20160910_Batch 17_CSE 7306c_Lab06_FullDay

Learning Outcomes from today's activity:

1. **Text Classification (Authorship Identification)**
2. **Page rank**
3. **TFIDF computation**
4. **Regular Expressions**

Steps for the activity:

1. Text Classification
    a. Reading the documents
    b. Pre-processing of text
        i. library(tm)
        ii. Preprocessing using tm_map
        iii. removeSignature – removes the author of the message,
        iv. stopwords(language='english') – removes stopwords for the language specified
        v. stripWhitespace – removes extra spaces,
        vi. tmTolower – transforms all upper case letters to lower case,
        vii. removePunctuation – removes punctuation symbols,
        viii. removeNumbers – removes numbers,
    c. Vectorization (Creation of DocumentTermMatrix): Once the data is cleansed, create the document term matrix using any of the methods cited below
        i. Use as is DTM matrix
        ii. Use TFIDF matrix
        iii. Use Binary form
    d. Convert the document term matrix as a data.frame to find features and perform modeling
    e. Create classification model on train data
        i. Naïve Bayes classifier (Probabilistic model)
    f. Evaluation Metric - Accuracy