Gebze Technical University
Computer Engineering

CSE 222 - 2018 Spring

HOMEWORK 5 REPORT

ALI HAYDAR KURBAN
151044058

Course Assistant: Ayse Serbetci Turan

1 INTRODUCTION

1.1 Problem Definition

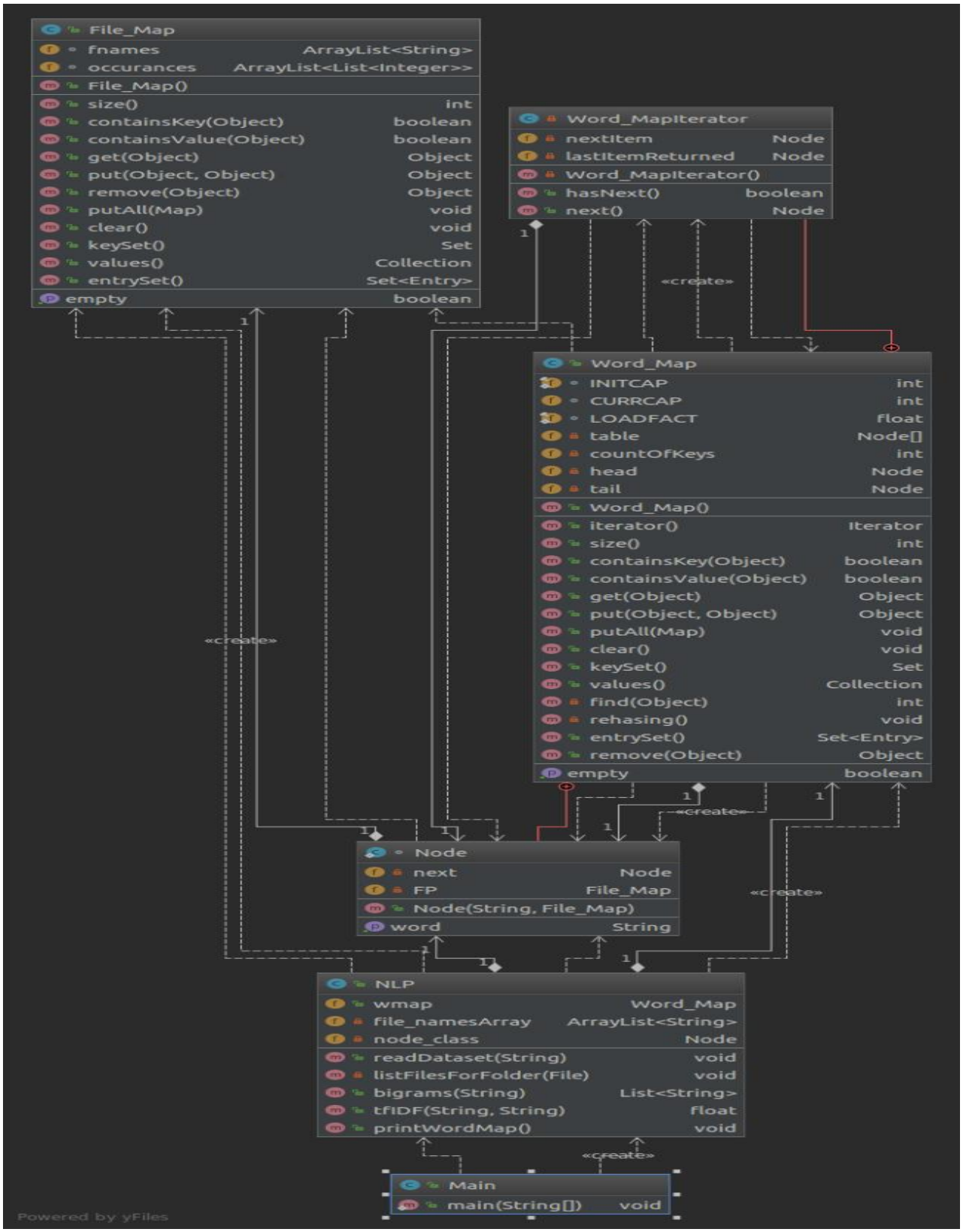
The problem is find bigrams a given word and a TFIDF of a given Word. There is a folder and inside of it there ara a lot of file. First of all read all files and put in a HashMap then find the bigrams of Word and TFID. In this homework I understood how a HashMap implementing, rehashing, how to use a HashMap, reading a folder with insede with its file. I understood the complexity of HashMap.

1.2 System Requirements

The solution does not require a hardware, or certain minimal amount of memory. To solve this problem I used some library of java, such as ArrayList, Set, Scanner… It can work a machine which has JVM.

2 METHOD

2.1 Class Diagrams



2.2 Use Case Diagrams

Users have to know that there is 2 operations once are bigrams of a word, and the other one is TFIDF of a given word and its file. Users have give a file to run these operations. The input file must be like the:
bigram word1
bigram word2
tfidf word3 filename

there is a whitspace of word and bigram takes only word, tfidf takes word and its file name. Users only have to prepare a file format of text obey the this rules.

2.3 Problem Solution Approach

To solve this problem I implement a hashmap and a map. The value of hashmap(Word_Map) is word which are in the files, the value of Word_Map is a map.The map which is value of Word_Map is File_Map. The File_Map maps' key is file name and value is an List of array. The list hols integer the integer means that the given word is inside the file's index. To implement this maps I override Maps' methods. In extra I implement 2 methods in Word_Map. These are find and rehasing. Find find an index of given element and rehasing is incremanting the Word_Map map and rehascode all items and put in new Word_Map. The real methods are bigrams and tfIDF. Lets explain them. First of all finding the tfIDF I used an formula given the homewrok pdf. Fist I found the value map of given word then calculate the number of terms appears in the given file. Then I found all given file with their size.(For example file1 is given file, I looked all value of file1 then their size and added them). Then I found all File_Map size, finally I found number of document with given word. And I made calculation it gave to me tfidf. To find bigram first I get the File_Map then it location array. With an iterator of Word_Map I look same file name and get its location array. If its location is bigger(just 1) than my word location, it means they are side by side. I made this design because there are a lot of word an file and to make execution time smaller I used HashMap.

Comlexity of all method of Word_Map is O(1), in worst case O(n). (n means that size of Word_Madp).

Compexit of File_Map size method is O(1), isEmpty is O(1), containsKey is O(n), containsValue is O(n), get mthod is O(n), put method O(1) in worst case O(n), remove method is O(1), putAll method is O(n), clear method is O(1), keySet method is O(n), values method is O(n) and finally entrySet is O(n). n means that size of fnames' size for all these methods.

3 RESULT

3.1 Test Cases

I tested my program with given example in homework pdf.
The test file1 is :
bigram very
tfidf coffee 0001978
bigram world
bigram costs
bigram is
tfidf Brazil 0000178

The test file2 is :
bigram difficult
bigram avoid
tfidf disclosed 0001184

3.2 Running Results

The output of file1 is:

[very difficult, very soon, very rapid, very aggressive, very promising, very attractive, very vulnerable]

0.004878173

[world market, world coffee, world made, world share, world markets, world price, world bank, world as, world cocoa, world prices, world for, world tin, world grain]

[costs have, costs and, costs of, costs Transport]

[is the, is not, is possible, is forecast, is caused, is expected, is depending, is at, is slightly, is projected, is estimated, is to, is due, is a, is well, is that, is no, is still, is imperative, is heading, is an, is difficult, is sold, is keeping, is defining, is time, is too, is uncertain, is proposing, is willing, is some, is unlikely, is fairly, is 112, is high, is going, is likely, is in, is basically, is also, is faced, is insisting, is unfair, is are, is only, is sending, is planned, is affecting, is trying, is harvested, is trimming, is improving, is Muda, is set, is meeting, is foreseeable, is beginning, is great, is precisely, is now, is one, is he, is after, is aimed, is committed, is insufficient, is put, is currently, is wrong, is unrealistic, is often, is it, is being, is searching, is showing, is helping, is why, is apparent, is open, is scheduled, is concerned, is more, is keen, is how, is downward, is sceptical, is favourable, is unchanged, is passed, is very, is getting, is ending, is down, is flowering]

0.0073839487
(I can not SS for that becuse there is no toString method like returns string like that.)

The output of file2 is :