CSC 317: Project 0

Bag of Words

Total: 50 points

In this project you will implement a bag of words (BoW). A BoW collects all words from input text files as a multi-set (meaning repeated words are maintained with multiplicity), and enables the calculation of certain statistics that help identify the importance of the words to a corpus (BoW). In this project you may represent it as a *set* (no repetition allowed) rather than a multi-set, while maintaining occurrence counts (frequencies). You should write a BoW class that is capable of doing the following:

1. **(Required)** Constructor: BoW(String text_file_name). This will create a BoW object initializing it with the words from the input text file.
2. **(Required)** Public Method 1: expand(String another_text_file_name). This will absorb into the BoW all words from the new text file.
3. **(Required)** Public Method 2: printTermFrequency(). This will print a list of all distinct words currently in the object's set, and their frequencies (number of occurrences).
4. **(Bonus)** Public Method 3: printInverseDocumentFrequency(). This will print a list of all distinct words currently in the object's set, and for each word, will print the ratio of the total number of documents (absorbed into the BoW so far) to the number of documents in which that word appears.

Feel free to write other (private) methods as needed. To keep this warm-up project simple, try not to create more than the one BoW class. However, if you do use multiple classes, then you must include a UML class diagram (which will be covered in class at a later date) to show their relationships. Your program should be interactive as shown below.

Sample Run

The text in bold preceded by '>' indicates sample user input. The rest is sample output from the program.

```
Please input a file name to initialize the BoW:
```

> **file1.txt**

```
Please select an operation from below:
        1. Expand
        2. Print term frequencies
```

```
        3. Print inverse document frequencies
        4. Exit
```

> **1**

```
Please input a file name to expand the BoW:
```

> **file2.txt**

```
Done.
Please select an operation from below:
        1. Expand
        2. Print term frequencies
        3. Print inverse document frequencies
        4. Exit
```

> **2**

```
Here are the term-frequencies so far:
```

…

What to submit

Submit the following on Canvas:

1. (**Optional**) A .pdf file showing the UML class diagram for the project, only if multiple classes are used.
2. (**Required**) A compiled jar file of the project. I should be able to run the .jar from command line like so: `java -jar P0.jar`
3. (**Required)** A .java file of the source code. If you have multiple .java files, zip them up and submit a .zip file instead. Don't include .java files in the jar.