Introduction to Information Retrieval and Text Mining

Programming Assignment 2 Report

tf-idf vectors

資管四 徐承志 B98705034・NTU・2012/10/23

1.HOW TO EXECUTE YOUR PROGRAM

Execution Environment：

Mac OSX、Windows、Linux with Ruby Programming Language. If you do not have Ruby setup, please install ruby first.

我使用的程式語言是Ruby programming language，OS是在MAC上的OSX10.8.2，使用的ruby version is 1.9.2 .

```
MichaelHsu-2:code michaelhsu$ ruby -v
ruby 1.9.2p290 (2011-07-09 revision 32553) [x86_64-darwin12.0.0]
```

Getting Start：使用CLI執行程式，執行順序如下

CLI 步驟	結果（output）
ruby 0_loop-to-extract.rb	/output/terms_hash/1.txt
ruby 1_construct-dictionary.rb	/output/dictionary_hash.txt
ruby 2_tf-idf.rb	/output/tf-idf_hash/1.txt
ruby 3_cosine_similaity.rb 1 2	cosine similarity of document 1 and 2 0.18284016135760092
ruby 4_format.rb	dictionary.txt vector1.txt

2.EXECUTION TIME WITH YOUR HARDWARE SPEC

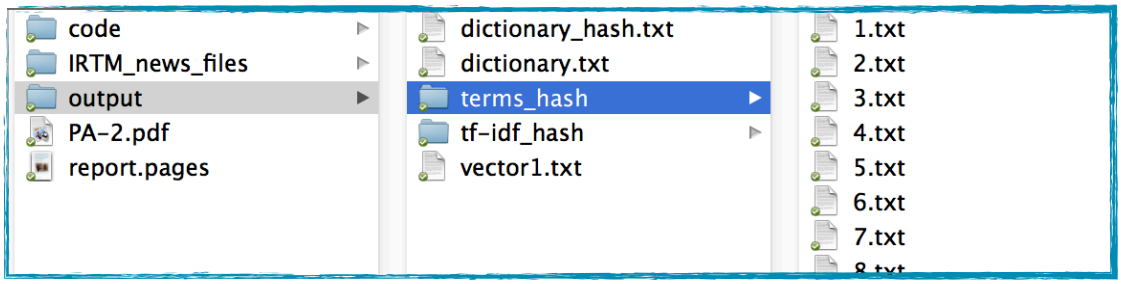
機型名稱	MacBook Pro
處理器名稱：	Intel Core i7
處理器速度：	2.3 GHz
總核心數目：	4
L3 快取記憶體	6 MB
記憶體：	8 GB
Execution Time	1 min

3.YOUR PROGRAM DESIGN & PROCEDURE

首先我挑選ADT dictionary 作為這次的資料形態，而我所挑選的oop語言有實作出這部分hash table(hash)，因為一個key對應到一個value，比較方便去使用，所以就直接拿來運用了。接著我把Assignment切割五個部分來完成，對應到source code依序為：

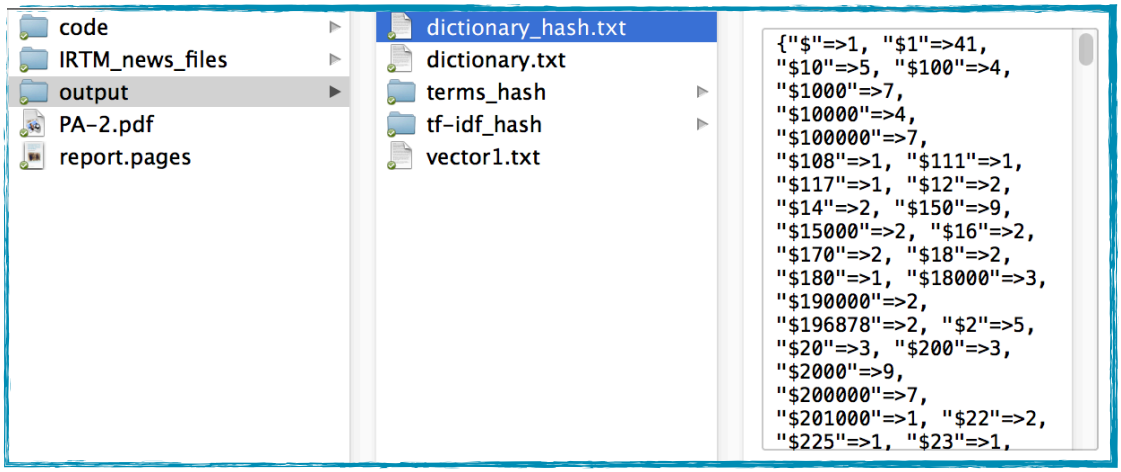
0_loop-to-extract.rb

利用pa1的extractor找出每一份document的terms，存在/output/terms_hash/中



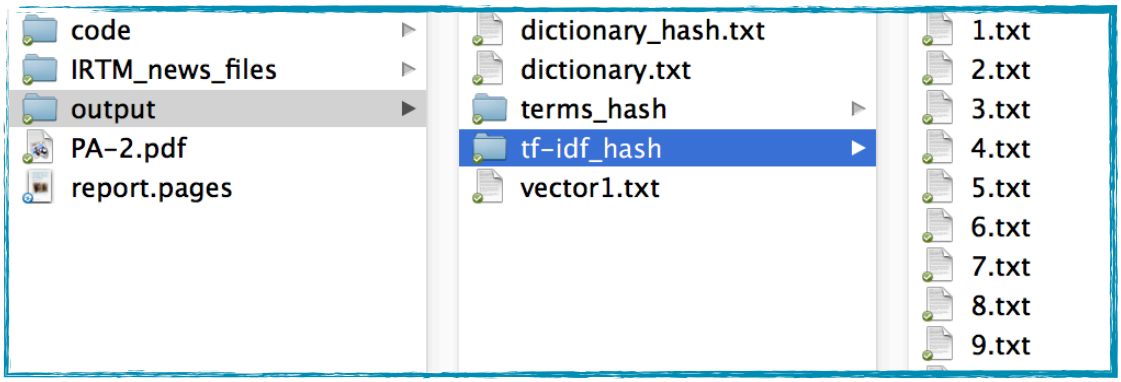
1_construct-dictionary.rb

把所有出現的term 做出一個dictionary，我在使用上先存進hash table，然後output file 為一個dictionary_hash.txt



2_tf-idf.rb

l讀取每一份document的terms，對應到dictionary_hash.txt，算出tf-idf-unit-vector，並且存放到/output/tf-idf_hash中



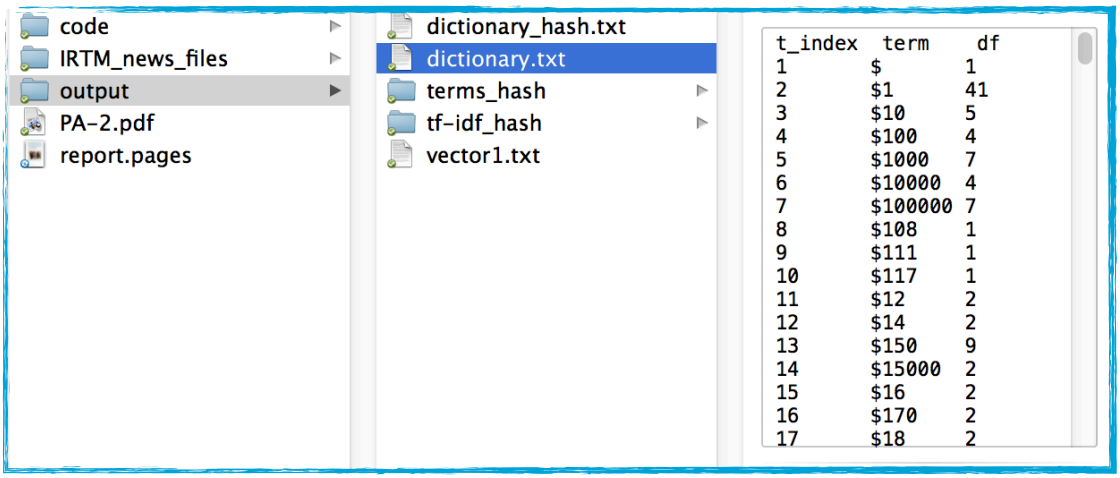
3_cosine_similaity.rb

丟進兩個argument，並且開啓這兩份的tf-idf_hash進行運算，得到cosine_similaity的值並且印出來

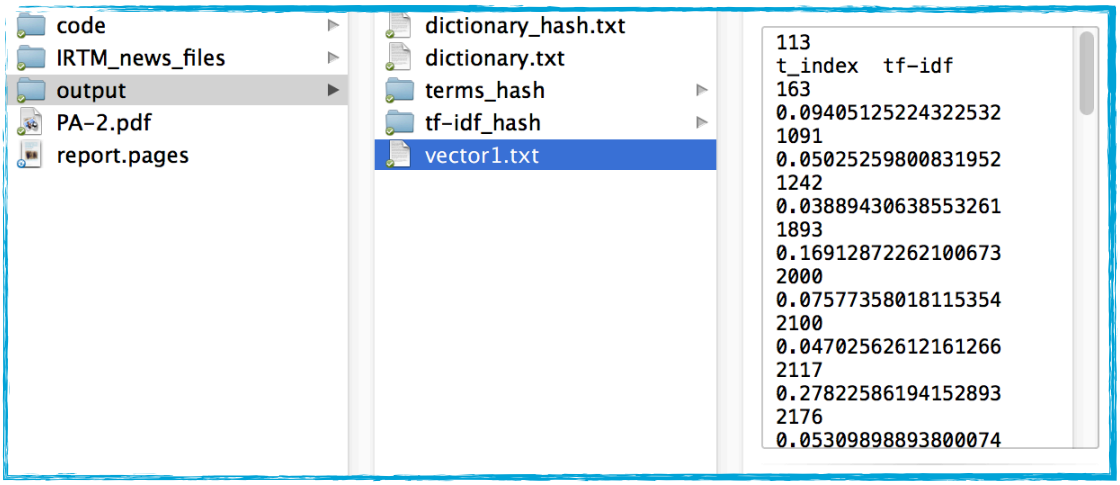
```
MichaelHsu-2:code michaelhsu$ ruby 3_cosine_similaity.rb 1 2
cosine similarity of document 1 and 2
0.18284016135760092
```

4_format.rb

因為繳交的格式的限制，另外寫一段code把dictionary_hash.txt轉換成易讀格式dictionary.txt



以及the vector file of document 1



4.ADVANTAGE OF YOUR PROGRAM

易讀性高，又不失效能。

5. DISCUSSIONS

我在做這次Assignment最大的困惑在於資料結構的挑選上，因為spec上輸出的格式並不會比較好拿來使用，最後決定還是先把所有資料存放在hash table中，拿來做完所有的運算後再輸出程比較易懂的spec要求格式。在考慮的過程中有發現有些同學是直接把資料放在database裡面，雖然這也是一個運用的好方法，但是執行效率卻大打折扣。

6. RESOURCE & REFERENCE

1. ruby http://www.ruby-lang.org/zh_TW/downloads/
2. ruby array http://www.ruby-doc.org/core-1.9.2/Array.html
3. ruby hash http://www.ruby-doc.org/core-1.9.2/Hash.html