CS6200 Information Retrieval
Fall 2020
Instructor: Omar Alonso

Practice Solr/ElasticSearch with a dataset

Objective: process a collection of documents, create an index on Solr or ElasticSearch, and run queries. These instructions will generally not spell out how to accomplish various tasks in Solr/ElasticSeach. You are encouraged to try to figure it out by reading the online documentation.

Prerequisites

1. Download and install Solr (http://lucene.apache.org/solr/) or ElasticSearch (https://www.elastic.co/)
2. Use the dataset and queries from homework #1 for creating the index and for issuing queries.

Homework: crawling and link statistics.
Deadline December 10.

Objective: Implement a simple focused crawler. Your task is to use a pick a topic (e.g., politics, sports, etc.) and provide a set of seeds urls for your crawler.

This assignment involves:

1. Define a topic and provide a list of urls (links) that you believe are relevant. Please use English only topics/links.
2. Implement a single-threaded crawler that uses the output of (1) to crawl the web with two parameters: number of links and depth (levels).
3. Normalize links. That is, eliminate parameters, redirects, etc.
4. Extract 20,000 links and compute the following statistics from the generated data set:
    a. Number of unique links extracted
    b. Frequency distribution by domain
    c. Breakdown of links by type (e.g., text, image, video)
    d. Average link depth
    e. For each crawled page, compute the number of incoming and outgoing links. Report the top-25 pages with the highest number of incoming and outgoing links.
    f. Plot the top-50 domains ranked by highest number of incoming links. Note that this is a computation for domains (e.g., cnn.com, bbc.co.uk) and not individual pages.

What you need to submit

1. Your single-threaded focused crawler
2. The output statistics and charts for item (4).
3. Documentation on how to compile and run the code.