CS221 Project 3 Milestone 1

JUN GUO, ID: 83322637
ZHEN CHEN, ID: 85282960

March 3, 2017

1 Data Structure

We design a map from the tokenized terms to the documents with the posting that includes (1) document ID, (2) tf, (3) $tf - idf$, (4) the positions in the head, and (5) the positions in the body. It is unnecessary to store the parameter *df* since the df for each term is just the size of the postings for the corresponding term, i.e., $df[term] = len(Index[term])$. More details about the data structure used in our index is illustrated by Fig. (1).

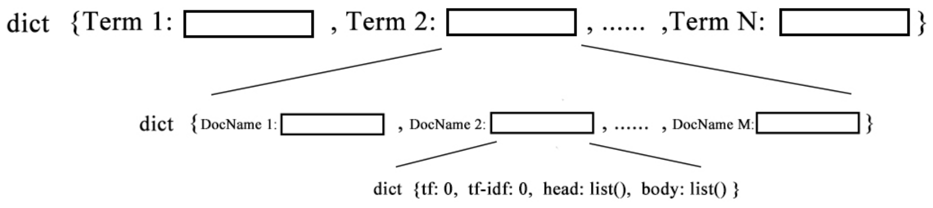
dict {Term 1: , Term 2: , ... ,Term N: }
dict {DocName 1: , DocName 2: , ... , DocName M: }
dict {tf: 0, tf-idf: 0, head: list(), body: list() }

Figure 1: Index structure.

2 Functions

We implement the following functions: (1) $listindexer()$: map the visible tokenized terms to the sorted index whose structure is introduced in the last section.
(2) *bool* $visible(element)$: decide if the given element is visible on web.
(3) *void* $get\_stem(content)$: extract stems from the content.
(4) *bool* $isstopword(word)$: decide if the given word is a stop word.
(5) *list* $termProcessing(content)$: remove invisible content and do tokenizing (including stemming, removing stop words, and the process of transferring the capital to the lower case).
(6) *int* $calculate\_Tfidf(tf, idf)$: calculate tf-idf according to the given tf and idf.

(7) *void compression(dict)*: compress data by utilizing delta encoding.
(8) *void save_data(dict)*: In this function, we provide two kinds of saving methods: (i) saving the complete index into one file and (2) saving the index into 36 files in terms of the initial alphanumeric characters. Moreover, we save the statistic results, such as the maximum df and the minimum df, into another file.

3 Results

Some important indexing results are shown in table 1.

Table 1: Simulation Parameters

Parameters	value
the number of documents	37497
the number of unique words	537510
the total size of the index	546.2M
the running time	4977.10s
the maximum df	19834
the minimum df	1
the most frequent term	“uci”
the maximum tf-idf	30.7711
the maximum tf-idf	0.2766