DATE __/__/____
PAGE NO _____

33304

myCompanion

Assignment - 07

* **Aim :-** Integrate R/ Python and Hadoop, and perform the following operations on forest fire dataset
  1. Text mining in R
  2. Data analysis using the map reduce in R

* **Theory :-**

- **Text Mining :**
  It helps computers understand the meaning of the text.
  Eg - Sentimental analysis, Text classification

R is described as a language and environment for statistical computing and graphics. In R, the packages useful in understanding and extracting insights from the text and text mining packages are as follows -
1. RSQLite SQLite interface for R
2. Syuzhet Sentiment analysis
3. Quanted N-grams
4. Word cloud. For making wordcloud visualization

- **Text Processing :**
Text data contains white spaces, special symbols, ending characters etc. These characters do not convey much information about the sentiment of the text, entities mentioned in the text or relationship b/w them.

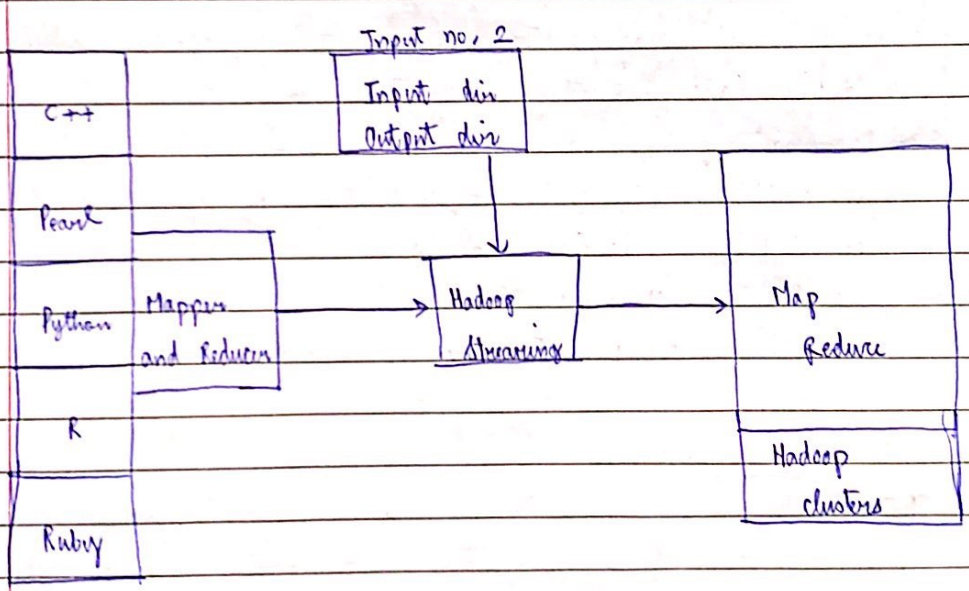
Depending on the task at hand, we deal with such characters differently. This will help isolate text mining in R on important words.
- Convert text to lower case
- Remove numbers
- Remove english stop words
- Remove symbols
- Stemming our text.

- Cleaning text in R →
  library ("tm")
  docs ← corpus (vectorsource ("email/Raw")

tm makes it very easy to create Term document matrix. With that, we can then proceed to build a word cloud

- Word cloud →
  It is a way of representing the frequency of terms in a document.
  library (wordcloud)

- Hadoop streaming →



Hadoop streaming is an utility for running the hadoop map reduce job with executable scripts such as Mapper and Reducer

* Conclusion :- Thus I have performed Text mining using R, and data analysis using MapReduce in R using R Hadoop.