ASSIGNMENT - 7

AIM: Integrate R and Hadoop

PROBLEM STATEMENT: Integrate R & Hadoop & perform the following operations on forest fire dataset.

a. Text mining in R
b. Data analysis using MapReduce in R.

THEORY:

1. TEXT MINING:

Natural languages are different from programming languages. The semantic meaning of statement depends on the context, tone & a lot of other factors

Text mining deals with helping computers understand the meaning of text. Some of the common text mining applications include sentiment analysis

In R, packages useful in understanding text & extracting insights from text & text mining package are as follows:

i. RSQLite, SQLite, for R
ii. tm, framework for text mining
iii. snowballc, text streaming library
iv. wordcloud for making visualisation
v. syuzhet for text sentiment analysis
vi. ggplot2 for data visualisation
vii. quanteda N-grams.

2. TEXT PREPROCESSING:
Text data contains white spaces, punctuations, stop words, etc.

depending on the task at hand, deal with different characters
- convert text to lower case
- Remove numbers
- Remove English stopwords
- Remove extra whitespaces
- Eliminate punctuation marks.

3. CLEANING TEXT IN R:
i. # transform & clean the text
ii. library("tm")
iii. docs ← corpus(vector source(email$a i))

Transformations are done with the tm_map() function to all elements of the corpus.

A document term matrix is important representation for text mining in R tasks & an important concept.

4. WORD CLOUD:
A word cloud is a simple yet informative way to understand textual data to do analysis

library(wordcloud)

For word cloud computing terms with a frequency greater than 30, use following commands:

Page No. :
Date :

wordcloud(names(freq), freq, min.freq = 30,
colors = brewer.pal(3, "Dark"))

CONCLUSION :
In this assignment, we learnt to integrate Hadoop & R & perform text mining & data analysis.