CSE508 : Information Retrieval
Assignment 1
Deadline : 19th Jan'18, 2359 hrs

Instructions

- Assignment is to be attempted individually. Please keep the discussions on an abstract level
- Language allowed : Python
- For Plagiarism, institute policy will be followed
- You need to submit ReadMe, code files and analysis.pdf
- You folder should be renamed in the NameRollNo_HW1 format before zipping

Question

You need to construct an unigram inverted index on the dataset given here
Do complete preprocessing of data : Stop word removal, tokenisation etc. You may use libraries for it but clearly state your assumptions.
Construct a word cloud of the complete dataset.
You need to build a CLI which supports following queries

1) x OR y
2) x AND y
3) x AND NOT y
4) x OR NOT y

Where x and y would be taken as input from the user

For query processing, You need to implement

- Merge postings algorithm
- Skip pointers algorithm (only for x AND y)

Analyse how the variation in number of skips affects the performance of your system and write in the report.