COMP9313 - 2019T2
Assignment #1 (MapReduce) - 25 points total

Problem Statement

Given a set of text documents (i.e. text files) in input, create an index with the list of *ngrams* (e.g. bigrams) contained in these documents along with the number of times the ngrams were found across all documents and the list of files where the ngrams appear.

Input

The list of files is provided in a directory (you can have an arbitrary number of files and file names), for example:

```
/tmp/input/
      file01.txt
      file02.txt
      file03.txt
      ...
```

The directory must be in your local file system (we will not use HDFS for this assignment). We provide a set of sample files here:
https://webcms3.cse.unsw.edu.au/COMP9313/19T2/resources/28190

Output

The output consists in a file that contains the list of *ngrams* (e.g. bigrams) identified in the documents in input, along with the number of times the *ngram* was found across all documents and the list of files where the ngrams where found. For example:

```
a collection  1  file01.txt
a network     1  file01.txt
a part        1  file03.txt
hadoop is     2  file01.txt file03.txt
...
```

In the example above, the *bigram* `hadoop is` was found 2 times in total, and it was found in files `file01.txt` and `file03.txt`.

Program Arguments
Your Java program *must* receive 4 arguments:

***args[0]*:**	The value *N* for the ngram. For example, if the user is interested only in bigrams, then *args[0]=2*.
***args[1]*:**	The minimum count for an ngram to be included in the output file. For example, if the user is interested only in ngrams that appear at least 10 times across the whole set of documents, then *args[1]=10*.
***args[2]*:**	The directory containing the files in input. For example, *args[2]="/tmp/input/"*
***args[3]*:**	The directory where the output file will be stored. For example, *args[3]="/tmp/output/"*

Notice that the order of the arguments above is important. When we test your code, we will assume that the arguments are in the right order, as explained above.

Building your Program

Your program will be built using *Apache Maven*, and we assume that the *only* dependency to be used in your project is *hadoop-core, version 1.2.1* (do not add and/or rely on other dependencies in your

project). In the link below, we provide a Maven project template that you should use for developing your solution:

https://webcms3.cse.unsw.edu.au/COMP9313/19T2/resources/28189

Your code must be included (in its entirety) in the file Assignment1.java. If you use multiple Java classes in your solution, all these classes must be entirely defined within the file Assignment1.java. You must ensure that the code you submit can be compiled and packaged using Apache Maven (we recommend you use the project template provided above). Any solution that has compilation errors will receive no more than 5 points for the entire assignment.

Assignment Submission

Deadline: 05 July 2019 20:59:59

Log in to any CSE server (e.g. williams or wagner) and use the *give command* below to submit your solution:

```
$ give cs9313 assignment1 z9999999.zip
```

where you must replace z9999999 above with your own zID. The zip file above must contain the following:

- The file Assignment1.java containing your Java code
- A PDF document (maximum 1 page, 10 points font-size Arial) that explains your solution (use of figures is highly encouraged to explain your solution).

You can also submit your solution using WebCMS, or Give:
https://cgi.cse.unsw.edu.au/~give/Student/give.php

If you submit your assignment more than once, the last submission will replace the previous one. To prove successful submission, please take a screenshot and keep it for your own record. If you face any problem while submitting your code, please e-mail the Course Admin (Maisie Badami, m.badami@student.unsw.edu.au)

Late submission penalty

10% reduction of your marks for the 1st day, 30% reduction/day for the following days.

Assessment

Your source code will be manually inspected and marked based on readability and ease of understanding. We will run your code to verify that it produces correct results. The code documentation (i.e. comments in your source code) and solution explanation (PDF document) are also important. Below, we provide an indicative assessment scheme (maximum mark: 25 points):

Result correctness	15 points
Documentation (PDF document)	5 points
Code structure and source code documentation (comments)	5 points

Plagiarism

This is an *individual assignment*. The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this course. If you knowingly provide or show your assignment work to another person for any reason, and

work derived from it is submitted, you may be penalized, even if the work was submitted without your knowledge or consent.

Reminder: Plagiarism is defined as using the words or ideas of others and presenting them as your own. UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties as severe as being excluded from further study at UNSW. There are several on-line sources to help you understand what plagiarism is and how it is dealt with at UNSW:

- Plagiarism and Academic Integrity
- UNSW Plagiarism Procedure

Make sure that you read and understand these. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible for ensuring that your assignment files are not accessible by anyone but you by setting the correct permissions in your CSE directory and code repository, if using one (e.g., Github and similar). Note also that plagiarism includes paying or asking another person to do a piece of work for you and then submitting it as your own work.

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW staff and students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW.