CS6200 Information Retrieval
Summer 2018
Instructor: Omar Alonso

Homework #1. Indexing and querying a TREC data set

Objective: process a collection of documents, create an index on Elasticsearch, and run queries. These instructions will generally not spell out how to accomplish various tasks in Elasticsearch; instead, you are encouraged to try to figure it out by reading the online documentation.

This programming assignment involves writing two programs:

1. A program to parse the corpus and index it with Elasticsearch (50 points)
2. A query processor, which runs queries from an input file (50 points)

Prerequisites

1. Install Elasticsearch and Kibana
2. Download AP89_DATA.zip from the Dropbox Data link shared via email. Email me if you have problems downloading the data set.

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 send them to your Elasticsearch instance.

The corpus files are in a standard format used by TREC. 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 document contents are between lines containing `<TEXT>` and `</TEXT>`.
4. All other file contents can be ignored.

Query Execution

Write a program to run the queries in the file query_desc.51-100.short.txt, included in the data .zip file. You should run all queries (omitting the leading number) and output the top 50 results (`DOCNO`) for each query to an output file. If a particular query has fewer than 50 documents with a nonzero matching score, then just list whichever documents have nonzero scores. Note that you need to construct the query using the information provided in the file.

What you need to submit

1. Your indexer's Java source code
2. Your query program's Java source code
3. Documentation on how to compile and run the code. I'll be testing your work with my own Elasticsearch instance and documents so please be precise on the instructions on how to test your code.