TDT4117 Information Gathering

Assignment 4

Erling Storaker Moen

**Task 1: Page rank and HITS**

a) *Compare page rank and HITS and briefly describe the main ideas of both approaches and point out their differences.*

The idea behind PageRank is that if a page is important if it is pointed out by other important pages. The importance of your page is decided by your page’s PageRank score, which is set by calculating the PageRank’s of all pages that point to your page.

PageRank use a recursive scheme like HITS algorithm, but the PageRank algorithm produces a ranking independent of a user’s query. HITS was developed as an algorithm that made use of the link structure of the web in order to discover and rank pages relevant for a particular topic. The hyperlink- topic search algorithm was developed more by how humans analyse a search process rather than just machines searching for a topic and returning everything that matched. HITS is also known as hubs and authorities. A page is called an authority for a query if it contains valuable information on the topic and was linked there by many hubs and a hub if the information on a page was not authoritative, but rather linked to many other pages.

**Strengths of HITS:**

* Ability to rank pages according to the query topic, resulting in relevant authority and hub pages.

**Weaknesses of HITS:**

* Does not detect advertisements, like sites that have commercial advertising sponsors that relates to your search.
* Can be spammed easily since people can add out-links on their own pages affecting the hub-score.

**Strengths of PageRank:**

* Very robust against spam

**Weaknesses of PageRank:**

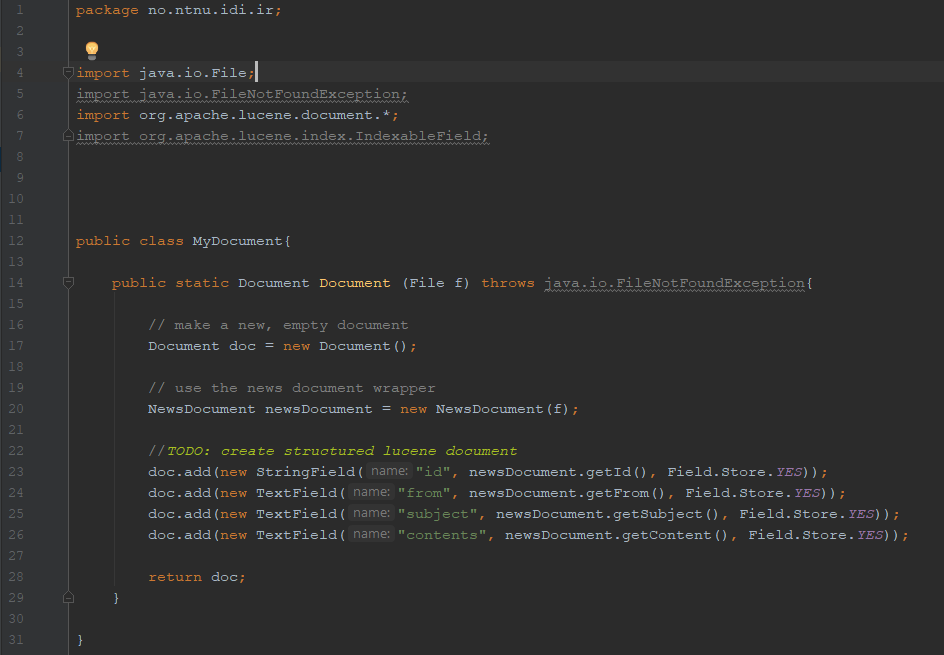
* Favours older pages since a new page, even if it is a very good page, will not have many links unless it is part of an already existing website.
* PageRank can easily be increased using “link farms”

**Main differences:**

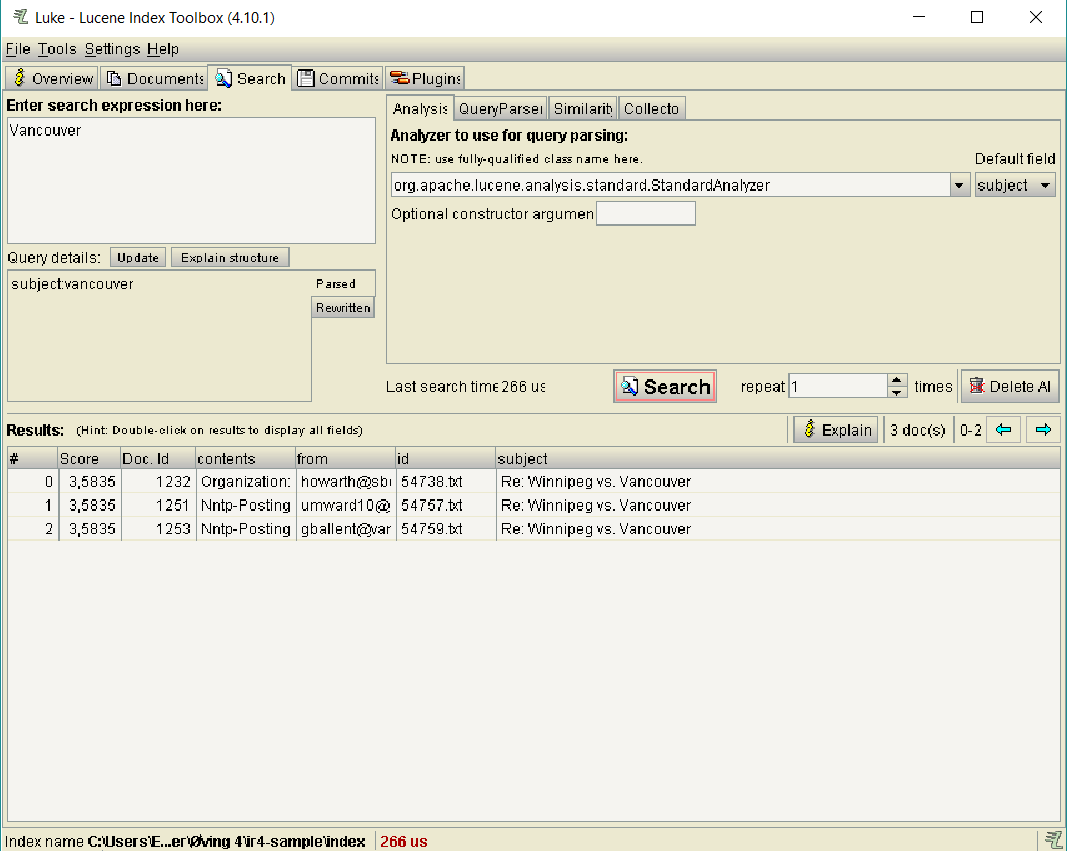
* HITS is query dependant, meaning the authority and hub scores are dependant on the search terms.
* HITS sets two scores per document, while PageRank sets one.
* HITS is sensitive to user query, while page PageRank is not.
* PageRank is less susceptible to link spam and more efficient.
* HITS does computations at query time, while PageRank is slow.
* PageRank assigns only one score to each page, HITS assign two scores, the authority of the page that estimates the contents value, and the hub value, that is the value of the page’s link to other pages.

**Task 2: Structured Indexing and Retrieval in Lucene**

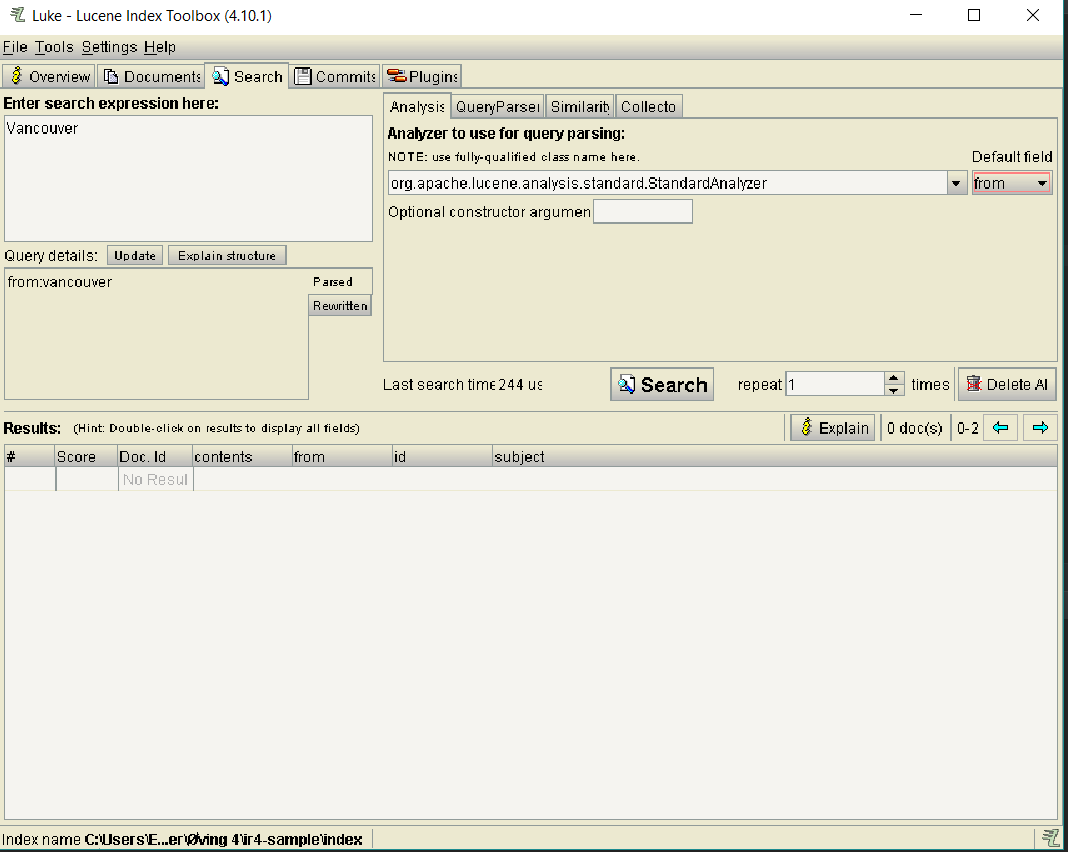
Subtask A: Implementation of MyDocument Class in Java

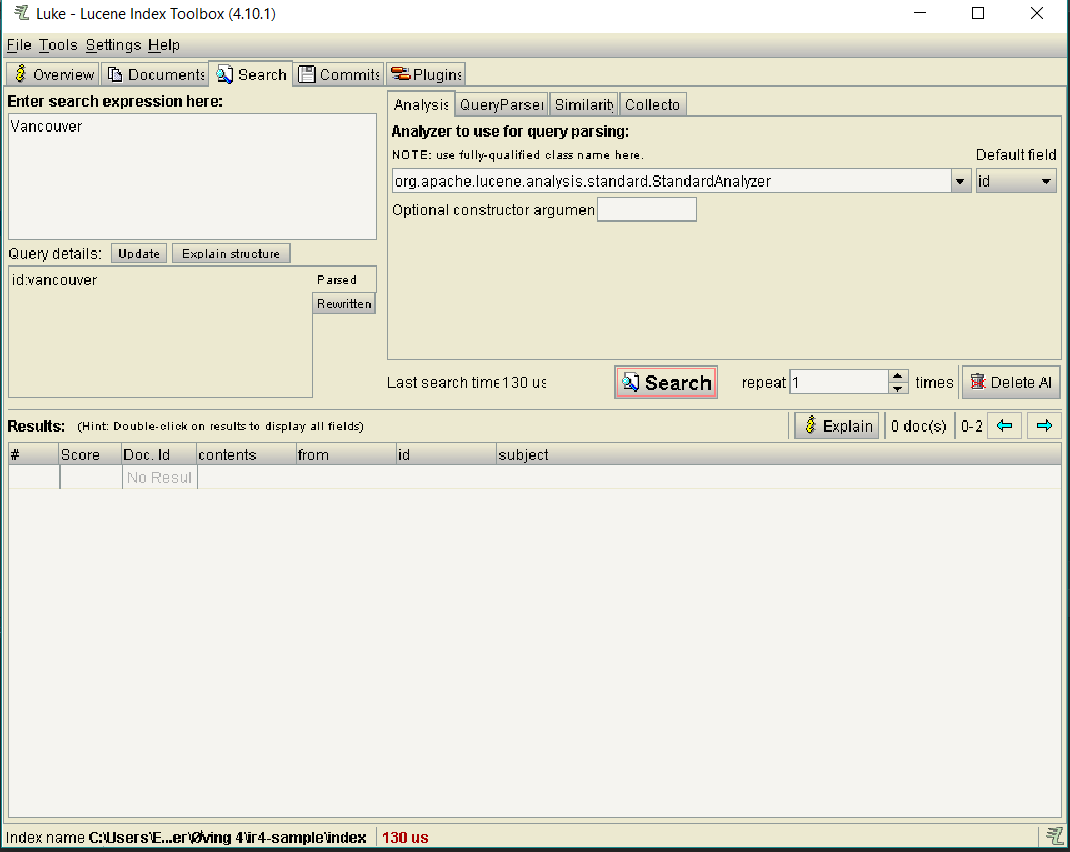


Subtask B

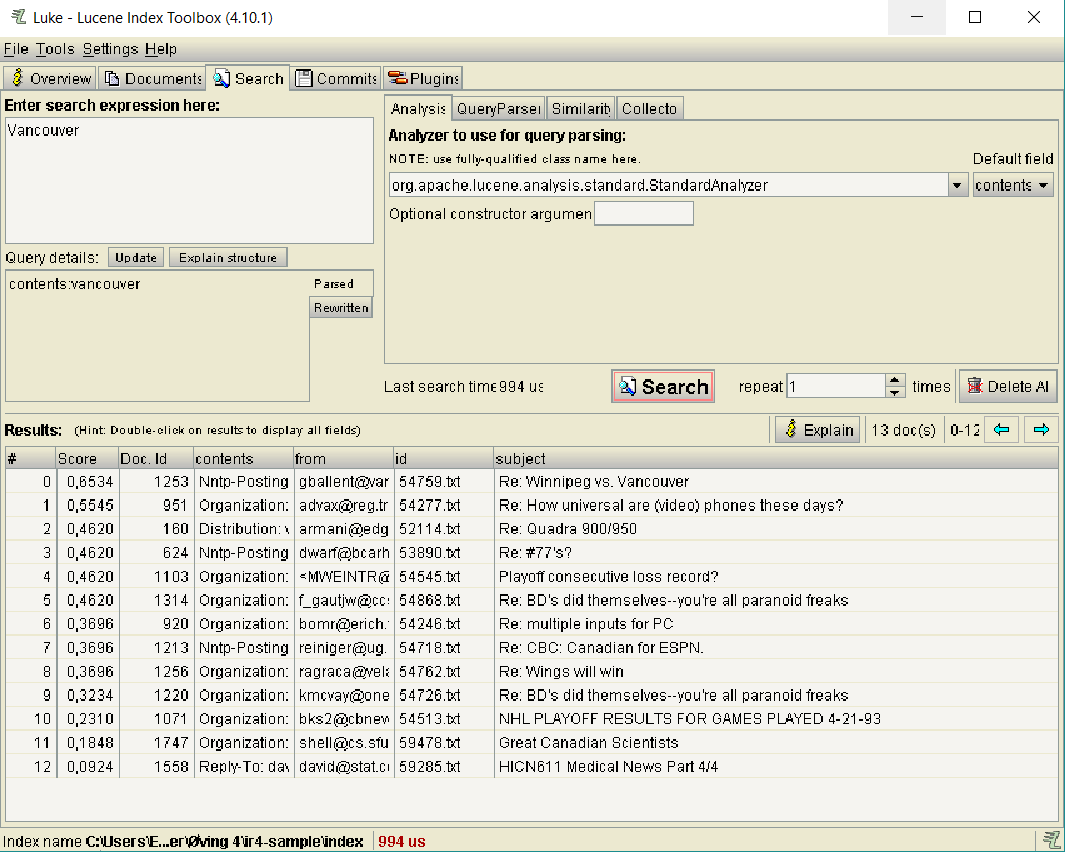


Subject returns the collection that contained Vancouver. All documents gets returned with the same score.





Neither “from” nor “id” fields consisted of the term Vancouver, so no documents were returned.



Contents returns every document that contained the Query. Also returns the respective weight score calculated