Practical 2

Text Preprocessing

(1) Generate a small collection of documents (minimum five) and perform following :
- Tokenization
- Stop Words Removal
- Stemming
- Lower Case conversion

(2) Represent the corpus using Document-Term matrix using boolean model.

(3) Represent term frequencies

(4) Visualize the data using histogram

(5) Plot word frequencies

(6) Show word cloud

(7) Cluster representations (optional)

*** The content in BLUE color is not yet covered in the theory sessions, so students are required to work on the other part during this lab.**

Moodle

- **Submit following (in zip code)**
  - **script of your code (py/ipnb)**
  - **Your code should display and save Image files for your three plots (as mentioned above)**

PDF

- **Methodology followed using flow chart/algorithm/pseudocode (depict the flow of your entire experiment graphically)**
- **List of necessary packages (whichever is applicable)**
- **the dataset used by you with a brief description (whether standard or manually created)**
- **Sample Input (the details of your datafile)**
- **Sample Output (to be incorporated in Moodle submission)**
- **your learning from this experiment**
- **changes made by you in algorithm parameters (if applicable)**
- **results and your observations/Conclusion**
- **How to make this experiment more interesting or how to extend it?**
- **Rate this experiment on a scale of 1 to 10.**

Important Links

- **Text Mining competition @ Kaggle**
- **Basic Text Mining in R**
- **R Data Mining**
- **Text Preprocessing**
- **Text Mining using Weka**

Demo of Text Miner