CS6370: Information Retrieval

Assignment 2

Description

In this assignment, you need to build Inverted Index for the corpus that you obtained in Assignment 1.

1 Section I

For Index creation, use the following steps.

- **Step 1:** Treat URL, email, date as individual tokens. Use Regular expression matching. Perform tokenization. Build Inverted Index.
- **Step 2:** Remove stopwords, and ensure each token should be of at least length 2. Build Inverted Index. Stopwords list can be found here: `http://www.lextek.com/manuals/onix/stopwords1.html`.
- **Step 3:** Perform Stemming (or lemmatization) and build Inverted Index.
- **Step 4:** Remove the least frequent terms (the terms which are occurring in $\leq 1\%$ of documents). Build Inverted Index.

At each and every step (From Step 1 to Step 4), compute the following statistics:

1. Number of Terms
2. Maximum Length of Postings List
3. Minimum Length of Postings List
4. Average Length of Postings List
5. Size of the file that stores the inverted index

Mention the above statistics in your report.

2 Section-II

After completing all the above steps (from Step 1 to Step 4), compute the following information.

1. Most frequent K words (Here frequent in terms of document frequency), say $K = 20$.
2. Postings List size for each of the above K words.

3. For each of these words, find *Average Gap Size* in the Postings List .

Repeat the above 3 steps for median K words, and least frequent K words.

Include these statistics in your reports.

3 Section-II

1. While constructing the index for Step 3 (see previous page), note down the number of tokens (T) seen and the number of terms in the dictionary (M) after you go through each document. Plot the curve for $\log_{10} M$ vs $\log_{10} T$.
2. Sort the terms in decreasing orders of their frequencies (after Step 3 of inverted index creation). Let y_i be the frequency of the i^{th} most frequent term. Plot the curve for $\log_{10} y_i$ vs $\log_{10} i$.
3. Consider the inverted index obtained after Step 3 of inverted index creation. Note down the gaps in the postings list. Create a separate text file where each row will contain the term id and the list of gaps (instead of actual document ids in postings). Plot the histogram of these gaps. You can decide the bin sizes of the histograms yourself.

What are your observations from the above plots?

Submission Instructions:

1. Submit a single zip file. It should contain your code, a README file to give instructions about executing your code, and the report.
2. File name convention: GroupID_<RollNos of group members> (without the angle brackets).
3. Compulsorily mention the names of the group members in each code file and report.
4. Report should be self contained.
5. Report should be in doc/docx/pdf format.
6. No plagiarism.