CS6200 Information Retrieval
Fall 2019
Instructor: Omar Alonso

Homework #4. Solr with a TREC data set

Objective: process a collection of documents, create an index on Solr, and run queries. These instructions will generally not spell out how to accomplish various tasks in Solr; 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 Solr (50 points)
2. A query processor, which runs queries from an input file (50 points)

Prerequisites

1. Download and install Solr (http://lucene.apache.org/solr/)
2. Use the AP89_DATA.zip collection.

Document Indexing
You will need to write a program to parse the documents and send them to your Solr instance.

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` and `TITLE`) for each query to an output file.

The file `qrels.adhoc.51-100.AP89.txt` contains the relevance assessments. By looking at the query description file, we know the content of topic 91:

```
91.   Document will identify acquisition by the U.S. Army of specified advanced
weapons systems.
```

The qrels tell us which files are relevant or not:

```
91 0 AP890111-0065 0
91 0 AP890113-0040 1
```

Looking on the files again:

```
Defense Minister Says Short-Range Nuclear Weapons to be Modernized (NR)
Military Buildup At a Glance (R)
```

What you need to submit

1. Your indexer's Java source code and how to load the documents to Solr.
2. Your query program's Java source code.
3. Compute the precision of the search results according to the qrels.