Sentimental Analysis 正反情緒分析

* Environment

Eclipse Java EE

JDK/JRE v1.6

UTF-8 File Encoding

* Process

pos\_by\_training.txt

training.txt

neg\_by\_training.txt

result.txt

answer.txt

adv.txt

positive.txt

negative.txt

opinion.txt

除了原有的外部情緒字典（正、反、程度詞），利用Training找出其他特別的正反詞彙

之後，根據句中「正反詞彙」出現的數量，搭配加重語氣的程度詞，給定一分數，作為評斷標準

* Frame Work

SentimentAnalyzer.java

9. Start Analyzing

SentimentalDictionary.java

KeyWordFinder.java

TextReader.java

7. Read and Arrange Words into Hash Map

5. Find out Frequent Words

3. Segment Word into Pieces

2. Store

8. Read

FrequencyRecorder.java

6. Append to Dictionary

4. Count Frequency

SegChinese.java

TextReader.java

1. Read

* Details about Training

Step1　斷詞

先由標點符號斷句，並使用Open Source的Library（MMSeg）

實現最大匹配、最大單詞長度的分詞

Reference: <http://function1122.blogspot.tw/2010/10/mmseg4j-java-55.html>

Step2　計算各單詞出現次數（頻率）

得到以「詞」為單位的資料後，計算整份Training Data中，各單詞出現的字數（頻率）

Step3　選擇一些在該類文章中，具代表性的正負面詞彙，加入字典

選擇的標準：SO值 > 4.5，加入Positive字典；SO值 < -4.5，加入Negative字典

* Details about Analyzing

Step1　斷句

以標點、各式符號斷句（不以分行斷句，因為一行視為一則評論或回覆）

Step2　找程度詞

將一個句子切分成小部分，判斷截斷後的詞彙是否屬於Dictionary中的程度詞

如果是，則將該句子的分數倍率乘以2

* Example

「這家旅館的爛服務非常差勁」會切成

「家旅館的爛服務非常差勁」、「這家旅館的爛服務非常差」…「常差」、「非常」…「這」

由長到短、後往前的截字方式（避免長詞關鍵字沒先抓到，反而抓到短詞）

抓到程度詞關鍵字後，會將倍率乘2，並把關鍵詞從句子中刪除

Step3　找正反面情緒用詞

截字、刪字方式同上一步，只是把截斷後的詞彙拿去Positive、Negative Dictionary中比對

比對後，如果是正面詞彙，分數+1，負面則-1（搭配程度詞的倍率，可能變為±2）

Step4　找出Shifter（不、沒）

比對句子中剩餘的字彙，是否包含「不」或「沒」

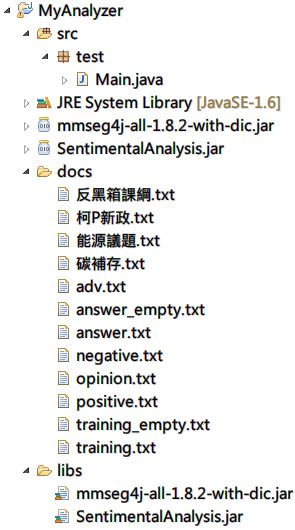
如果有，則將該句子的分數乘上-1

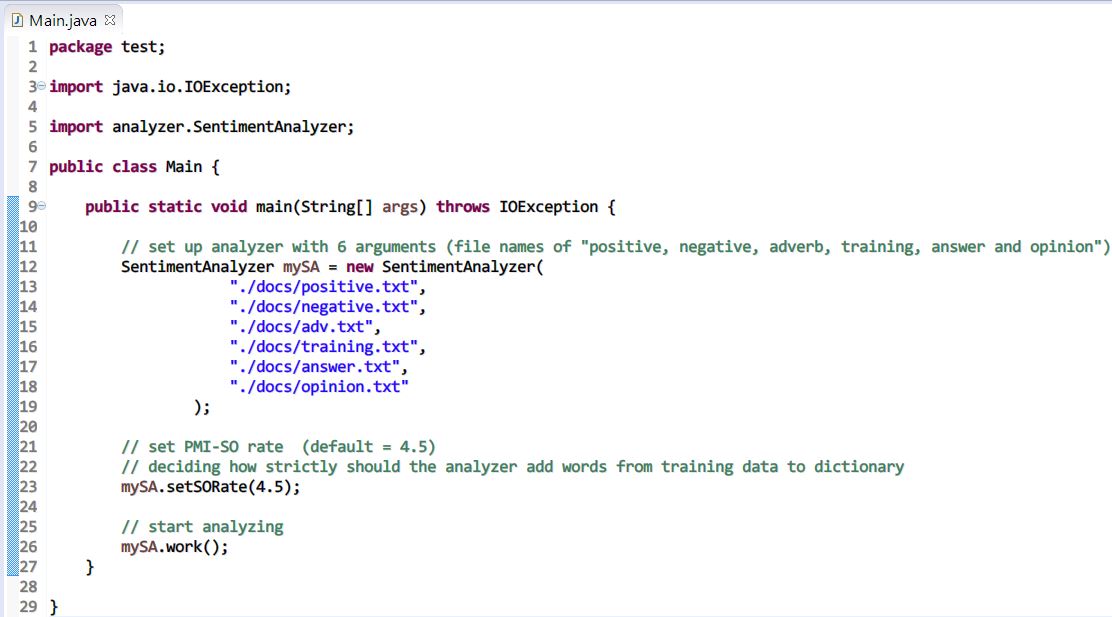
Step5　判斷整則評論的正反傾向

整則評論的分數=各句子的分數加總，若評論分數 ≧ 0，判為正面傾向，反之負面

* API

Source Code已打包成SentimentalAnalysis.jar，外加mmseg4j-all-1.8.2-with-dic.jar

Library Setting好之後，使用SentimentAnalyzer()建構子和method－work()來run



* Input File Format
  + Training用的文字檔 預設為docs/training.txt

一則評論占一行，不加編號（若無，須建立空檔案）

* + Training的答案 預設為docs/answer.txt

行數與training.txt相同，一行一字，以半形大寫P/N來表示（若無，須建立空檔案）

* + 正、反、程度字典 預設為docs/positive.txt, docs/negative.txt, docs/adv.txt

一個單詞（單字）占一行，不加編號

* + 欲分析的評論 預設為docs/opinion.txt

格式與Training的檔案相同，一則評論占一行，不加編號

* Output File Format

分析後會於當前目錄產生result.txt

result.txt中每則評論的分析占4行：

Line1 「NO.%d rate = %d (Positive)」 或　「NO.%d rate = %d (Negative)」

Line2 原評論的斷詞結果

Line3 「Keywords Found: 」+數個「%s(+1、-1或adv) 」，為找到的關鍵字和其意義

Line4 空行