

# BIM207 2021-2022

## Homework Due Date: 30/11/2021

- Your program takes the following options and arguments using `args4j` plugin to pass to the program when required:
  - Files as argument i.e. `java -jar target\WordCounter.jar sample1.txt sample2.txt`
  - `-task` : task name to run and should be one of following `"NumOfTokens"` `"FrequentTerms"` `"TermLengthStats"` `"TermsStartWith"`
  - `-r` : refers to reverse order
  - `-u`: refers to unique values
  - `-topN`: Number of items to be printed
  - `-start`: Terms starts with given String.
- You should do at least **3 commits with a commit message** to your repository. (10 pts)
- First of all, you should apply to preprocess to the text in the files.
  - Read all lines from the files.
  - Tokenize the lines and convert the tokens into lowercase (Hint: You can use `StringTokenizer` and `toLowerCase()` )
  - For each token, remove the characters that are not Letter or Digit. To do this, you can use `Character.isLetterOrDigit()`
  - At the end of those operations, you will get the terms to be used in the tasks.
- **Task: NumOfTokens (20 pts)**
  - You should print the number of terms. Example outputs:  
`java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task NumOfTokens`  
Number of Tokens: 761  
  
`java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task NumOfTokens -u`  
Number of Tokens: 309
- **Task: FrequentTerms (30 pts)**
  - You should print frequent terms with their occurrences. The order and the number of items to be printed may change with `-r` and `-topN` options. If multiple terms have an equal number of occurrences, then they should be sorted. Example outputs:  
`java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task FrequentTerms`  
the 39  
a 28  
and 26  
to 24  
is 14

```
java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task FrequentTerms -topN 10
```

```
the    39
a      28
and    26
to     24
is     14
of     13
this   12
an     11
by     11
are    9
```

```
java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task FrequentTerms -topN 10
```

```
-r
```

```
ad      1
adapter 1
administration 1
advantage 1
advent 1
allows 1
animation 1
another 1
application 1
away 1
```

- **Task: TermLengthStats (20 pts)**

- you should print the length of the longest term, the shortest term and the average length of the terms. When -u option is passed, then the program computes the statistics over unique terms.

```
java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task TermLengthStats
```

```
Max Token Length in Character: 16, Min Token Length: 1, Average Token Length: 5.2772
```

```
java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task TermLengthStats -u
```

```
Max Token Length in Character: 16, Min Token Length: 1, Average Token Length: 6.4757
```

- **Task: TermsStartWith (20 pts)**

- You should print the term that starts with the given prefix by -start option. The order and the number of items to be printed may change with -r and -topN options.

```
java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task TermsStartWith -start de
```

```
clarations
defines
defining
derive
design
```

```
java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task TermsStartWith -start de  
-r
```

devoted  
devote  
designing  
designers  
design

```
java -jar target\WordCounter.jar sample1.txt sample2.txt sample3.txt -task TermsStartWith -start de  
-r -topN 10
```

devoted  
devote  
designing  
designers  
design  
derive  
defining  
defines  
declarations

## Pom.xml configuration

Please use the plugins `maven-shade-plugin` and `maven-compiler-plugin`

```
<plugin>  
  <groupId>org.apache.maven.plugins</groupId>  
  <artifactId>maven-compiler-plugin</artifactId>  
  <version>3.8.1</version>  
  <configuration>  
    <source>1.8</source>  
    <target>1.8</target>  
    <encoding>UTF-8</encoding>  
  </configuration>  
</plugin>
```

## Important !

**Projects that fail to pass the below scripts will not be graded.**

**Such projects will take 0 points**

**Make sure the following commands are running**

```
git clone https://github.com/AnadoluUniversityCeng/bim207-hw1-2021-USERNAME.git  
cd bim207-hw1-2021-USERNAME  
mvn clean package
```