Introduction to Information Retrieval and Text Mining

Programming Assignment 3 Report

Multinomial NB Classifier

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，並且取得輸出結果

CLI 步驟	輸出（output）
ruby 0_Training_Extractor.rb	/output/training_dictionary_hash.txt
ruby 1_Compute_LLR_value.rb	/output/LLR_value.txt
ruby 2_SelectFeatures.rb	/output/features.txt
ruby 3_training_phase.rb	/output/condprob.txt
ruby 4_testing_phase.rb	/output/testing_result.txt
ruby 5_output_format.rb	/output/output.txt

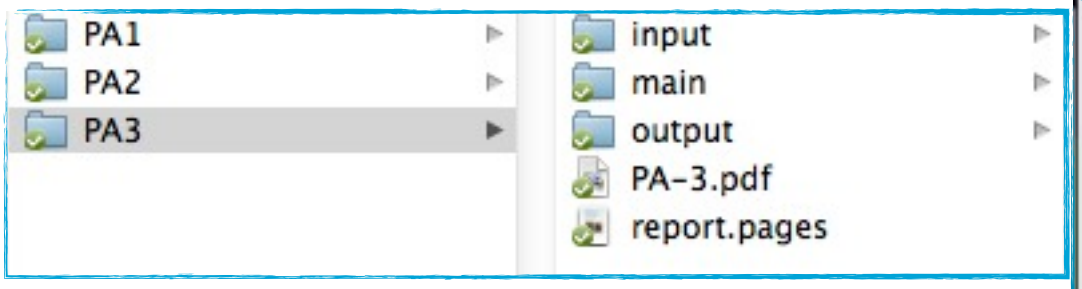
2.EXECUTION TIME WITH YOUR HARDWARE SPEC

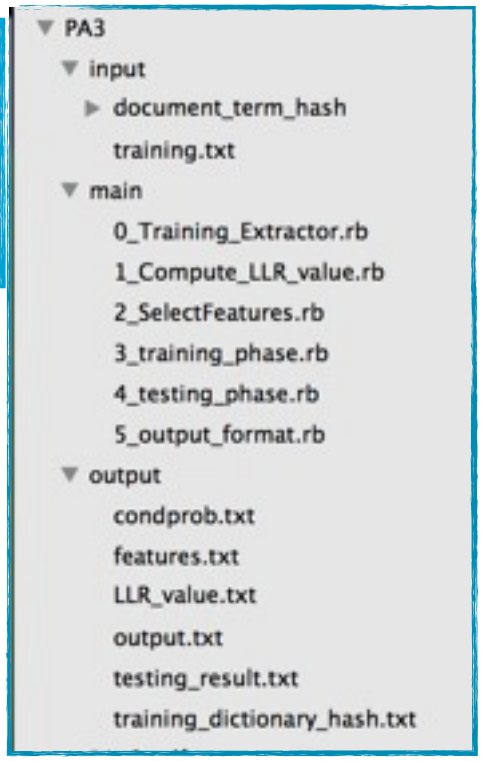
機型名稱	MacBook Pro
處理器名稱：	Intel Core i7
處理器速度：	2.3 GHz
總核心數目：	4
L3 快取記憶體	6 MB
記憶體：	8 GB
Execution Time	1 min

3.YOUR PROGRAM DESIGN & PROCEDURE

依據Multinomial NB，將training、testing的演算法實作出來， 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。

4. FRAMEWORK





/input	
document_term_hash	PA2取出的每一份document的terms，格式hash table
training.txt	老師給的train set分類

/main	
0_Training_Extractor.rb	把所有Training Document set 出現的term 做出一個dictionary出來
1_Compute_LLR_value.rb	算出每一個term 的LLR value
2_SelectFeatures.rb	平均選擇，select the top k/n features for each n classifiers
3_training_phase.rb	training phase
4_testing_phase.rb	testing phase
5_output_format.rb	輸出作業需求的格式

/output	
training_dictionary_hash.txt	training set 取出的dictionary 用hash table存放
LLR_value.txt	每個term的LLR value
features.txt	最後取出的 feature 500 terms
condprob.txt	training phase 每個term的是哪個class的機率
testing_result.txt	testing phase
output.txt	最後結果

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