Twitter Keyword Search

Shiza Ali and Mohammad Hammas Saeed

Problem Statement

Keyword search engine that accepts a text file containing a number of tweets as input (each line is one tweet), and provides a simple user interface for querying the list of tweets against keywords.

The system replies to keyword queries with the top 10 tweets that are most relevant to user query keyword(s).

Overview

- Extract tweets using web crawler
- Word segmentation
- Build inverted index to search the word
- Store the inverted list using relevant data structure such as Hash Tables
- Input query
- Sort the similar tweets, and select the top 10 most similar

Inverted Index

- To gain the speed benefits of indexing at retrieval time, we have to build the index in advance. The major steps in this are:
  - Collect the tweets to be indexed
  - Tokenize the text
  - Turn each document into a list of tokens
  - Do linguistic preprocessing
  - Index the documents that each term occurs in by creating an inverted index, consisting of a dictionary and postings.

Inverted Index