CS6200 Information Retrieval
Fall 2019
Instructor: Omar Alonso

Homework #1. Build a simple inverted file

Objective: process a collection of documents, create an inverted index on titles, and run queries.

This programming assignment involves writing three components:

1. An indexer
2. Term distribution output file for top 100 terms.
3. Graph term distribution for the entire output (a screenshot would work)

Prerequisites

1. Use the AP89_DATA.zip collection.

Document Indexing

Create an index of the downloaded corpus. The documents are found within the ap89_collection folder in the data .zip file. You will need to write a program to parse the documents and index the title field (<HEAD></HEAD>).

The corpus files are in a standard format used by TREC (Text REtrieval Conference). Each file contains multiple documents. The format is similar to XML, but standard XML and HTML parsers will not work correctly. Instead, read the file one line at a time with the following rules:

1. Each document begins with a line containing <DOC> and ends with a line containing </DOC>.
2. The first several lines of a document's record contain various metadata. You should read the <DOCNO> field and use it as the ID of the document.
3. The tags <HEAD> and </HEAD> describe the title of the document.
4. The document contents are between lines containing <TEXT> and </TEXT>.
5. All other file contents can be ignored.

What you need to submit

1. Your indexer's Java source code
2. A description of your indexing design. How are you approaching the problem, your specific design choices, and other information that you think is important to mention?
3. Screenshot of the term distribution
4. Documentation on how to compile and run the code. I'll be testing your work.