Page No.:
Date:

Assignment No: 7

* **Problem Statement :-**
Integrate R/python & Hadoop and perform following operations on forest fire dataset
(a) Text mining in R
(b) Data analysis using the mapreduce in R

* **Theory :-**
Text mining :-
- Natural languages are different from programming languages.
- The semantic or the meaning of a statement depends on the context, tone & a lots of other factors.
- Unlike programming languages, natural language are ambiguous.
- Text mining deals with helping computers understand the meaning of text.
- It is successfully described as a language and environment for statistical computing and graphic which makes it worth knowing if you are dabbling in the data science or statistics & exploratory data analysis.

* **Text Preprocessing :-**
- Before we dive into analyzing text,

We need to preprocess it.
- Text data contains white spaces punctuations, stop words etc.
- These characters do not convey much information & so hard to process.
- Stemming is the process of reducing inflected words to their stem, base or root form, e.g. Changing 'car', 'cars' to 'car'.

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

Each row of matrix is document vector with one column for every term in the term frequency.

With the document term matrix made, we can then proceed to build a word cloud for Hillary's emails, highlighting with words the most are frequently made.

Page No.
Date

Word ~~Count~~ cloud :-

A word cloud is a simple yet informative way to understand textual data & to do text analysis.

Word cloud is another way of representing the frequency of terms in a document

Here is a size of word indicates its frequency in document corpus.

* Library word ~~counts~~ cloud :-

For word cloud comprising terms with a frequency greater than 30, use following commands.

wordcloud (name (freq), freq, min.freq. = 30, colors = brewer.pal (8, "Dark2"))

* Conclusion :- Thus, I have understand & implement the text mining in R & text preprocessing.