Assignment 1

Information Retrieval

- Mayank Chauhan, MT18008

Dataset

Link: [Twenty Newsgroup data set]

Steps to build a Inverted Index

1. Collect the data.
2. Apply tokenization - remove punctuation, accent marks and other non alphanumeric characters like (, >, $, \, \\n, /t, etc.
3. Apply Normalization - convert every text to lower case.
4. Stemming and Lemmatization.
5. Index the tokens and create the posting lists.

Analysis

	Time	Vocabulary Size
No Preprocessing	43 seconds	729603
Only Stemming	4.6 min	185679
Only Lemmatization	1.76 min	203744
No Stop words, Stemming Lemmatization	4.7 min	185376
Stop words, Stemming, Lemmatization	5.5 min	185430

Table 1. Running time and Vocabulary Size for different text processing styles

Positional Inverted Index

Vocabulary Size : 37757 (With Stemming and Lemmatization)