**John Paul Garcia**

**Preetham Wilfred John**

**Big Data Analytics**

**Assignment Number: 1**

**Dr. Boetticher**

**Ms. Rekha Sampangiramaiah**

**Question 4:**

Run the python files topInfluentialPaperMapper.py and topInfluentialPaperReducer.py on Hadoop. Check for output. Datasets: dblp-ref-0.json, dblp-ref-1.json, dblp-ref-2.json, dblp-ref-3.json and dblp-ref-veryshort.json

**Question 5:**

Run the python files top5InfluentialPaperMapper.py and top5InfluentialPaperReducer.py on Hadoop. Check for output. Datasets: dblp-ref-0.json, dblp-ref-1.json, dblp-ref-2.json, dblp-ref-3.json and dblp-ref-veryshort.json

**Question 6:**

Run the python files authorsPaperMapper.py and authorsPaperReducer.py on Hadoop. Check for output. Datasets: dblp-ref-0.json, dblp-ref-1.json, dblp-ref-2.json, dblp-ref-3.json and dblp-ref-veryshort.json

**Question 7:**

Run the python files top5InfluentialAuthorMapper.py and top5InfluentialAuthorReducer.py on Hadoop. Check for output. Datasets: dblp-ref-0.json, dblp-ref-1.json, dblp-ref-2.json, dblp-ref-3.json and dblp-ref-veryshort.json

**Question 8:**

Step1: Run the authormapper.py and authorreducer.py file for getting the list of authors mapped to each paper they have written. Make a note of the output location.

Step2: Run paperMapper.py and paperreducer.py code file for getting the list of influential paper and make a note of the output location.

Step3: Use command “-Dstream.num.map.output.key.fields=2” to sort the output of the map function based on two keys. Run joinMapper.py and joinReducer.py for clubbing authors based upon their citations by providing the input as the two files that was generated above.

Step4: Run authorAvgCitationMapper.py authorAvgCitationReducer.py with the input file as the file generated from the above step, to get the final output of authors(top 10) with their average of top 3 most cited paper.

Datasets: dblp-ref-0.json, dblp-ref-1.json, dblp-ref-2.json, dblp-ref-3.json and dblp-ref-veryshort.json