Kevin Courey

Prof. Sabine Bergler

COMP 479

November 24, 2025

Project 2 Demo

The main method for this Project is located in the Driver module. Within this method, an inverted index is constructed and its instance method run() is invoked. The run method invokes a sequence of methods responsible for the following: determining whether the index will be populated using BSBI or SPIMI (at the moment, SPIMI is the only choice that can be made), determining the limit on the number of terms to be stored within the index, populating the index, sorting the index, and writing the sorted contents to a file labelled "InvertedIndexContents.txt" located within the directory \Comp479_Project2\Results for the sake of future analysis. The populate method is responsible for activating the web crawler from within the WebCrawler module located in the directory \COMP479_Project2\MyCrawler\MyCrawler\spiders\.

Within this module is the logic responsible for handling all web crawling through the spectrum.library.concordia.ca domain in search of Master's and PhD theses, as well as other pdf files. After analyzing the spectrum website, starting from "https://spectrum.library.concordia.ca", I discovered that the following pathway would lead my crawler to all target documents using a minimal number of steps:

1) https://spectrum.library.concordia.ca/
2) https://spectrum.library.concordia.ca/view/document_subtype/
3) https://spectrum.library.concordia.ca/view/document_subtype/thesis/
    a) https://spectrum.library.concordia.ca/view/document_subtype/thesis=5Fmasters/

b) https://spectrum.library.concordia.ca/view/document_subtype/thesis=5Fphd/

After examining these hyperlinks and taking note of the key words that distinguish them from the other links located within the spectrum website, I was able to establish a set of rules that would ensure my crawler's ability to locate the target documents in a timely manner. The web crawler is also responsible for extracting the text within each target document, tokenizing it, and adding each token to the inverted index that was constructed in the main method. Additionally, the web crawler keeps track of the downloaded files, as well as their contents and stores this information within \COMP479_Project2\Downloads\. Once the web crawler has successfully downloaded a user-specified number of pdf documents, the web crawler will deactivate, marking the end of the inverted index's populate method. Shortly thereafter, the inverted index's run() method will have completed its operations, which leads to the next portion of the main method to be executed.

This portion involves the construction of a QueryProcessor, and invoking its run method, which is responsible for accepting a user's query, normalizing it, and determining which documents within the inverted index are relevant to the user's normalized query using the intersect method. A user can enter as many queries as they want, with each one resulting in a list of zero or more "relevant" documents being returned to the user. The retrieved documents are also stored within a text file labelled "My-collection.txt" located in the directory \Comp479_Project2\Results.

Once the user has decided he no longer wants to enter any more queries, the final portion of the program is executed and the documents that were collected by the web crawler are clustered using the k-means module from the scikit-learn library. Three clusterings are performed

on these documents (k=2, k=10, k=20), the results of which are saved to a text file labelled "clusterInfo.txt" located in the directory \COMP479_Project2\Results\.