What to Submit (Updated 10/02)

2 JAR files (1 for each question), 5 output files (1 for each sub-problems: 3 for Q1, 2 for Q2).

Instructions on using the same jar file to run different subproblems:

Take three arguments with the following code:

```
if (otherArgs.length != 3) {
    System.err.println("Usage: Q1Analysis <in> <out> <part>");
    System.err.println("part: A (WordCount), B (TargetWords), or C
(Pattern)");
    System.exit(2);
}

String part = otherArgs[2];
```

Use a simple if-else statement to check the `part` variable and execute the job for the appropriate sub-program.

NOTE: It is recommended to keep all subproblems within a single Java file and a single JAR file (1 for each problem, total 2 java files, 2 jar files). This practice promotes DRY (Don't Repeat Yourself) code and simplifies evaluation. While you may submit separate Java files if necessary, be aware that marks will be deducted.

Following this method, the code execution command in Hadoop might look as follows where the third argument decides the subproblem:

Q1

docker exec -it resourcemanager hadoop jar /tmp/Q1Analysis.jar /inputA /q1_output_A A
docker exec -it resourcemanager hadoop jar /tmp/Q1Analysis.jar /inputA /q1_output_B B
docker exec -it resourcemanager hadoop jar /tmp/Q1Analysis.jar /inputA /q1_output_C C

Q2

docker exec -it resourcemanager hadoop jar /tmp/Q2Analysis.jar /inputB /q2_output_A A
docker exec -it resourcemanager hadoop jar /tmp/Q2Analysis.jar /q2_output_A /q2_output_B B

Output Format (Five Files):

Q1A

a 4697
aback 2
abaft 2
abandon 3
.....
.....

Q1B

sea 1234
ship 123
whale 34

Q1C

.....
.....
10,a 104
10,b 91
10,c 180
10,d 101
.....
.....

Q2A

.....
.....
abbeville 1769, 2759, 6617, 16360, 31497, 36928
abbey 33949, 48196
abbie 21810, 39655
.....
.....

Q2B

abc 9999

FAQ

1. The requirement to put everything into only two output files has been waived, as this was more complex than intended.
2. If you have already completed the assignment with the output in two files (one for each problem), you do not need to redo it.
3. For Q2A, you only need to consider the words; you do not need to include dates or numbers. A simple regex filter can achieve this in one or two lines. However, marks will not

be deducted if you include them.

4. For Q2A, the number at the beginning of each line should be treated as the line number.
5. For Q2B, use the output file from Q2A as the input.
6. For both Q1 and Q2, you may convert the text to lowercase. Marks will not be deducted if you choose not to.