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:

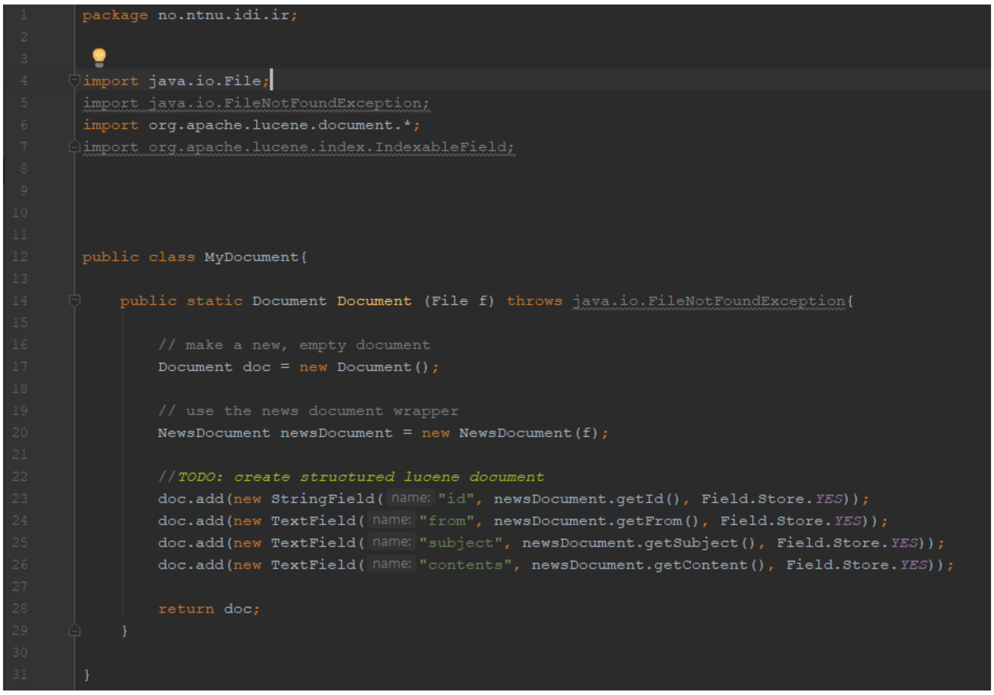
- Favours older pages since a new page, even if it is a very good page, will not have many links unless it is part of an already existing website.
- PageRank can easily be increased using “link farms”

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.
-

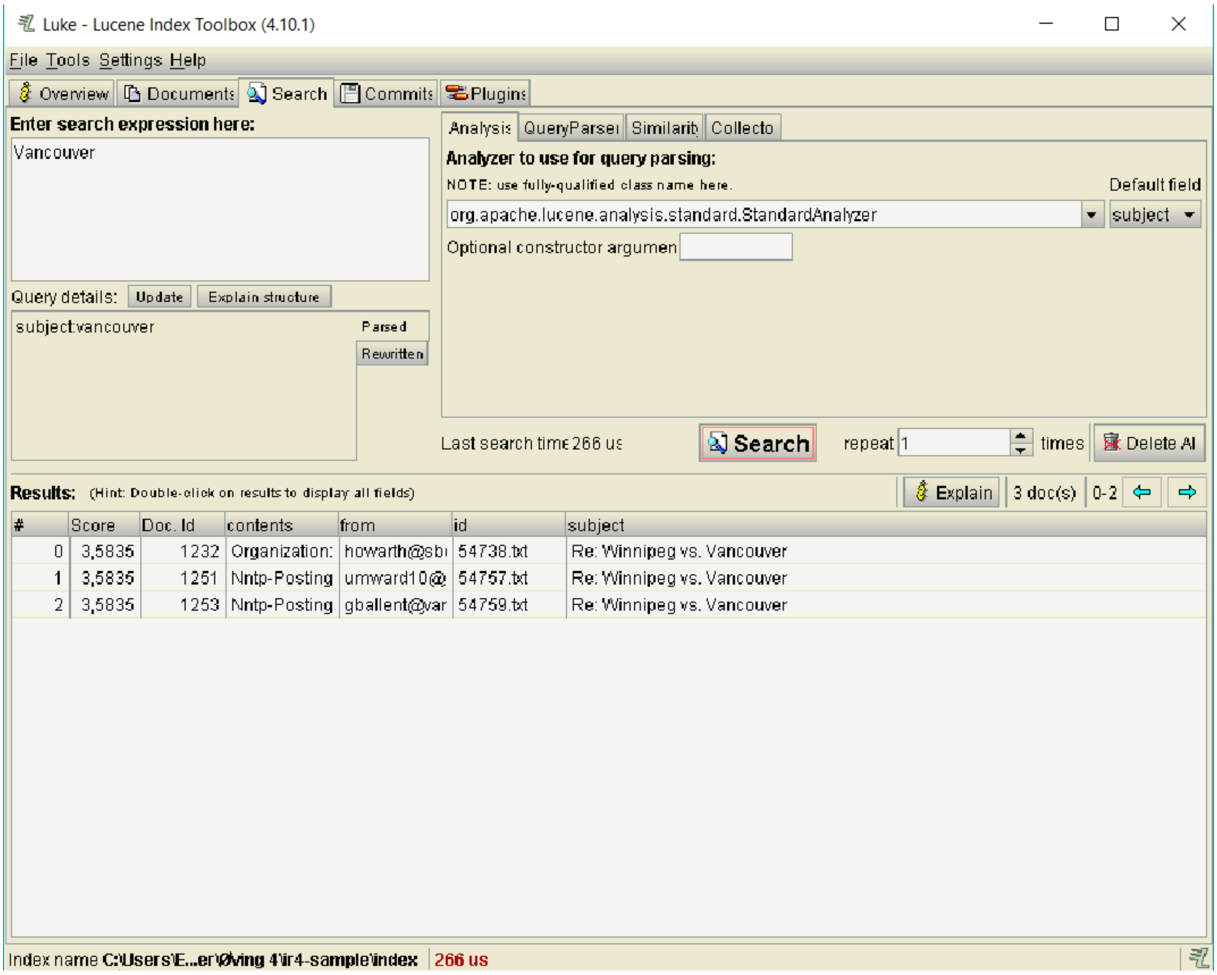
Task 2: Structured Indexing and Retrieval in Lucene

Subtask A: Implementation of MyDocument Class in Java



```java
package no.ntnu.idi.ir;


import java.io.File;
import java.io.FileNotFoundException;
import org.apache.lucene.document.*;
import org.apache.lucene.index.IndexableField;




public class MyDocument{

    public static Document Document (File f) throws java.io.FileNotFoundException{

        // make a new, empty document
        Document doc = new Document();

        // use the news document wrapper
        NewsDocument newsDocument = new NewsDocument(f);

        //TODO: create structured lucene document
        doc.add(new StringField( name: "id", newsDocument.getId(), Field.Store.YES));
        doc.add(new TextField( name: "from", newsDocument.getFrom(), Field.Store.YES));
        doc.add(new TextField( name: "subject", newsDocument.getSubject(), Field.Store.YES));
        doc.add(new TextField( name: "contents", newsDocument.getContent(), Field.Store.YES));

        return doc;
    }

}
```

Subtask B



Enter search expression here:

Vancouver

Query details: subject:vancouver

Analyzer to use for query parsing:

NOTE: use fully-qualified class name here.

org.apache.lucene.analysis.standard.StandardAnalyzer

Default field: subject

Optional constructor argumen

Last search time 266 us

repeat 1 times

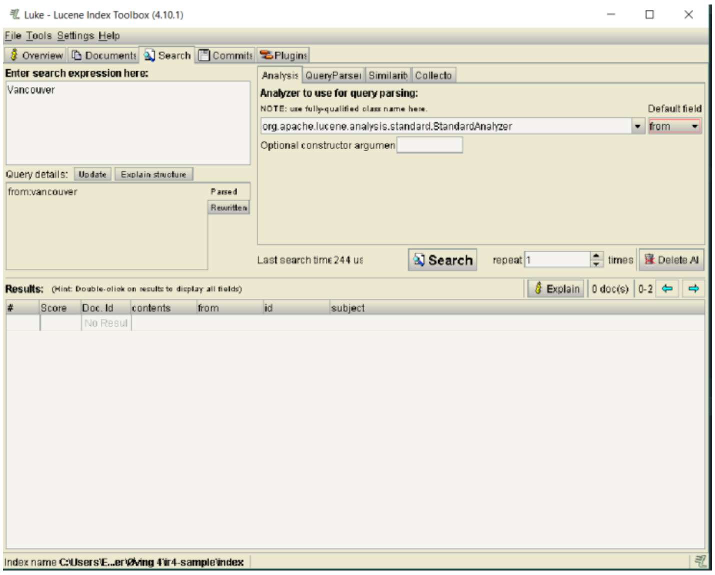
Results: (Hint: Double-click on results to display all fields)

3 doc(s) 0-2

#	Score	Doc. Id	contents	from	id	subject
0	3,5835	1232	Organization:	howarth@sbi	54738.txt	Re: Winnipeg vs. Vancouver
1	3,5835	1251	Nntp-Posting	umward10@	54757.txt	Re: Winnipeg vs. Vancouver
2	3,5835	1253	Nntp-Posting	gballent@var	54759.txt	Re: Winnipeg vs. Vancouver

Index name C:\Users\E...er\Øving 4\ir4-sample\index 266 us

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



Enter search expression here: Vancouver

Query details: from:vancouver

Analyzer to use for query parsing: NOTE: use fully-qualified class name here.

org.apache.lucene.analysis.standard.StandardAnalyzer

Default field: from

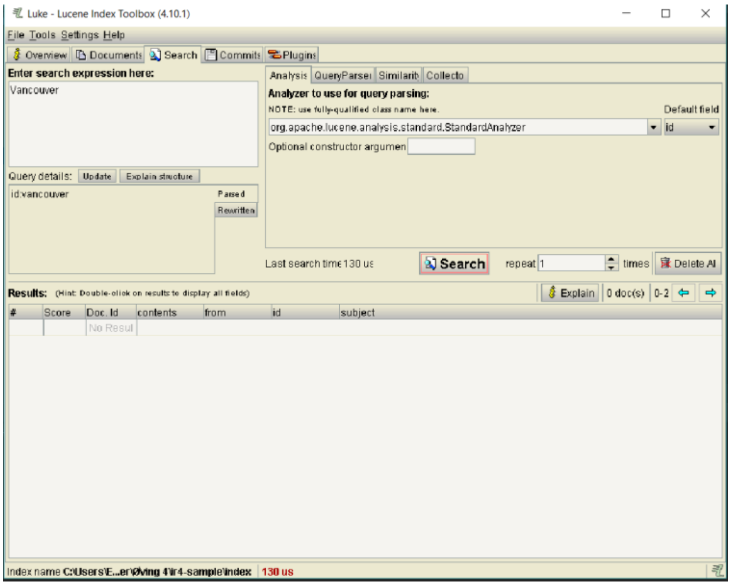
Optional constructor argumen

Last search time 244 us repeat 1 times

Results: (Hint: Double-click on results to display all fields) 0 doc(s) 0-2

No Resul

Index name C:\Users\E...er\Øving 4\ir4-sample\index



Enter search expression here: Vancouver

Query details: id:vancouver

Analyzer to use for query parsing: NOTE: use fully-qualified class name here.

org.apache.lucene.analysis.standard.StandardAnalyzer

Default field: id

Optional constructor argumen

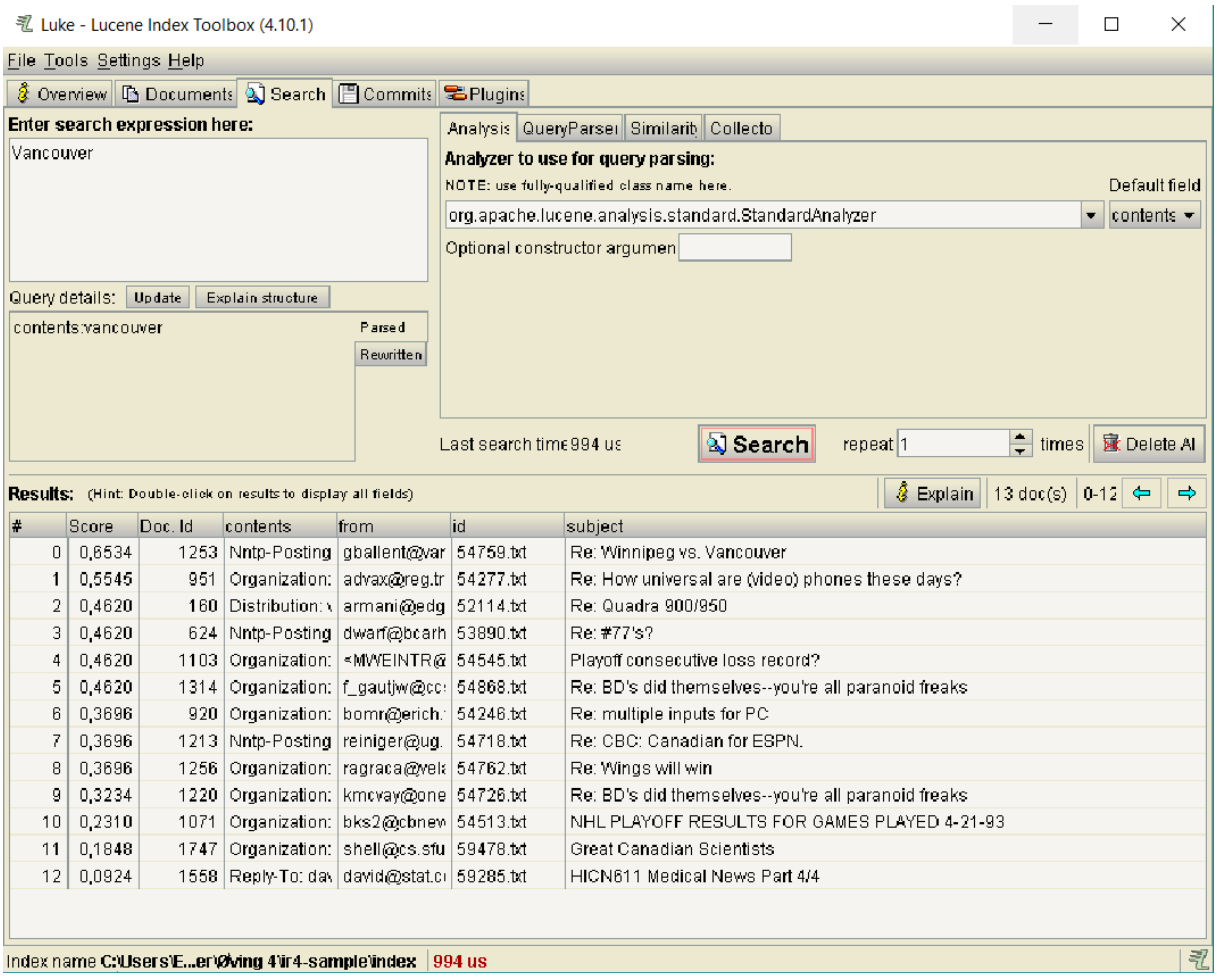
Last search time 130 us repeat 1 times

Results: (Hint: Double-click on results to display all fields) 0 doc(s) 0-2

No Resul

Index name C:\Users\E...er\Øving 4\ir4-sample\index 130 us

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



Enter search expression here: Vancouver

Query details: contents:vancouver

Analyzer to use for query parsing:

NOTE: use fully-qualified class name here.

org.apache.lucene.analysis.standard.StandardAnalyzer

Default field: contents

Optional constructor argumen

Last search time 994 us

repeat 1 times

Results: (Hint: Double-click on results to display all fields)

#	Score	Doc. Id	contents	from	id	subject
0	0,6534	1253	Nntp-Posting	gballent@var	54759.txt	Re: Winnipeg vs. Vancouver
1	0,5545	951	Organization:	advax@reg.tr	54277.txt	Re: How universal are (video) phones these days?
2	0,4620	160	Distribution: \	armani@edg	52114.txt	Re: Quadra 900/950
3	0,4620	624	Nntp-Posting	dwarf@bcarh	53890.txt	Re: #77's?
4	0,4620	1103	Organization:	<MWEINTR@	54545.txt	Playoff consecutive loss record?
5	0,4620	1314	Organization:	f_gautjw@cc:	54868.txt	Re: BD's did themselves--you're all paranoid freaks
6	0,3696	920	Organization:	bomr@erich.'	54246.txt	Re: multiple inputs for PC
7	0,3696	1213	Nntp-Posting	reiniger@ug.	54718.txt	Re: CBC: Canadian for ESPN.
8	0,3696	1256	Organization:	ragraca@vel:	54762.txt	Re: Wings will win
9	0,3234	1220	Organization:	kmcvay@one	54726.txt	Re: BD's did themselves--you're all paranoid freaks
10	0,2310	1071	Organization:	bks2@cbnew	54513.txt	NHL PLAYOFF RESULTS FOR GAMES PLAYED 4-21-93
11	0,1848	1747	Organization:	shell@cs.sfu	59478.txt	Great Canadian Scientists
12	0,0924	1558	Reply-To: da\	david@stat.cı	59285.txt	HICN611 Medical News Part 4/4

Index name C:\Users\E...er\Øving 4\ir4-sample\index 994 us

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