

COMP336 - Big Data

Assignment 2

Semester 1, 2020

Macquarie University, Department of Computing

Due: Week 8 (Sunday 3 May)

Demonstration and Marking: Week 9, during the Workshop (Zoom)

Weighting: 20%

In this assignment you will implement MapReduce techniques for the processing of Big Data. You will build your assignment on top of Hadoop (i.e. an open-source version of MapReduce written in Java).

Note: If you cannot setup Hadoop on your local system, you can use online services such as:
<https://studio3t.com/>


This Assessment Task relates to the following Learning Outcomes:

Apply Map-reduce techniques to a number of problems that involve Big Data.

Task 1: (15%)

- Dataset: 10000 Tweets; dataset on iLearn “tweets.zip”
- MapReduce: Calculate the count of number of occurrences of each word in the text of Tweets.
- Create a short documentation in which you briefly describe your implementation:
 - What to write in the mapper(s) ? Flowchart and Pseudocode !
 - What to write in the reducer(s) ? Flowchart and Pseudocode !



```
...  
"link" : "http://twitter.com/ ",  
"text" : "THE FOLLOWING TAKES PLACE BETWEEN ...",  
"object" : {  
...  
}
```



Task 2: (15%)

- Dataset: 10000 Tweets; dataset on iLearn “tweets.zip”
- MapReduce: Calculate the count of number of tweets for a list of different cities in Australia.
- Create a short documentation in which you briefly describe your implementation:
 - What to write in the mapper(s) ? Flowchart and Pseudocode !
 - What to write in the reducer(s) ? Flowchart and Pseudocode !


```
...  
"twitterTimeZone" : "Sydney",  
"verified" : false,  
"utcOffset" : "39600",  
"preferredUsername" : "losebabyweight1",  
"languages" : [  
  "en"  
],  
"location" : {  
  "objectType" : "place",  
  "displayName" : "Australia"  
},  
...  
}
```



Task 3: (35%)

- Dataset: 10000 Tweets; dataset on iLearn “tweets.zip”
- MapReduce: Implement the **Merge Sort**¹ algorithm using Map-Reduce.
- MapReduce: Implement the **Bucket Sort**² algorithm using Map-Reduce.
- Create a short documentation in which you briefly describe your implementation:
 - How many MapReduce Jobs? Why?
 - What to write in the mapper(s) ? Flowchart and Pseudocode !
 - What to write in the reducer(s) ? Flowchart and Pseudocode !
- Sort Tweets, using the object.id :

```
...  
"object" : {  
  "objectType" : "note",  
  "id" : "1345715690143449899009",  
  ...  
}
```



Task 4: (35%)

- Dataset: 10000 Tweets; dataset on iLearn “tweets.zip”
- MapReduce: Implement the **TF-IDF** algorithm using Map-Reduce for the term “health” in the text of the Tweets.
- Create a short documentation in which you briefly describe your implementation:
 - How many MapReduce Jobs? Why?
 - What to write in the mapper(s) ? Flowchart and Pseudocode !
 - What to write in the reducer(s) ? Flowchart and Pseudocode !

```
...  
"link" : "http://twitter.com/frosenpizza/statuses/715691298909368321",  
"text" : "Michael Kidd: primary health care performance initiative establishing goals for the development of global health #fgp16",  
"object" : {  
  "objectType" : "note",  
  ...  
}
```

Submission:

Submit a zip file including:

- A documentation for each task including the Flowchart and Pseudocode
- Source code for the mapper(s) and reducer(s)
- Output for each task

¹ https://en.wikipedia.org/wiki/Merge_sort

² https://en.wikipedia.org/wiki/Bucket_sort