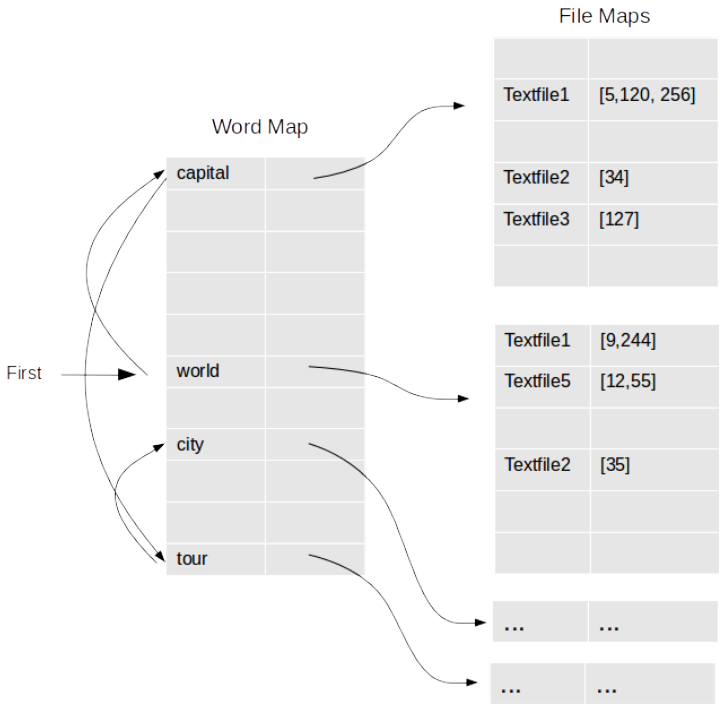
GEBZE TECHNICAL UNIVERSITY
Department of Computer Engineering
CSE 222-Data Structures
2018-2019 Spring
HW #6
Due Date: 22.04. 2019 06:00

In this homework, you will implement two different HashMap classes to perform basic Natural Language Processing operations. Specifically, you will read a text dataset consisting of multiple input files and keep the words in the Word HashMap. The key for the word hashmap will be the words and the value will refer to another hashmap (file hashmap) which keeps the occurrences of the word in different files. The key for the file hashmap is the filename and the value is an arraylist containing the word positions in that file. In order to be able to traverse the word map in an efficient way, each entry should keep a pointer to the next inserted entry, allowing the structure to be Iterable (implements Iterable interface). By this way, the methods such as containsValue(), containsKey() and keySet() can be implemented in a more efficient way, without the need for visiting empty cells in the table. In Figure 1 a sample snapshot of the hashmaps is presented.



After obtaining this structure, you will implement two basic operations used in NLP : retrieving bi-grams and calculating TFIDF values, which are explained below, respectively.

Bi-grams: 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. Let us find the bi-grams in the following text:

“For several years Uganda had been unable to meet its ICO
export quota as rebel activity disrupted the coffee industry”

Bi-grams:

For several
several years
years Uganda
Uganda had
had been
...
the coffee
coffee industry

In Figure 1, the dataset contains a bi-gram “capital city” since capital occurs in the Textfile2 at position 34 and city occurs in the same file at position 35.

TFIDF(Term frequency-inverse document frequency) : This is a score which reflects the importance of a word for a single document. In NLP, a word is informative for a file to be categorized if it occurs frequently in that file but has very few occurrence in other documents in the dataset. To calculate TFIDF, we first calculate the term frequency:

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

Then the inverse document frequency is calculated to weigh down the words that are frequent also in other files while scale up the rare ones.

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

TFIDF = TF*IDF

Your project will read an input file which consists of multiple queries and print the result of the standard output. There are two main query types. One for retrieving the bi-grams of a word and the other one is for calculating the TFIDF value of a given word for a given filename. A sample input file is given below:

```
bigram very
tfidf coffee 0001978
bigram world
bigram costs
bigram is
tfidf Brazil 0000178
```

The output should be:

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

0.004878

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

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

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

0.0073839
```

Submit your homework with file name <stdID>.zip which includes your IntelliJ project and your report in pdf format. You can ask your questions via asturan@gtu.edu.tr or moodle discussion forum.