 資管碩二 黃文鴻

1. Environment

```
VSCode
```

2. Programming Language

```
Python 3.8
```

3. Usage

```
pip3 install nltk tqdm numpy
python3 ./pa4.py
```

4. Introduction

`In tokenize_data():` , tokenize document with porter stemmer and remove punctuation exclude hyphens

`In read_dataset():` , read all dataset into a list of tokens

`In cosine_similarity():` , calculate cosine similarity with tf-idf vectors of 2 documents

`In line 68-114` , calculate tf-idf of all documents and convert into a numpy array

`In simple_hac():` , use the hac algorithm with complete link for similarities between clusters, outputs: similarities matrix & hac merges list

`In line 164-183` , group doc id into clusters with hac merges list

`In line 187-192` , save clusters into files with k=8, 13, 20