Gebze Technical University
Computer Engineering

CSE 222 - 2018 Spring

HOMEWORK 6 REPORT

NURULLAH GÜNDOĞDU
161044108

AYŞE ŞERBETÇİ TURAN

1 INTRODUCTION

1.1 Problem Definition

In this project, we were asked to find the bigram and TFIDIF of the words we read from the file. In order to achieve these values, an unspecified number of files are provided in a folder. By reading each of the given files, we need to keep the list of words and words in each file in the index of the file. We need to keep these lists in different clas. We have 2 clas, which are the Word_map and File_map classes. We keep the words in this class and an object of the File_map class. In the file_map class, we keep the file names and the locations where the word passed. Then print the bigram or TFDIF results on the screen according to the commands given in the file.

Formul of TFIDIF;

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF(t) = log(Total number of documents / Number of documents with term t in it)

TFIDF = TF*IDF

Bi-grams explanation;

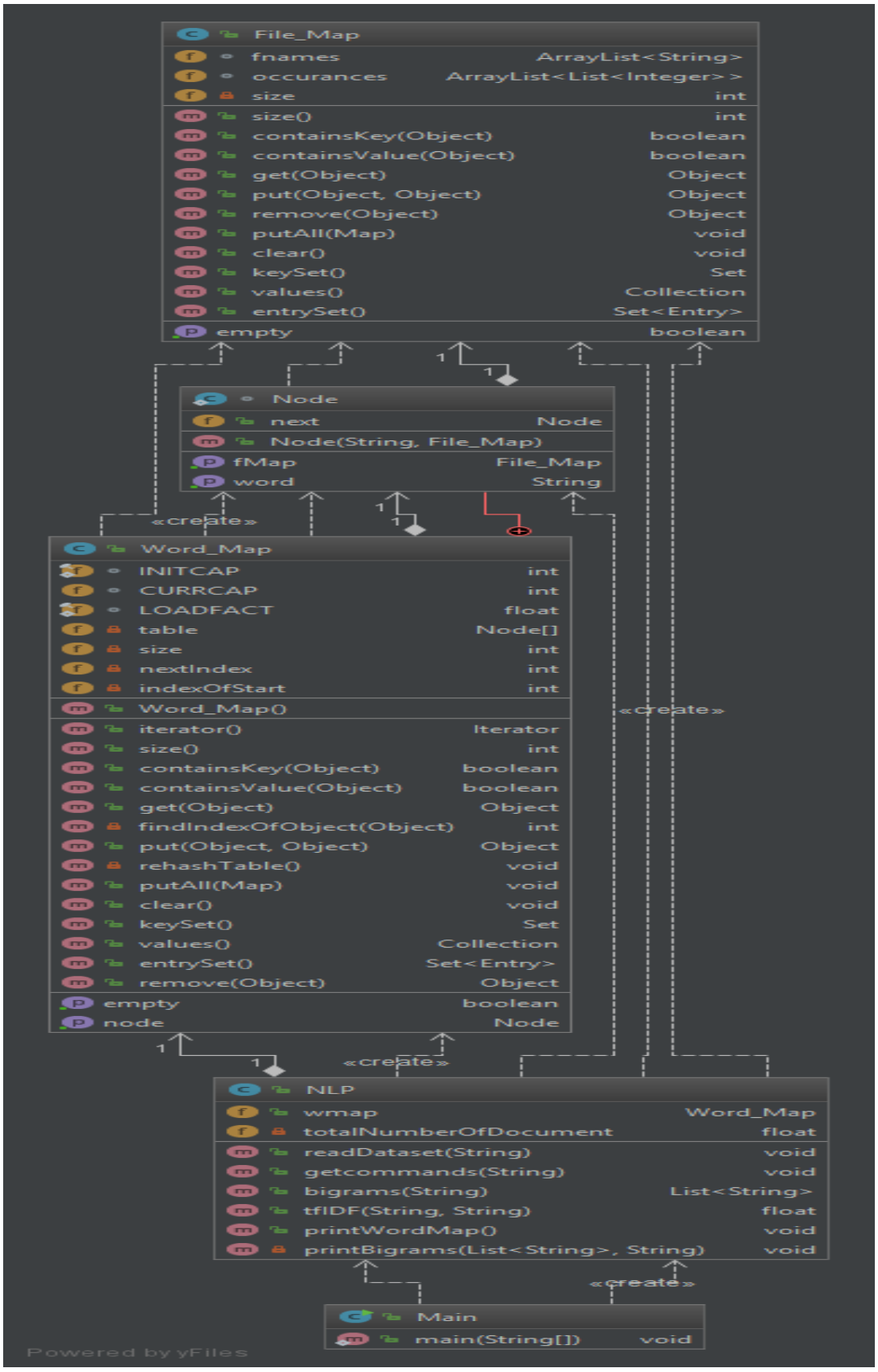
A bi-gram is simply a piece of text consisting of two sequential words which occurs in a given text at least
once. Bi-grams are very informative tools to reveal the semantic relations between words.

1.2 System Requirements

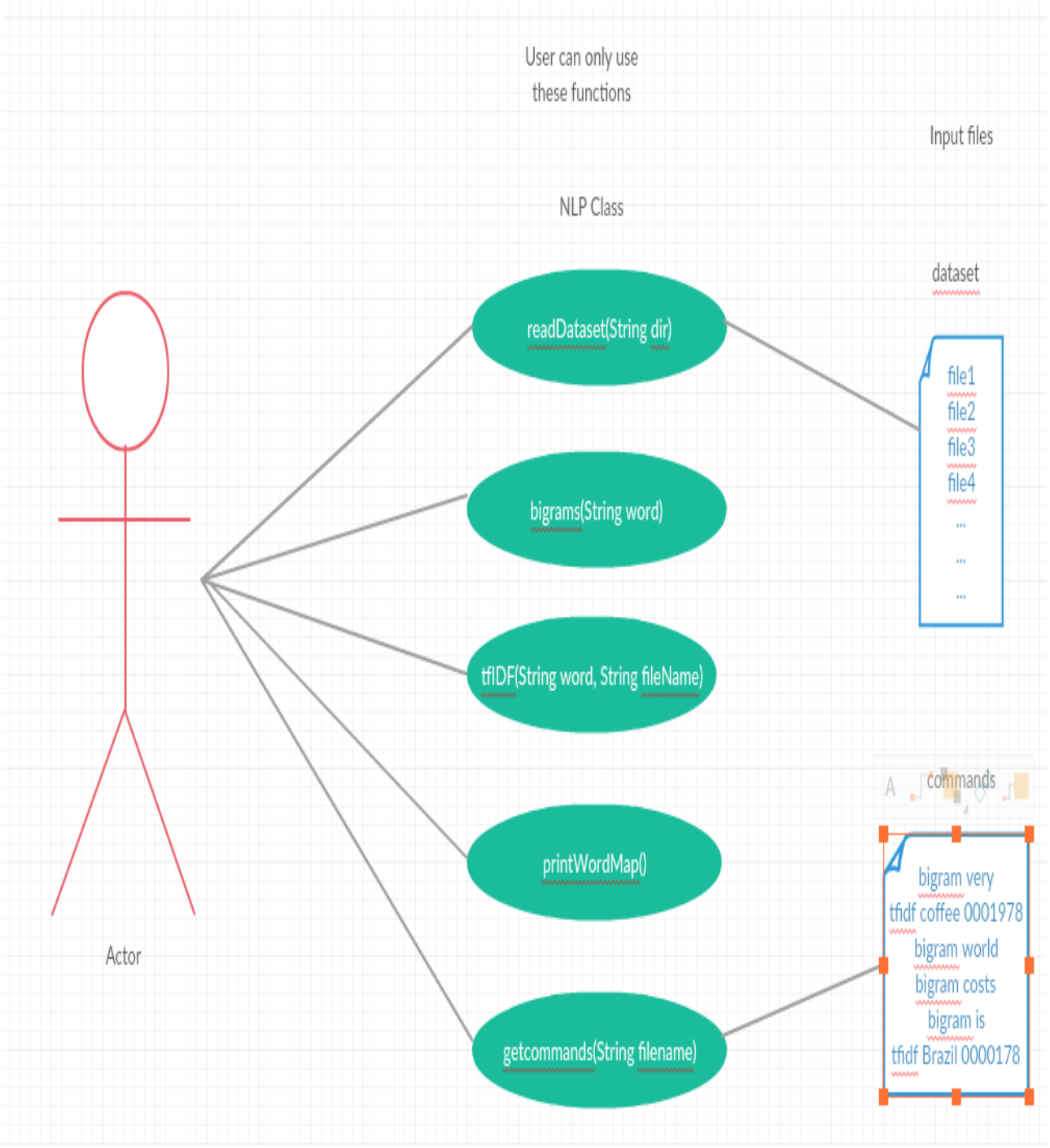
You have to download İntellij IDEA for run the programs. The program requirements are listed below.
Microsoft Windows 10/8/7/Vista/2003/XP (incl.64-bit)
2 GB RAM minimum, 4 GB RAM recommended
1.5 GB hard disk space + at least 1 GB for caches
1024x768 minimum screen resolution
You can download program here;
https://www.jetbrains.com/idea/download/#section=windows (Community Version)

2 METHOD

2.1 Class Diagrams



2.2 Use Case Diagrams



2.3 Problem Solution Approach

In this Project I create 3 classes. These are Word_Map, File_Map, NLF classes. Firstly I create NLF object in main and call readDataset function with filename. Program reads files in dataset directory and put every word with their index in Word_Map object. When every words puts theWord_Map, call the getcommands function. In this function program reads input files. There are command in this file. If command is bigram, program call bigram function and find all bigrams of given Word. If command is TFDIF, program call tfdif function and calculate TFDIF value of given word in given file. After that program finish.

NLP Class;

public void readDataset(String dir)

In that function read files name from directory. After that read all files string by string. Each word is checked in Word_Map. If the word is not in Word_Map, the word using the put () function and the created File_Map class object is added to Word_Map. If the word is reached, File_Map is reached and File_Mape is added to the location of the word. When creating File_Map, File_Map is added with the put () function.

Time complexity :
n = Total number of words in all files
T(n) = O(n).

void getcommands()

This function reads the file with the commands. The number of cycles in the loop returns. In each command he calls the command's function and performs the command.

Time complexity :
n = Number of commands in files
T(n) = O(n).

List<String> bigrams(String word)

This function returns a list of all the bigrams in the files read. This creates a list of the word in Word_Map by looking at File_Map. File_Map looks at the index of the files it contains and looks at the File_Map of other words in Word_Map and finds the next words from itself and adds the words to the list.

Time complexity :
n = Number of commands in files
T(n) = O(n).

float tfIDF(String word, String fileName)

This function finds and returns the TFIDF value. First there is the TF value. This value finds the number of words in the file from the same word, then finds the value found by dividing it by the number of words in the file. Then it finds the IDF value. In this value, the file number is divided by the number of files passing the word and finds the logarithm. Multiplies the Tf and IDF values and returns the result.

Time complexity :
n = Number of files
T(n) = O(n).

void printWordMap()

This function prints on the Word_Map to screen.

Time complexity :
n = Number of words
T(n^2) = O(n^2).

void printBigrams(List<String> list, String word)

This function prints the bigram of the word on the screen.

Time complexity :
n = Number of bigarm
T(n) = O(n).

Word_Map Class;

Iterator iterator()

This function return iterator object.

Time complexity :
T(1) = O(1).

int size()

This function return size of Word_Map.

Time complexity :
T(1) = O(1).

boolean isEmpty()

This function return true if Word_Map is empty. If is not it returns false.

Time complexity :
T(1) = O(1).

boolean containsKey(Object key)

This function return true if key is in Word_Map. If is not it returns false. It calculates the hashcode of key and check the index and check if it is null.

Time complexity :
T(1) = O(1).

boolean containsValue(Object value)

This function return true if value is in Word_Map. If is not it returns false. It looks all File_Map in the Word_Map and checks whether they are equal to File_Map.

Time complexity :
n = Number of FileMaps in Word_Map
T(n) = O(n).

Object get(Object key)

It returns the File_Map of given key. It calculates the hashcode of key and check the index and check if it is null.

Time complexity :
T(1) = O(1).

int findIndexOfObject(Object key)

This function find the index of key with its hashcode. And returns index.

.

Time complexity :
If it can find with hashcode complexity;
T(1) = O(1).

if it can’t complexity

T(capacity * ¾) = O(capacity * ¾)

Object put(Object key, Object value)

This function adds key and value to Word_Map. It calls findIndexOfObject() function and find the index of key after that it adds to table index.

Time complexity :
T(1) = O(1).

Object putAll(Map m)

This function add all words and File_Map to Word_Map with put() function.

Time complexity :
n = Number of words in Map
T(n) = O(n).

void rehashTable()

This function resize the table array if it is nearly full.

Time complexity :
n = Number of words in Word_Map
T(n) = O(n).

void clear()

This function clear the all data in the Word_Map

Time complexity :
n = Number of words in Word_Map
T(n) = O(n).

Set keySet()

This function puts all words in the Word_Map to a set and return set.

Time complexity :
n = Number of words in Word_Map
T(n) = O(n).

Collection values()

This function puts all File_Map in the Word_Map to a Collection and return Collection.

Time complexity :
n = Number of words in Word_Map
T(n) = O(n).

Set<Entry> entrySet()

This function adds the all data in Word_Map to Set<Entry> and returns it.

Time complexity :
n = Number of words in Word_Map
T(n) = O(n).

Object remove(Object key)

This function delete given object from WordMap.

Time complexity :
T(1) = O(1).

File_Map Class;

int size()

This function return size of File_Map.

Time complexity :
T(1) = O(1).

boolean isEmpty()

This function return true if File_Map is empty. If is not it returns false.

Time complexity :
T(1) = O(1).

boolean containsKey(Object key)

This function return true if key is in File_Map. If is not it returns false. It looks every index of list for find object.

Time complexity :
n = Number of file in FileMap
T(n) = O(n).

boolean containsValue(Object value)

This function return true if value is in File_Map. If is not it returns false. It looks every index of list for find object.

Time complexity :
n = Number of file in FileMap
T(n) = O(n).

Object get(Object key)

This function return list of index in ArrrayList. It looks every index of list for find the list.

Time complexity :
n = Number of file in FileMap
T(n) = O(n).

Object put(Object key, Object value)

This function adds key and value to fnames and occurances list.

Time complexity :
T(1) = O(1).

Object putAll(Map m)

This function adds all key and value in map to fnames and occurances list.

Time complexity :
n = Number of files in Map
T(n) = O(n).

void clear()

This function clear all data in File_Map .

Time complexity :
T(1) = O(1).

Set keySet()

This function puts all file in the File_Map to a set and return set.

Time complexity :
n = Number of files in File_Map

T(n) = O(n).

Collection values()

This function puts all list of index in the File_Map to a Collection and return Collection.

Time complexity :
n = Number of files in File_Map
T(n) = O(n).

Set<Entry> entrySet()

This function adds the all data in File_Map to Set<Entry> and returns it.

Time complexity :
n = Number of files in File_Map
T(n) = O(n).

Object remove(Object key)

This function delete given object from File_Map.

Time complexity :

T(1) = O(1).

3 RESULT

3.1 Test Cases

```
bigram very
tfidf coffee C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0001978
bigram world
bigram costs
bigram is
tfidf Brazil C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000178
bigram a
tfidf International C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000026
bigram International
```

3.2 Running Results

3.2.1 Input.txt Result

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

0,004878

[world market, world coffee, world made, world share, world price, world markets, 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 expected, is caused, is depending, is at, is
 estimated, is slightly, is projected, is to, is due, is a, is still, is no, is that, is well, is
 heading, is imperative, is an, is difficult, is time, is too, is keeping, is defining, is sold, is
 uncertain, is some, is fairly, is unlikely, is willing, is proposing, is 112, is high, is likely,
 is going, is in, is also, is faced, is basically, is are, is insisting, is unfair, is only, is
 sending, is planned, is affecting, is trying, is harvested, is trimming, is improving, is Muda, is
 meeting, is set, is foreseeable, is beginning, is great, is precisely, is now, is one, is he, is
 after, is aimed, is committed, is put, is insufficient, is currently, is wrong, is unrealistic, is
 it, is often, is being, is searching, is showing, is helping, is why, is apparent, is scheduled,
 is open, is concerned, is more, is keen, is how, is downward, is sceptical, is favourable, is
 unchanged, is passed, is very, is ending, is getting, is down, is flowering]

0,007384

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

0,006442

[International Coffee, International Merchant, International Services, International Petroleum,
 International Monetary, International Cocoa, International agreements, International Natural,
 International Tin]


Process finished with exit code 0
```

3.2.2 printWordMap() Function Result

```
Word    :250000     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007513       Index of word   :[13]

Word    :195000     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007513       Index of word   :[24]

Word    :14     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007513       Index of word   :[42]
Word    :14     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008364       Index of word   :[457]
Word    :14     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0009271       Index of word   :[6]
Word    :14     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0009280       Index of word   :[46]

Word    :3845000        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007513       Index of word   :[59]

Word    :4070000        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007513       Index of word   :[63]

Word    :saw        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[4]

Word    :unsatisfactory     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[21]

Word    :Especially     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[25]

Word    :East       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[29]

Word    :grade      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[42]
Word    :grade      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008859       Index of word   :[44]

Word    :robustas       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[43]

Word    :met        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[45]
Word    :met        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008477       Index of word   :[537]
Word    :met        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008606       Index of word   :[40]
Word    :met        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008777       Index of word   :[82]
Word    :met        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008792       Index of word   :[82]

Word    :Sporadic       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[48]

Word    :origin     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[68]

Word    :covered        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[76]

Word    :enter      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[82]

Word    :purchases      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[87]
```

```
Word    :Bremen     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[1]

Word    :attracted      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[5]

Word    :neglected      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[18]

Word    :Buyers     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[22]

Word    :awaiting       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[24]
Word    :awaiting       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007709       Index of word   :[292]
Word    :awaiting       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008364       Index of word   :[188]
Word    :awaiting       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008425       Index of word   :[164]

Word    :qualities      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[40]

Word    :FNC        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[54]

Word    :shippers       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[57]

Word    :attractive     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[65]

Word    :Americans      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[73]

Word    :afloat     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[79, 116]

Word    :nearby     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[84]
Word    :nearby     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007964       Index of word   :[42]

Word    :scarce     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[88]

Word    :turnover       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[90]

Word    :offers     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[97]

Word    :hand       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[111]
Word    :hand       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007569       Index of word   :[55]

Word    :prompt     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[118]
Word    :prompt     File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007110       Index of word   :[140]

Word    :fob        File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005743       Index of word   :[125]
```

```
Word    :hurt       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000527        Index of word   :[438]

Word    :ICOs       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000527        Index of word   :[443]
Word    :ICOs       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008350        Index of word   :[118]
Word    :ICOs       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008477        Index of word   :[76, 164]
Word    :ICOs       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008606        Index of word   :[63]

Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000527        Index of word   :[448]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000677        Index of word   :[28]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0001327        Index of word   :[134]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0001603        Index of word   :[31, 456, 855]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0002972        Index of word   :[37]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0004598        Index of word   :[3]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0006272        Index of word   :[141]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007678        Index of word   :[287]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007882        Index of word   :[311]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008656        Index of word   :[27, 319, 463]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008777        Index of word   :[230]
Word    :analysts       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008792        Index of word   :[229, 459]

Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000527        Index of word   :[461]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000641        Index of word   :[54]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000828        Index of word   :[607]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0001603        Index of word   :[759]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0001631        Index of word   :[209]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0002732        Index of word   :[27]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0002820        Index of word   :[97]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0003195        Index of word   :[213]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0006796        Index of word   :[43]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007678        Index of word   :[285]
Word    :much       File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0008606        Index of word   :[60]

Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0000527        Index of word   :[463]
Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0001978        Index of word   :[172]
Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0003322        Index of word   :[82]
Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0005750        Index of word   :[296, 397]
Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0006429        Index of word   :[84]
Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007218        Index of word   :[249]
Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007709        Index of word   :[354]
Word    :three      File    :C:\Users\nurullah\IdeaProjects\2019_HW6\dataset\0007964        Index of word   :[105]
```