Ruby Rocchio Classifier

The Programming Assignment report from NTU102-1 DMIR course

by NTU Michael Hsu

1. How to execute 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 OSX10.9，使用的 ruby version is 2.0.0p195 .

```
[~ ]
$ ruby -v
ruby 2.0.0p195 (2013-05-14 revision 40734) [x86_64-darwin12.3.0]
```

Getting Start：使用 CLI 執行程式，執行順序如下

```
$ cd code && ruby Rocchi_Classifier.rb
```

-

```
[~/Dropbox/15. 碩一上課業/02. DMIR 資料探勘與資訊檢索/PA/code ]
$ ruby Rocchi_Classifier.rb
>> training phase @ class 1
>> training phase @ class 2
>> training phase @ class 3
>> training phase @ class 4
>> training phase @ class 5
```

Output result `output.txt`

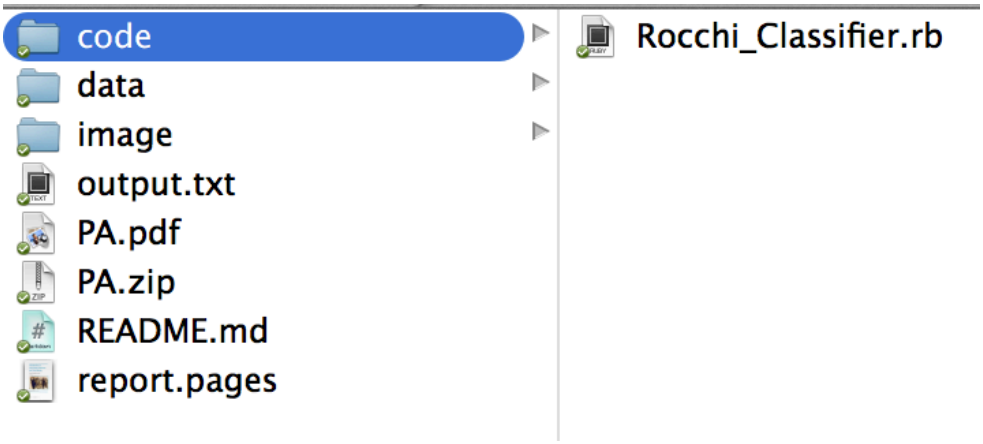
```
doc_id  class_id
17  2
18  2
20  2
21  2
22  2
...
```

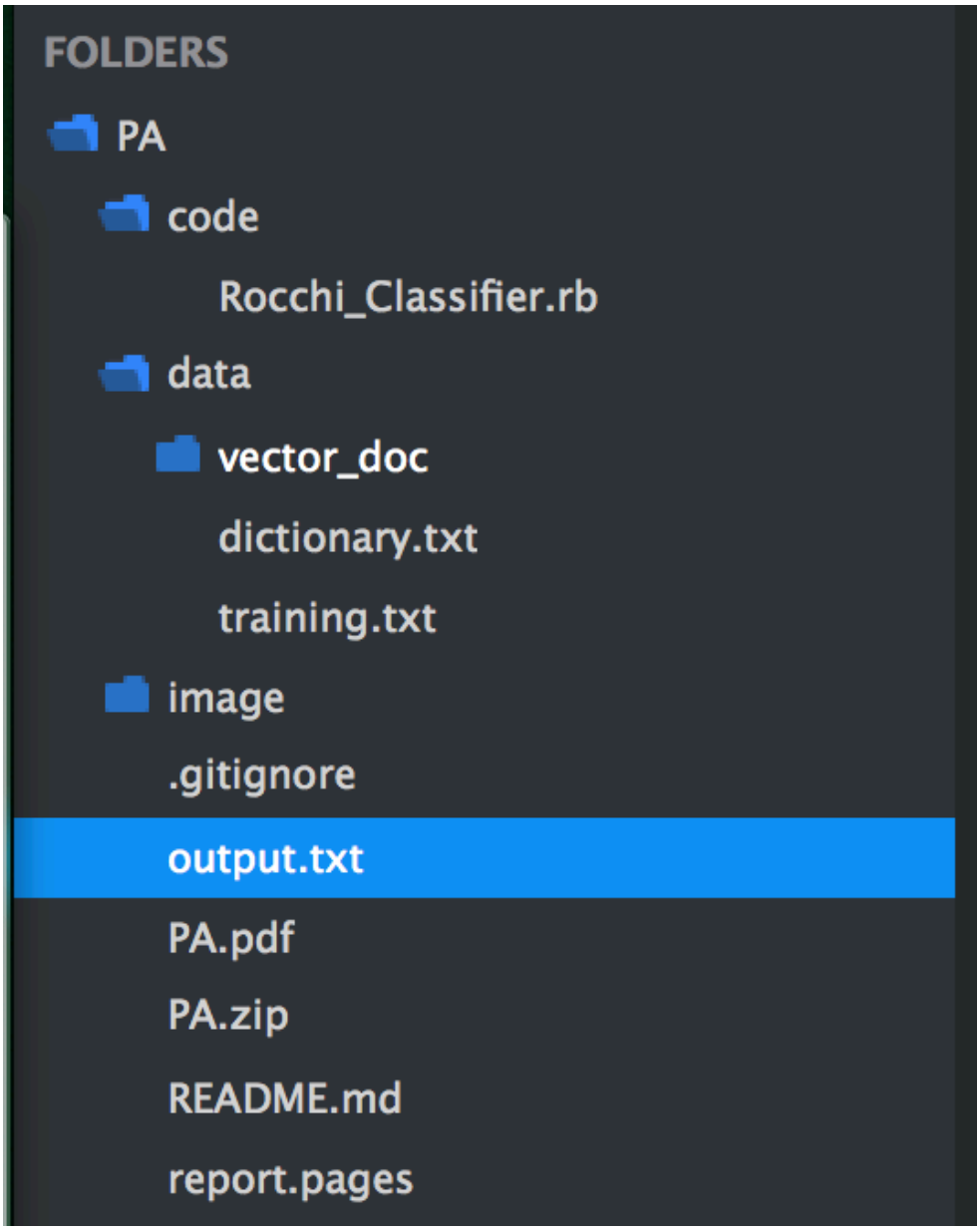
- *其中用* `\t` *tab 符號做分隔。*

2. Your program design & procedure

我用最簡單的資料形態去實作，然後用single-link做merge。 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。

3. Project framework





- 依照投影片上的演算法分做 training phase、testing phase 兩段來實作。

- Rocchio training:

```
TrainRocchio(C,D)
  for each c_j in C
  do
       D_j = {d : <d,c_j>in D}
       u_j = Σ_{d in Dj} v(d)/|D_j|
  return {u_1,...,u_j}
```

- Rocchio testing:

```
ApplyRocchio({u_1,...,u_j},d)
  return argmin_j|u_j-d|
```

4. Advantage of your program

易讀性高，又不失效能。

5. Discussions

我在做這次 Assignment 實做演算法的過程中沒有發現到原來 centroid 沒有出現的字但是 testing document 的 term 也需要減零做計算；相反亦同，當 testing document 有出現的 term 但是 centroid 沒有出現，也要減零做計算。但這部分一直沒有想到比較好的方法，如果要將兩邊都個跑一遍迴圈會使得效率大幅降低，因此日後還是有改善的空間。

6. Resource & Reference

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