亞洲大學Asia University

雲端資料處理與探勘

Cloud Data Management and Mining

期末專題：PubMed 資料分析專題

指導老師:

王經篤

系級:

資傳3B

學號姓名:

102063099洪子涵

102063107黃郁嵐

**目錄:**

1. 執行過程與結果
2. 找出『標題』出現最多的前20名的英文單字
3. 找出『摘要』出現最多的前20名的英文單字
4. 比較『標題』和『摘要』的結果有何不同
5. 根據『PubMedID』的最後一碼
6. 統計檔案出現次數
7. 找出『PubMedID』和你學號最後一碼相同的『PubMedID』
8. PubMed (Pattern History)
9. 計算一個pattern在不同時間區間的出現次數
10. [PatternHistory\_SumByMonth.7z](http://dns2.asia.edu.tw/~jdwang/TeachingCourses/CloudDataManagement/PatternHistory_SumByMonth.7z)改寫成PatternHistory\_SumByYear.7z
11. 找出只有出現在2013年的pattern
12. 討論與問題
13. 小組成員工作分配
14. 個別貢獻度百分比
15. 個人心得
16. **執行過程與結果**
17. 找出『標題』出現最多的前20名的英文單字

檔案：WordCount\_Project\_jdwang2014\_11\_12

修改程式：Mapper4a\_jdwang

/\*if (OneArticle.length == 3) {

// This article is without the "abstract"

SrcText = Title;

}

else

{

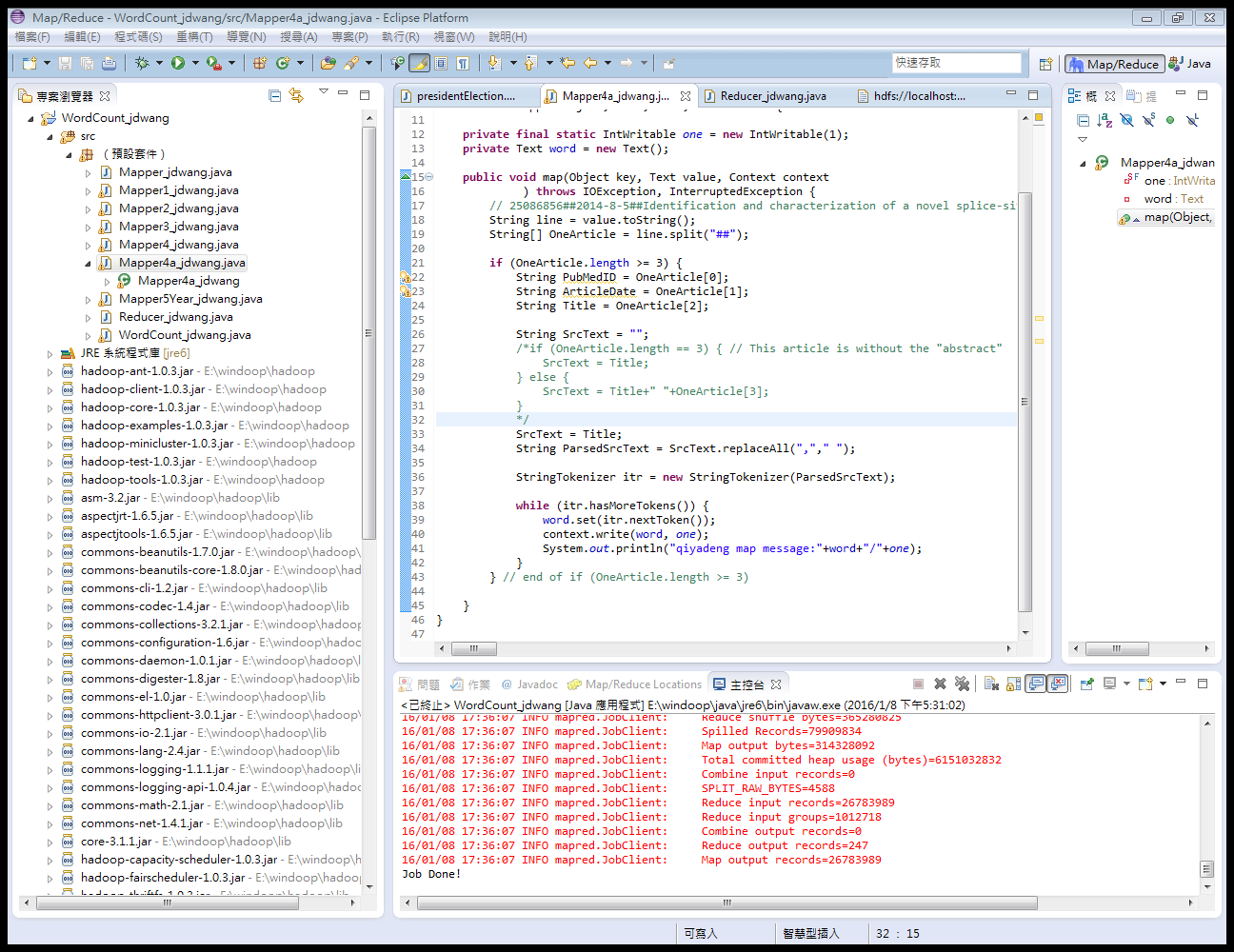
SrcText = Title+" "+OneArticle[3];

}

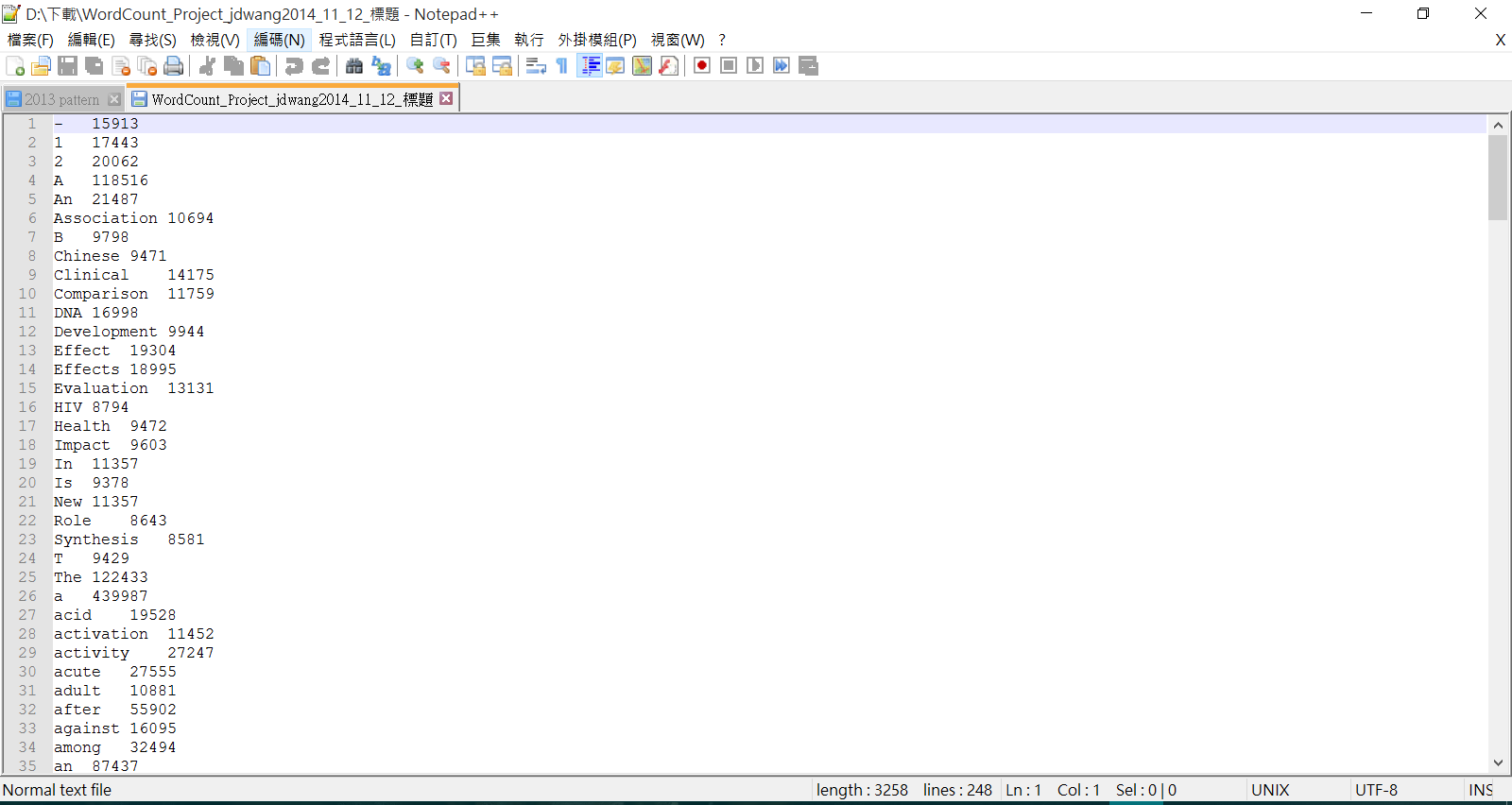
\*/

SrcText = Title;

程式執行成功：



程式執行結果：



『標題』出現最多的前20名的英文單字：

|  |  |  |
| --- | --- | --- |
| 1 | of | 1578952 |
| 2 | in | 1035248 |
| 3 | and | 1028642 |
| 4 | the | 663684 |
| 5 | a | 439987 |
| 6 | for | 358358 |
| 7 | with | 340978 |
| 8 | to | 253715 |
| 9 | on | 207100 |
| 10 | by | 147300 |
| 11 | from | 124034 |
| 12 | The | 122433 |
| 13 | A | 118516 |
| 14 | patients | 89125 |
| 15 | an | 87437 |
| 16 | as | 66703 |
| 17 | cell | 62152 |
| 18 | is | 61741 |
| 19 | using | 59809 |
| 20 | after | 55902 |

1. 找出『摘要』出現最多的前20名的英文單字

檔案：WordCount\_Project\_jdwang2014\_11\_12

修改程式：Mapper4a\_jdwang

String SrcText = "";

**if** (OneArticle.length == 4) {

// This article is without the "abstract"

SrcText = OneArticle[3];

}

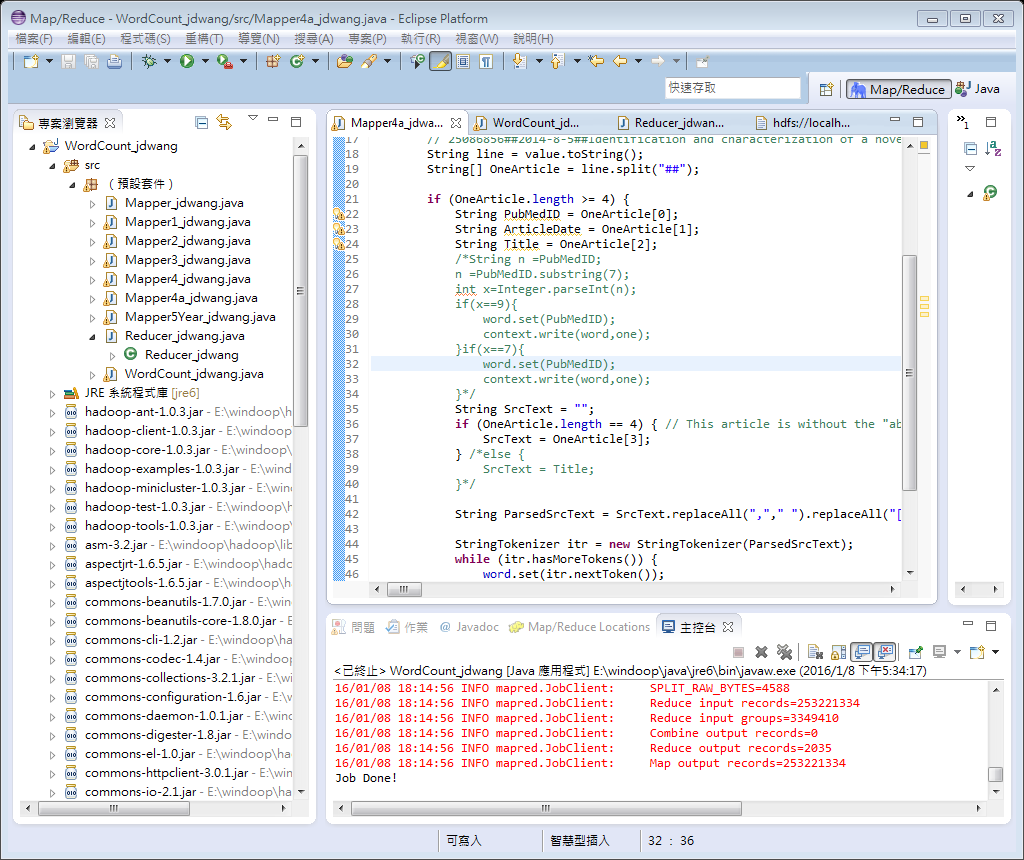
/\*else

{

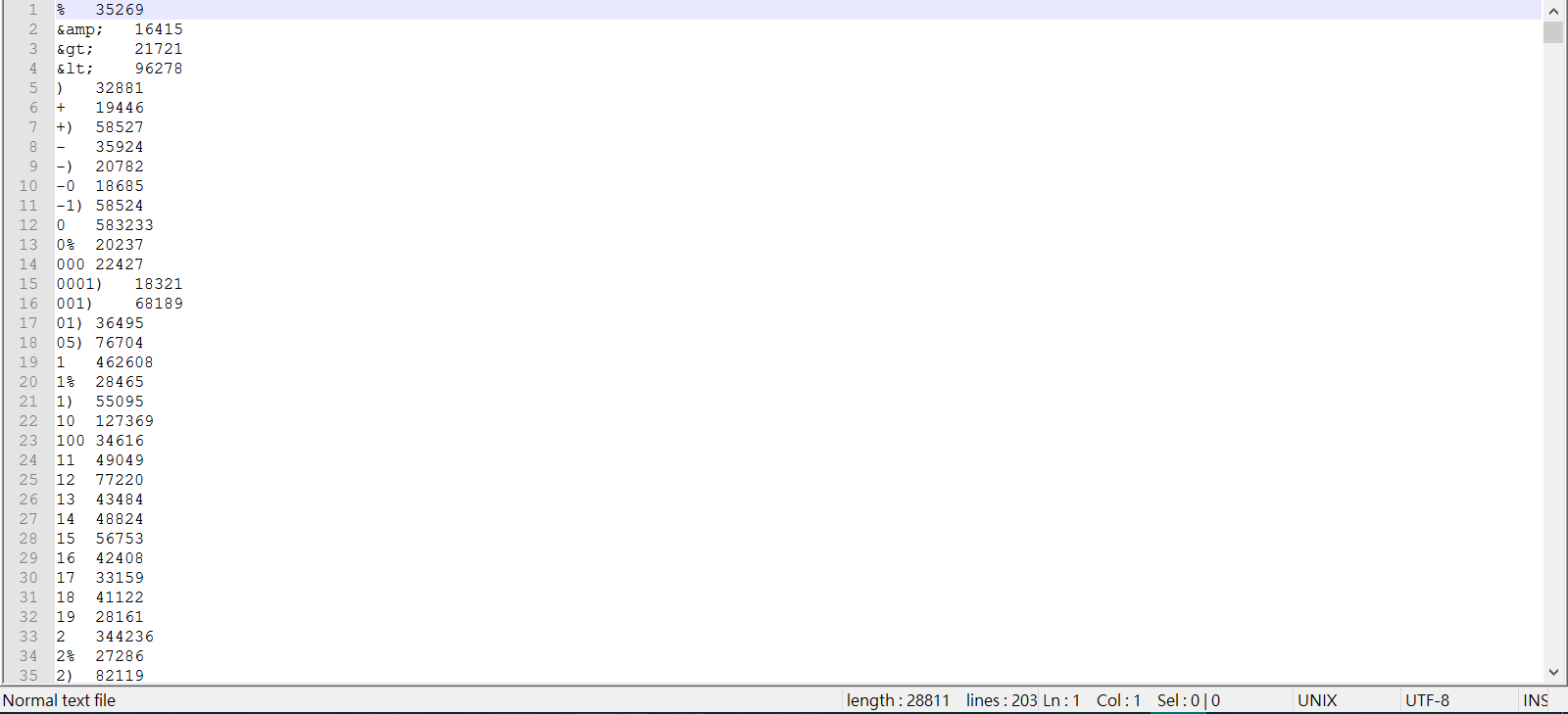
SrcText = Title;}\*/

String ParsedSrcText = SrcText.replaceAll(","," ").replaceAll("[.]", " ").replaceAll("[(]", " ");

程式執行成功：



程式執行結果：

****

『摘要』出現最多的前20名的英文單字：

|  |  |  |
| --- | --- | --- |
| 1 | the | 11259883 |
| 2 | of | 10509228 |
| 3 | and | 9049897 |
| 4 | in | 5969319 |
| 5 | to | 4681356 |
| 6 | a | 3753928 |
| 7 | with | 2850255 |
| 8 | for | 2365606 |
| 9 | that | 1980071 |
| 10 | is | 1971963 |
| 11 | The | 1936708 |
| 12 | was | 1932091 |
| 13 | were | 1792868 |
| 14 | by | 1507758 |
| 15 | as | 1284537 |
| 16 | on | 1243726 |
| 17 | are | 1073697 |
| 18 | from | 1025867 |
| 19 | an | 875068 |
| 20 | be | 869531 |

1. 比較『標題』和『摘要』的結果有何不同



**用底線標示的是重複出現一樣的單字**

1. 根據『PubMedID』的最後一碼
2. 統計檔案出現次數

檔案：WordCount\_Project\_jdwang2014\_11\_12

修改程式：Mapper4a\_jdwang

if (x==9){

word.set(n);

context.write(word, one);

}

If(x==8){

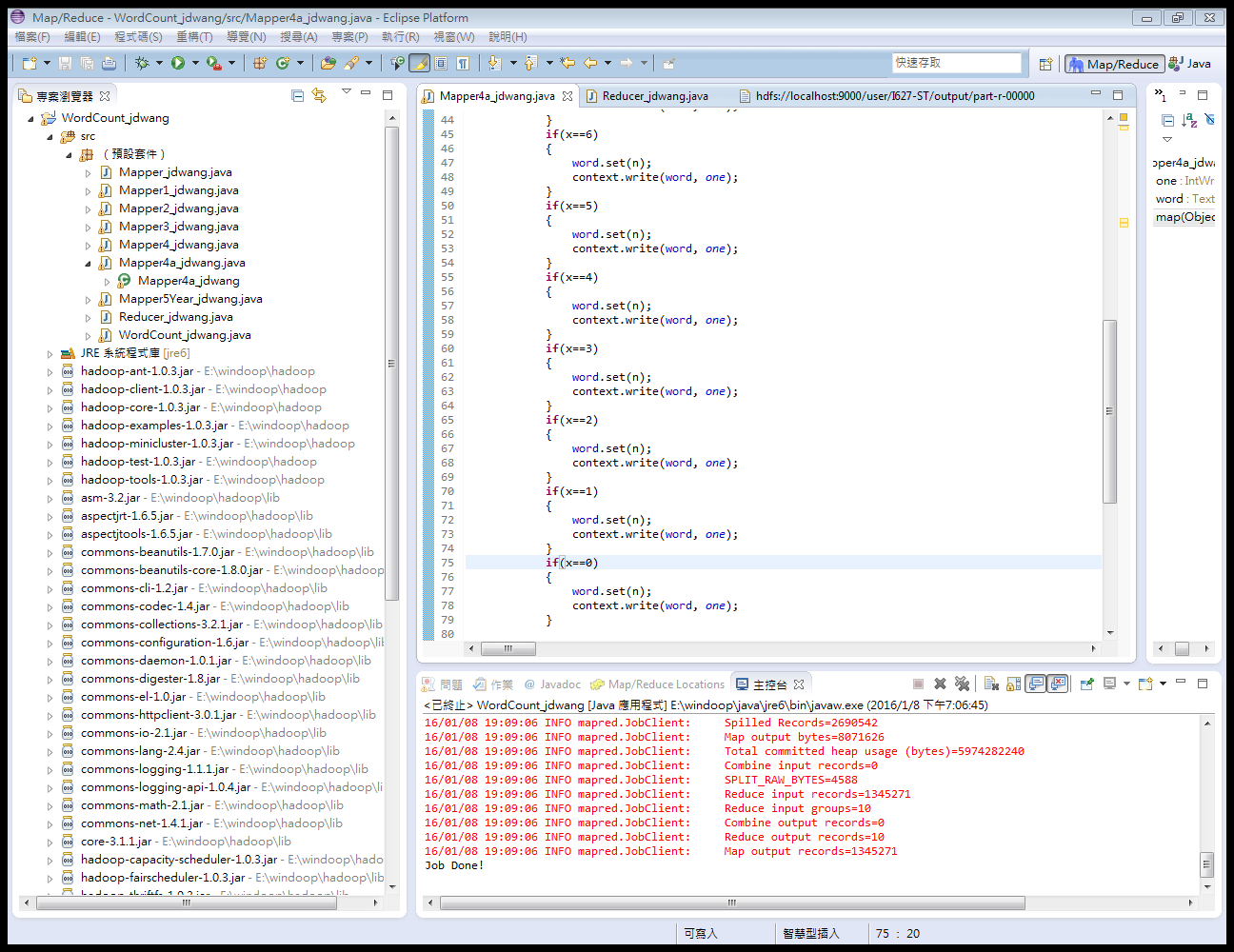
word.set(n);

context.write(word, one);

}

一直到(x==1)

程式執行成功：



程式執行結果：



1. 找出『PubMedID』尾碼和你學號最後一碼相同的『PubMedID』

檔案：WordCount\_Project\_jdwang2014\_11\_12

洪子涵學號：102063099尾數9

黃郁嵐學號：102063107尾數7

修改程式: Mapper4a\_jdwang

String n=PubMedID;

n=PubMedID.substring(7);

**int** x=Integer.*parseInt*(n);

**if**(x==9){

word.set(PubMedID);

context.write(word, *one*);

}

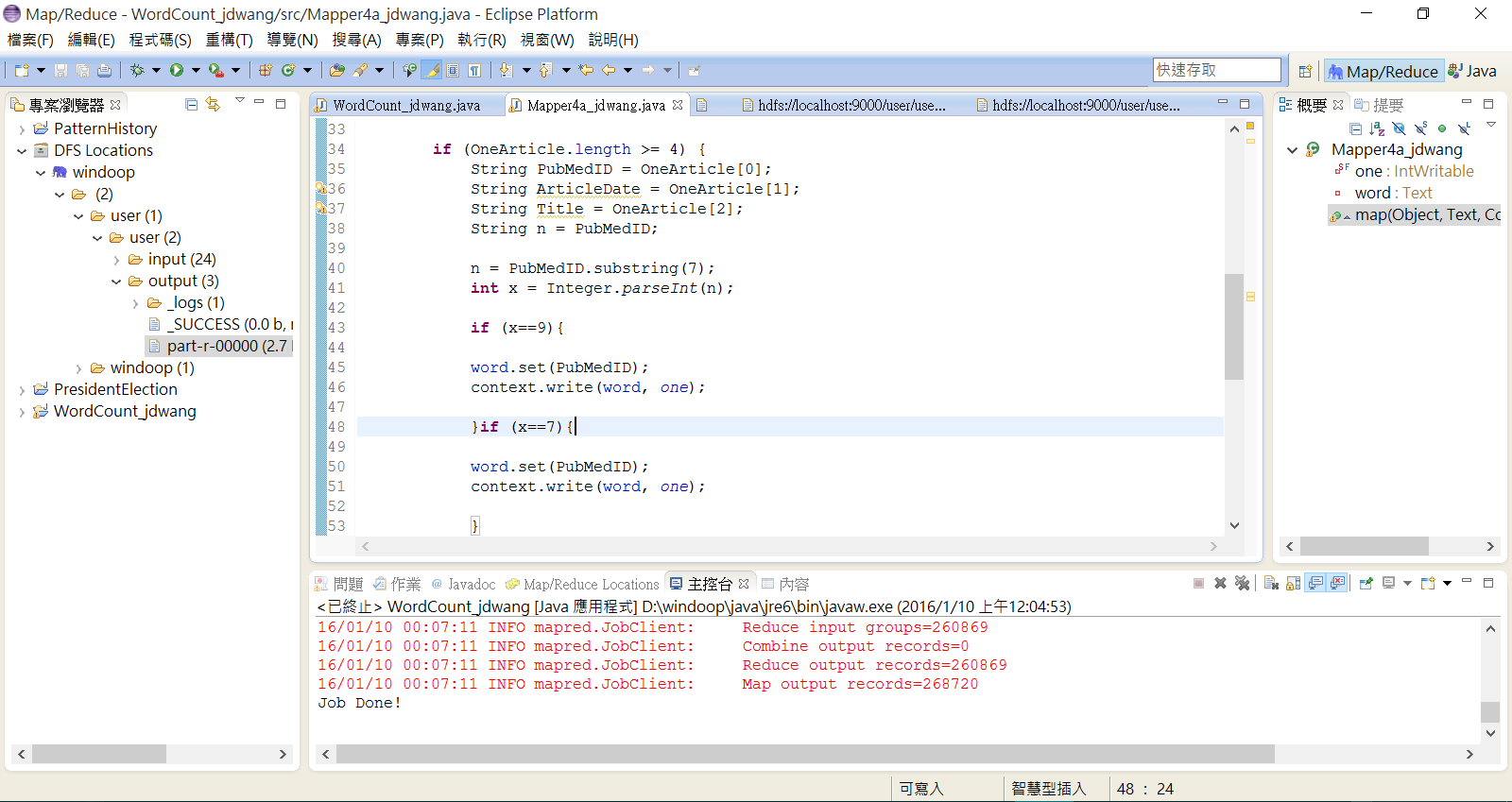
**if**(x==7){

word.set(PubMedID);

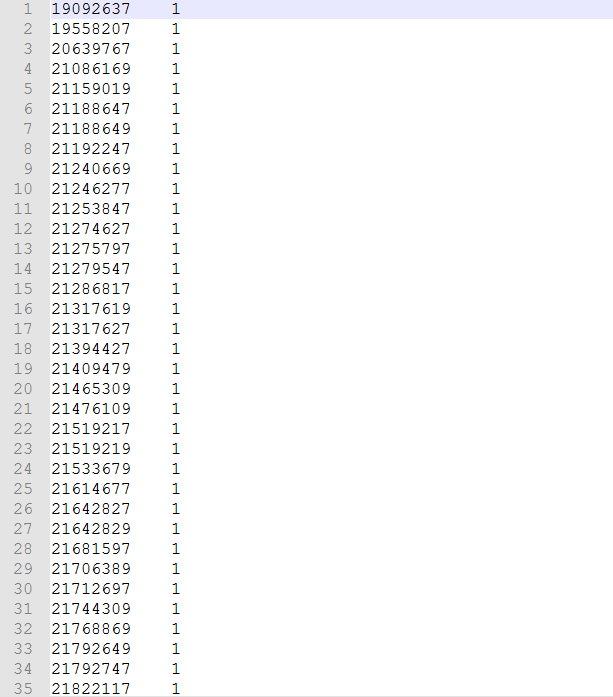
context.write(word, *one*);

}

程式執行成功：



程式執行結果：



1. PubMed (Pattern History)
2. 計算一個pattern在不同時間區間的出現次數

檔案：PatternHistory

程式修改：Mapper\_SumByYear

String Year = TimeTag[0]; // "2011"

//String Month =TimeTag[1];// "1"

**if** (MonthTFSum.containsKey(Year==”2013”)) {

Integer Orginal = MonthTFSum.get(Year);

Integer sum = Orginal+TempTF;

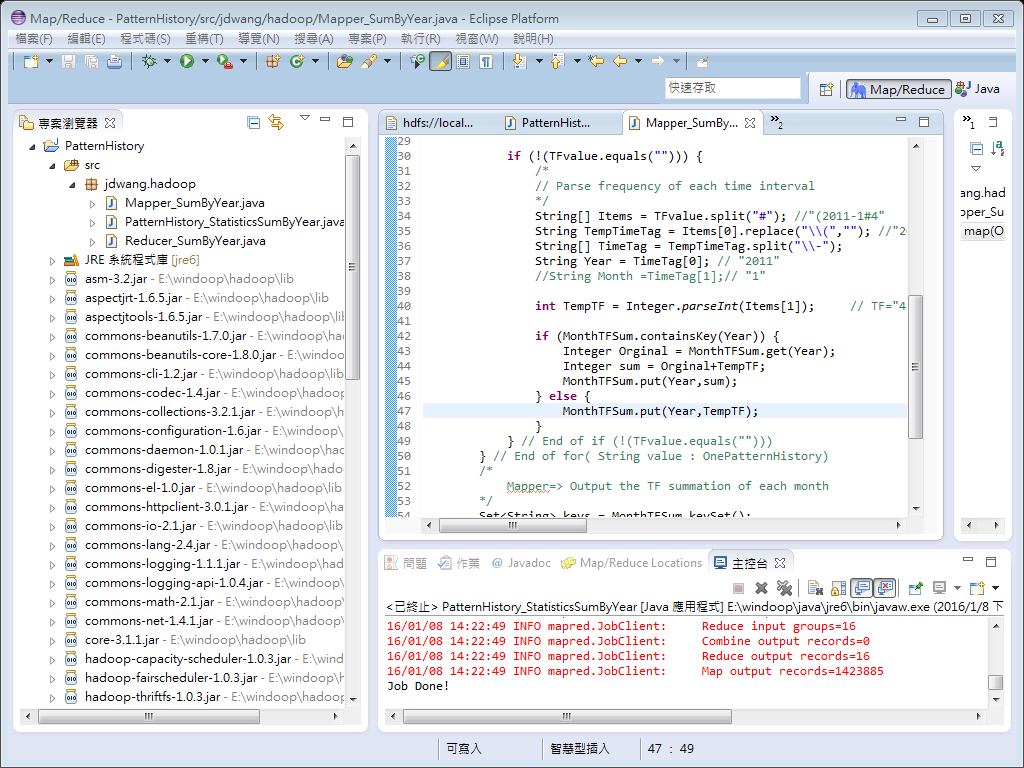
MonthTFSum.put(Year,sum);

} **else** {

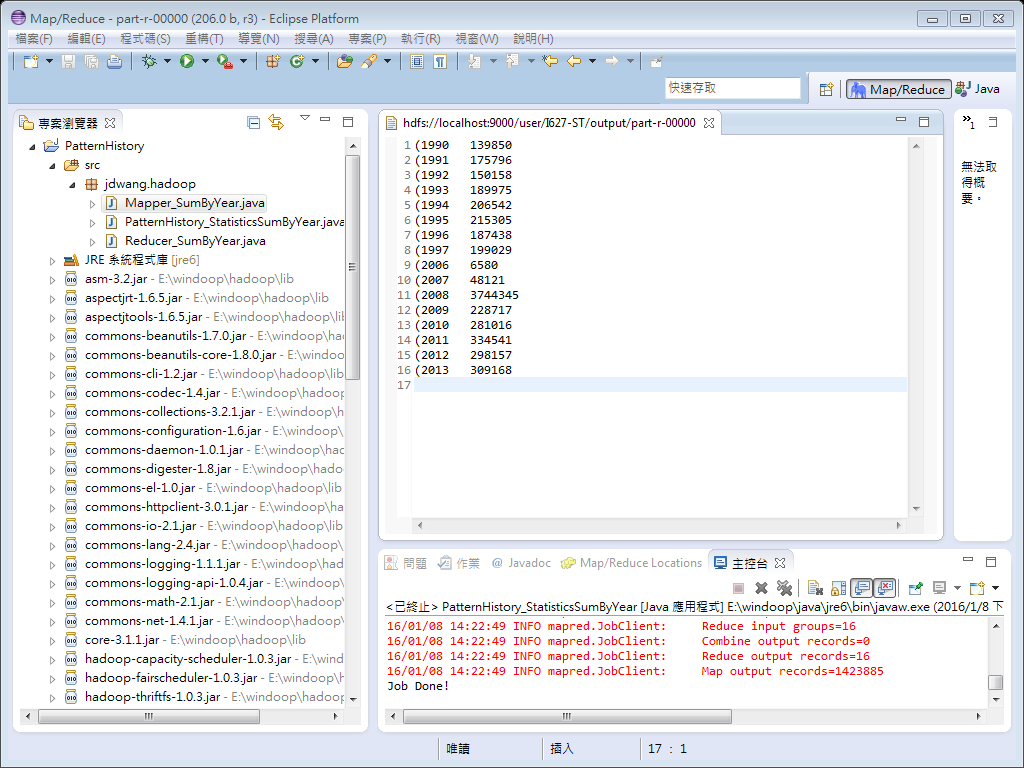
MonthTFSum.put(Year,TempTF);

}

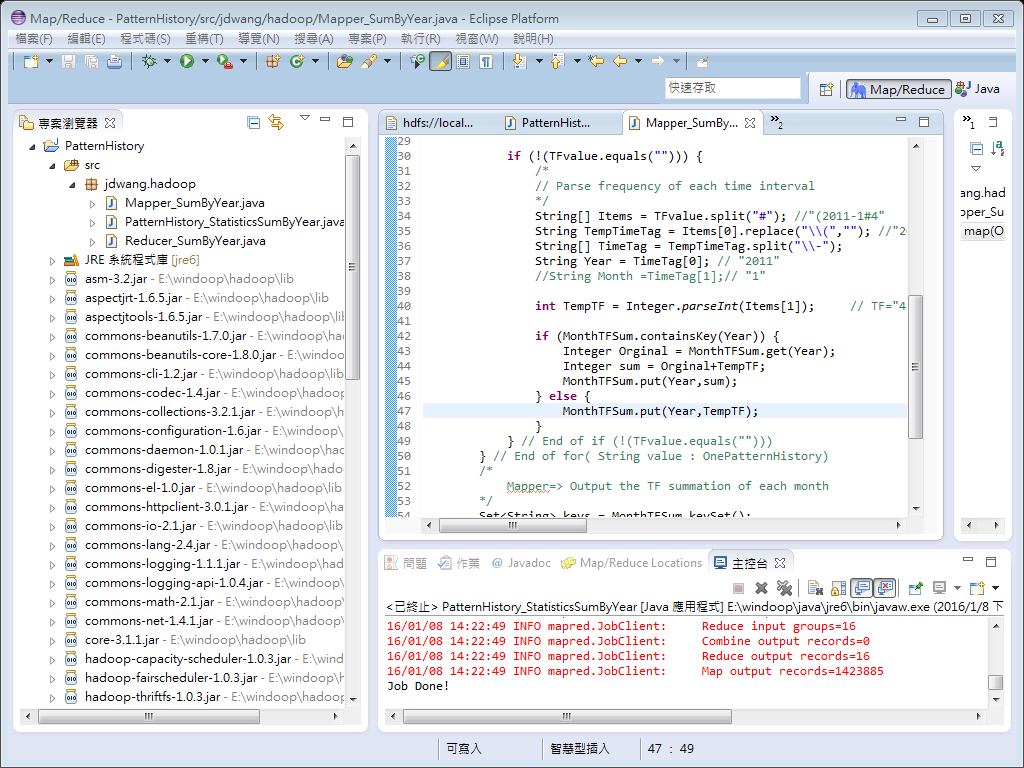
資料：output\_PubMed\_Paresed\_100\_Pass2

程式執行成功：

程式執行結果：



1. [PatternHistory\_SumByMonth.7z](http://dns2.asia.edu.tw/~jdwang/TeachingCourses/CloudDataManagement/PatternHistory_SumByMonth.7z)改寫成PatternHistory\_SumByYear.7z



將Mapper\_SumByMonth.java

改成Mapper\_SumByYear.java

將PatternHistory\_StatisticsSumByMonth.java

改成PatternHistory\_StatisticsSumByYear.java

將Reducer\_SumByMonth.java

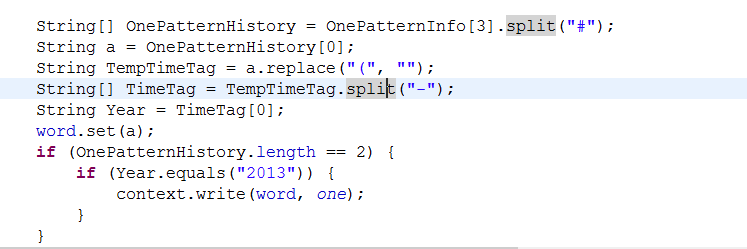
改成Reducer\_SumByYear.java

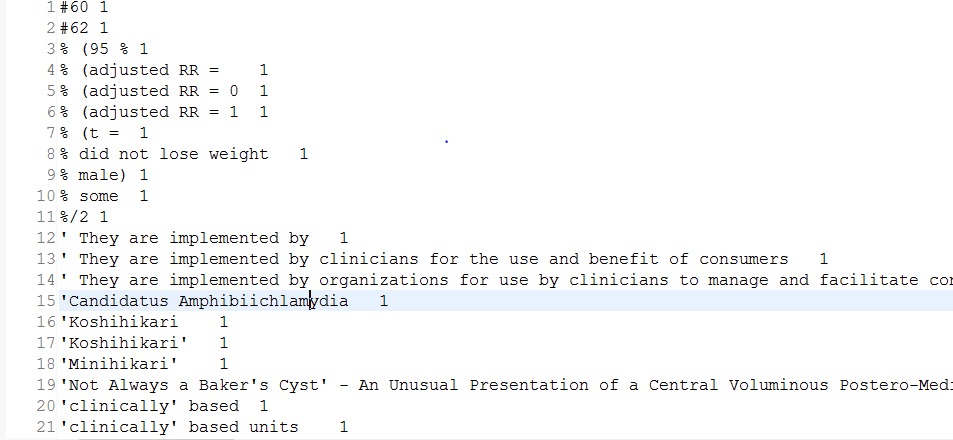
1. 找出只有出現在2013年的pattern

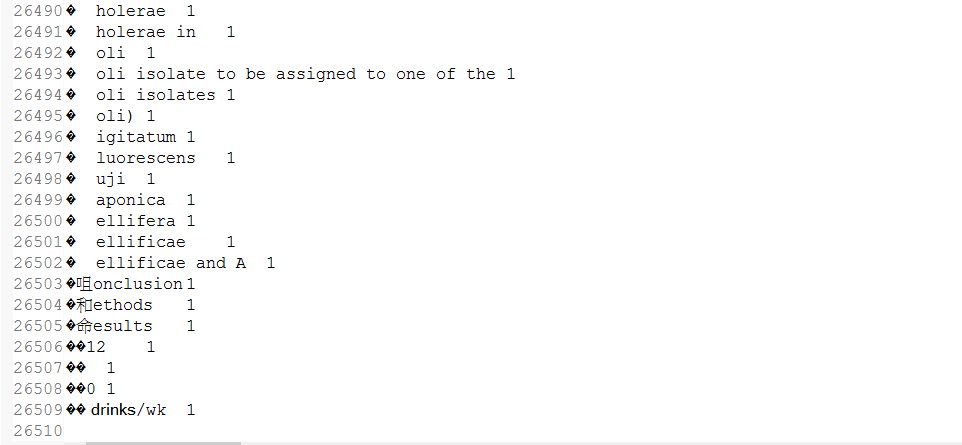
檔案：PatternHistory

程式修改：Mapper\_SumByYear

資料：output\_PubMed\_Paresed\_800\_Pass2



程式執行成功：



最後一題感謝沈東政同學幫忙及協助！下面是與他協助的程式碼完成的，最可惜的是無法自己完成！

1. 討論與問題

問題: 看不懂老師的題目居多，不知道老師的問題要的是甚麼！

還有程式再做更改的時候，需要去更改。

希望老師多說明清楚要的部分是什麼?

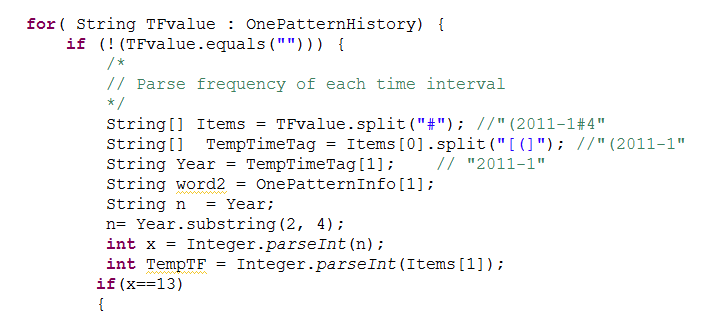
好比在第一題的計算標題拆單字的時候，**標題算出2個A和a有大寫和小寫。這也算是同個單字吧！**

但是拆解的時候卻把它列為不同的字。

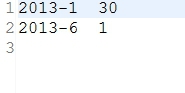
計算year的程式，剛開始讓windoop全部執行去跑，卻還是有括弧出現。應該是沒有replace進去。

後來私下又重新修改了一下，變成跑出只跑出年份和月份的。**用[( ]去分割。**

如果用 [ - ]去分割，會將前面的[ ( ]也讀進去！



這樣的結果是可以成功跑出年份和月份！(這是下面擷取其中一段)

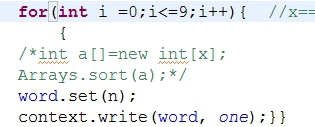
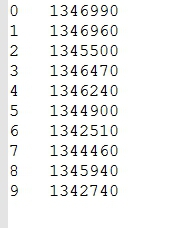


然後最後一題，老師說要算出2013pattern，沒有說明清楚是要2013的單字還是年分！

如果是要單字的話，上面的2013pattern只是部分。但是還是沒有辦法完整地將單字辨識出來！

老師說要的是只出現含有「2013年的月份」的話，那應該就可以完成。

計算PubMedID的部分，我試著用for迴圈讓讓程式跑出0~9的個數的PubMedID算出來，是成功了一半！數據10倍卻多了。

在最後一題卡關了！特別請教算出來的同學，感謝他的協助！終於解出來了！

這次是和資應系的同學們一起參與及討論修改完成的！

這些問題都解出來，覺得有進步了！只差在最後一題。

1. 小組成員工作分配

洪子涵：

程式修改、資料統整、機上操作、容錯發現

黃郁嵐：

資料紀錄、機上操作、容錯發現

1. 個別貢獻度百分比

|  |  |  |
| --- | --- | --- |
|  | 洪子涵 | 黃郁嵐 |
| 程式修改 | 100% | 25% |
| 機上操作 | 50% | 50% |
| 容錯發現 | 100% | 25% |
| 資料紀錄 | 30% | 70% |
| 資料統整 | 45% | 65% |

1. 個人心得

102063107黃郁嵐個人心得：

雖然說程式碼不是我修改的，但這個期末專題是我這學期最麻煩的報告。我不懂java要怎麼寫，所以這次專題經過老師講解後，大概知道它們是在做什麼，但是沒辦法去修改程式，在此非常感謝隊友carry我～

要修改別人的程式碼真的很難，尤其是要修改明明能用別的比較簡單卻偏要用一堆重複複雜的語法。想要重新寫整個程式又很麻煩＆時間不夠。好吧，以上都是抱怨，接下來我要去學java了。

102063099 洪子涵:

老師給的這項作業，剛開始看真的不知道該怎麼下手。

曾想過用spark去算，但也只能算出那些年出現的次數最多。

用spark很快就算好了沒錯，但是它的語言是python和scala這兩種都不熟，所以再用spark去改程式的時候，覺得有些吃力。

後來還是用windoop去改寫程式，在windoop執行、更改。

後來也有和朋友討論及修改，最後終於完成了！

這次沒有辦法全部解出老師要的，也解出自己想要的結果！

謝謝老師的帶領及同學的協助。