**Project Report**

1. Team Composition and Contributions:

Mr. Akshat Sehgal (UB Person Id – 50198939) – Pink Code (2 pm)

Ms. Saatchi Nandwani (UB Person Id – 50207363) – Blue Code (8 am)

|  |  |
| --- | --- |
| TASK | CONTRIBUTOR |
| CCR Cluster Setup | Saatchi |
| Gathering Inputs | Akshat/Saatchi |
| Map Reduce Program to compute Trigrams | Akshat |
| Empirical Results and Analysis | Saatchi |
| Project Report | Akshat/Saatchi |

2. Connection to CCR Cluster:

We updated the HDFS NameNode, Resource Manager and MapReduce JobHistory hosts and port numbers in the configuration files namely core-site.xml, hdfs-site.xml, yarn-site.xml and mapred-site.xml.

3. Data sets:

We are dealing with 39 articles as we get sufficient number of trigrams for our empirical analysis. We decided this number on the basis of last two digits of the Person ID 50198939. We chose an online text file merging utility for merging all individual articles into one file:  
<http://www.ofoct.com/merge-text-files-online>

As per the project requirements, we replicated 2 individual articles and merged them manually to our final input file.

4. Design Issues:

- setting the number of mappers  
We were not able to make the YARN deploy multiple mappers for our input file.

Reason:  
The default block size of the Hadoop configuration deployed on CCR is 128MB. Our input file size is ~~700kb which is less than the block size. Hence YARN will deploy only 1 mapper for our input text file.

How we reached to this conclusion:  
We tried changing the number of mappers using Job.setNumMapTasks in our code – but this didn’t work as this method is deprecated in latest Hadoop version that is deployed on the cluster (2.7.2)

(2) We tried using –D mapreduce.job.maps=2 as a parameter in our jar command.

(3) Its mentioned in the Hadoop Documentation that the number of splits in the file govern the number of mappers deployed. We tried modifying the split size of the file

5. Empirical Results and Discussions:

//GRAPHS TO DO

How many unique trigrams were identified by your program?

72753

Does a change in the number of mappers and reducers affect the run time of your program?

* Case 1: Using 1 reducer:

Trigrams identified:72753

Execution time of the map reduce job:5.117 sec

* Case 2: Using 2 reducers:

Trigrams identified:72753

Execution time of the map reduce job:2.059 sec

Did you have substantially long wait times before your job was processed?

//TO DO

6. Practical Experiences

Graph