CSI4107 - Information Retrieval
Winter 2019
Search Engine Programming Project

1. Overview of the Search Engine Programming Project

The table below provides the important information about the logistics of the project. Please read
carefully to make sure you do not miss any information and deadlines.

Purpose	Develop your own Search Engine (SE). Have hands-on experience with the various modules which could be part of a search engine: retrieval model, user interface, relevance feedback, document clustering, etc.
Development	The project will be done in 2 phases. In phase 1, you must develop a Vanilla System which will be a first complete search engine with limited capabilities. In phase 2, you must include additional features to your system and perform a formal evaluation.
Teams	The programming project can be done in teams (max 2) or individually. The number of modules to be included in your Search Engine will be in relation to the number of students working together.
Programming language	MUST be either Java or Python.
Percentage	Vanilla system (15%), Final system (15%)
Important dates	Vanilla system → Feb. 20, 2019, Final system → April 12, 2019 *No extensions will be given. -10% per day late.*
What to submit ?	You must submit a single zipped file containing: (a) A small report (pdf) with a title Page indicating the names and student numbers of team members, the name of your system, its functionalities, and a diagram of your system architecture. (b) a README file that will explain how to run your code (c) the Java or Python classes of your system The UI must be self-explanatory so that your search engine could be easily tested. If you work in teams, a single student should submit.
How to submit ?	Two links will be provided in Brightspace, in the Search Engine Project Module, one in February for the Vanilla system and one in April for the final system.
Mandatory source acknowledgement	In today's world of available snippets of code everywhere on the Internet, you will certainly copy/paste from various sites in addition to writing your own code. You MUST provide the source (URL links) of each site where you copied code from and say "Inspired from X" or "modified from X" in your class comments. Important: Failure to include links to sources will result in a grade of 0 for the project.

2. Collections

Your Search Engine will have to index and retrieve documents from 3 collections.

Name	Description	Link
UofO-CSIcourses	A list of course descriptions. This collection will be used for the vanilla system.	https://catalogue.uottawa.ca/en/courses/csi/
Reuters 21578	A news collection, often used for text categorization and information retrieval. This will be required for the final system.	http://www.daviddlewis.com/resources/testcollections/reuters21578/
Web collection	Your choice of a site (or a few sites) to crawl.	

3. Development of the Vanilla System

The vanilla (or baseline) system must perform retrieval operations on the UofO-CSIcourses collection. The vanilla system must include the following mandatory modules, and either 1 (for people working alone) or 3 (for people working in teams) optional modules.   All modules are individually described in file CSI4107-SE-Modules.pdf.

Mandatory modules
1. Corpus pre-processing
2. User Interface
3. Dictionary building
4. Inverted Index Construction
5. Corpus Access
6. Boolean model of information retrieval
7. Vector Space Model of information retrieval

Optional modules
● Phrase query indexing
● Spelling correction with Edit Distance
● Spelling correction with Soundex
● Wildcard management with additional bigram indexing
● Relevance feedback
● BM25 (probabilistic model)

4. Development of the Final System

Requirements for the final system will be provided at a later time (during study break).