CSCE 5300 Introduction to Big Data and Data Science

ICE - 5

Lesson Title: ***Solr & Lucene***

Lesson Description: ***Parallel Indexing: Solr and Lucene***

Lesson Overview:
Solr is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest internet sites.

In Class Exercise:
Source code: https://github.com/apache/solr/tree/main/solr/example

1. ***Book Dataset***
   Dataset:
   https://github.com/apache/solr/blob/main/solr/example/exampledocs/books.csv
   A. List all books and all their attributes in the dataset
   B. Find books whose book name is a game of thrones and a clash of kings and Display the Author of those book.
   C. Find books whose Series_t is the chronicle of prydain and their genre_s.
   D. Find books which have more than 1 sequence_i.
   E. Find all book whose genre_s is Fantasy.
2. ***Film Dataset***
   Dataset: https://github.com/apache/solr/blob/main/solr/example/films/films.csv
   A. List all films movies with initial_release_date, genre, and director_by.
   B. Find films whose director_by is Zack Snyder.
   C. Find films whose genre is Fantasy and director is David Yates.
   D. Find films whose initial release date is before 2005-12-10 and sort the result-set in descending order by date of release.
   E. List all director_by which is released on year 2006.

ICE Submission Guidelines

1. ICE Submission is individual.
2. ICE code has to be properly commented.
3. The documentation should include the screenshots of your code and results.
4. Provide the explanation of the exercise for each question as per your understanding.
5. The similarity score for your document should be less than 15%.
6. Submit the source code (properly commented) and documentation (.pdf/.doc) with explanation and screenshot of source code/queries having input logic and output results.
7. Submission after the deadline is considered as late submission.