Lab Final, Session 2020-21

Distributed System - 1.5 Hours

---

1. Find the Longest Word(s) in a Text File [30 marks]

Write a Hadoop MapReduce program to identify the longest word(s) in the input text file.

Requirements:

- Ignore punctuation.
- Words are case-insensitive, but output should preserve the original case from the file.
- If multiple words have the same maximum length, list them all.

Example Input:

Hadoop powers big data applications.
MapReduce is a powerful tool in distributed computing.

Expected Output:

applications
distributed

2. Word Count – Case Insensitive [40 marks]

Write a Hadoop MapReduce job to count the number of times each word appears in the text file, **ignoring case**.

Requirements:

- Treat words like Hadoop, hadoop, and HADOOP as the same.
- Output the words in lowercase (standardized form).

Example Input:

Hadoop is a big data tool.
hadoop helps process data using MAPREDUCE.

Expected Output:

a 1
big 1
data 2
hadoop 2
helps 1
is 1
mapreduce 1
process 1
tool 1
using 1

3. Count Word Lengths [30 Marks]

Write a Hadoop program to count how many words of each **length** are in the input text file.

Requirements:

- Ignore punctuation.
- The output should show the word length and how many words of that length appear in total.

Example Input:

Hadoop is fast and scalable.
MapReduce handles large volumes of data.

Expected Output:

2 1
3 2
4 1
5 2
6 3
8 1
9 1

Explanation:

- “is” → length 2
- “and”, “volumes” → lengths 3 and 7 respectively, etc.