

# COMP9313 21T3 Project 1 (10 marks)

## Problem statement:

Build the inverted index for a given set of documents (compute the term weights by TF-IDF as shown in slide 46 of Chapter 3.1, using base 10 logarithm). Ignore the letter case, i.e., convert all words to lower case first.

### Input files:

Each line is in format of “DocID DOC”, where DocID is the ID of the document, and DOC is the document content (a list of terms). For example (the DocIDs may not be continuous):

0 Construct Inverted Index
1 using MapReduce
2 inverted index code
3 index index file

This sample file “tiny-doc.txt” can be downloaded at:

<https://webcms3.cse.unsw.edu.au/COMP9313/21T3/resources/68063>

### Output:

If a term  $W$  is contained in  $N$  documents, the corresponding output for this term would have  $N$  lines, in format of: “ $W \text{ DocID}_1, \text{weight}_1$ ”, “ $W \text{ DocID}_2, \text{weight}_2$ ”, ..., and “ $W \text{ DocID}_N, \text{weight}_N$ ”. Please use **yield  $W, \text{DocID}_i$** , “**+str(weight<sub>i</sub>)**” in your reducer. For each term, the DocIDs are sorted in ascending order numerically. Given the example document, the output file is like (assumed in one file):

"code" "2, 0.6020599913279624"
"construct" "0, 0.6020599913279624"
"file" "3, 0.6020599913279624"
"index" "0, 0.12493873660829992"
"index" "2, 0.12493873660829992"
"index" "3, 0.24987747321659984"
"inverted" "0, 0.3010299956639812"
"inverted" "2, 0.3010299956639812"
"mapreduce" "1, 0.6020599913279624"
"using" "1, 0.6020599913279624"

### Code format:

Name your folder as “zID\_proj1” (e.g. z5123456\_proj1) and name your python mrjob script as “weightInvertedIndex.py”. Finally, package all your python files as a zip file with name “zID\_InvertedIndex.zip”.

## Compile:

To reduce the difficulty of the project, you are allowed to pass the number of documents to your job. We will also use more than 1 reducer to test your code. Assuming there are 10,000 documents and we use two reducers, we will use the following command to check the correctness of your solution:

```
$ python3 weightInvertedIndex.py -r hadoop hdfs_input -o hdfs_output --  
jobconf myjob.settings.docnumber=10000 --jobconf mapred.reduce.tasks=2
```

You can access the number of documents in your program by accessing “`jobconf_from_env('myjob.settings.docnumber')`”.

Please ensure that the code you submit can be compiled. Any solution that has compilation errors will receive no more than 4 points for the entire assignment.

## Documentation and code readability

Your source code will be inspected and marked based on readability and ease of understanding. The documentation (comments of the codes) in your source code is also important. Below is an indicative marking scheme:

Result correctness: 5
Efficiency and memory usage: 4 (the use of design patterns you have learned)
Code structure, Readability, and Documentation: 1

## Submission:

Deadline: Sunday 17th Oct 11:59:59 PM

You can submit through Moodle:

If you submit your assignment more than once, the last submission will replace the previous one. To prove successful submission, please take a screenshot as assignment submission instructions show and keep it by yourself. If you have any problems in submissions, please email to [yu.hao@unsw.edu.au](mailto:yu.hao@unsw.edu.au).

## Late submission penalty

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

## **Plagiarism:**

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 an offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Relevant scholarship authorities will be informed if students holding scholarships are involved in an incident of plagiarism or other misconduct.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this subject. 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.