Project 1 Phase 3 – Hadoop, MapReduce, Spark, Fasttext

Due: 6th 11:59pm

Learning Outcomes for Phase 3:

1. Create local or remote instances of Hadoop and load your dataset – 5 marks
2. Identify the feature you want to analyze and perform word count or any other task using spark context or MapReduce. – 10 marks
3. Explain the working of MapReduce on the data you selected for task2. – 5 marks

Deliverables:

1. **Loading data in Hadoop [5 Marks]:**

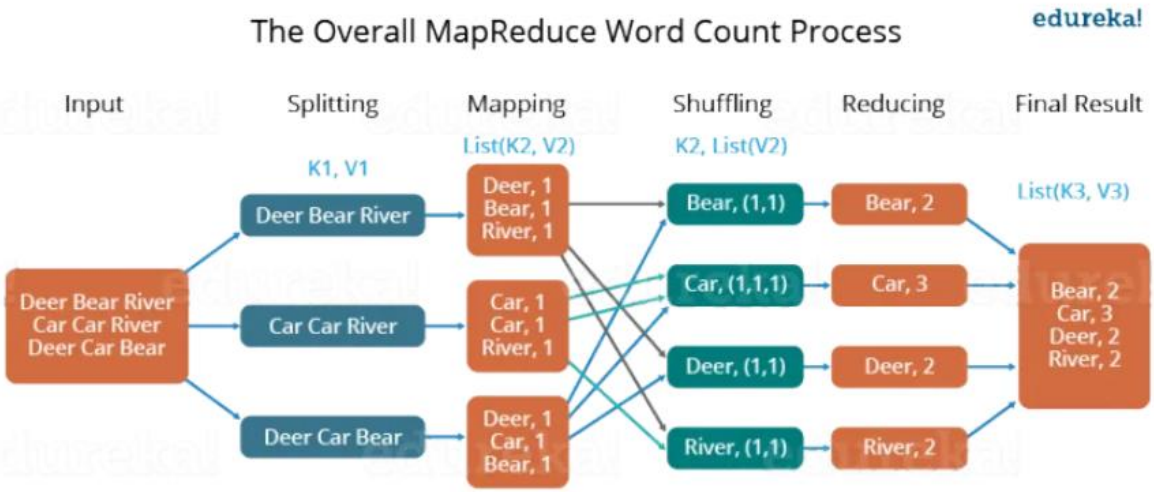
- Create local or remote instance of Hadoop.
- Load all the data that you collected doesn't have to be just 2000 rows.
- Load all the data that you collected into a pandas dataframe.
- Time comparison between Hadoop and dataframe loading.

2. **Using MapReduce [10 marks]:**

- Select a feature from your dataset that you want to analyze using MapReduce.
- Perform word count or any other task of your choice on all the data using MapReduce or Sparkcontext (not just 2000 rows).
- Perform above step without MapReduce.
- Time comparison for above 2 steps.

3. **Working of MapReduce [5 marks]:**

- Similar to the flow chart we discussed in lectures.



MapReduce Example - MapReduce Tutorial

Extra Credit:

1. **Word2Vec/Fasttext [10 marks]:**

- Use word2vec or Fasttext to perform sentiment analysis or similarity search on textual data from your dataset. (limit to 2000 rows)