TDT4117
Assignment 4

Pål-Edward Larsen (paaledwl)

Johannes Kvamme (johannkv)

18.11.2018

Task 1: Page Rank and HITS

Comparison and main ideas

Page Rank
PageRank gives a webpage an importance rating based on how many links lead to it, the importance of the pages that lead to it, resulting in the probability that a user randomly visits the page. Also, PageRank includes a dampening factor, which is the probability that a user will continue to click around, and not stop visiting pages.

HITS
HITS, or Hyperlink-induced Topic Search, gives a webpage two values as a rating.
Authority is the page's value as an authority on a subject, calculated by the value of the hubs that link to it.
Hub is the page's as a link hub, calculated by the value of the authorities it links to.
To do this, it gets an answer set often called as root set. From this, it generates a subset named base set, which is one link adjecent, both tin/out from the set.

Differences
While HITS calculates two different values, PageRank cares only for one. HITS use the two values to calculate importance as a link hub and importance as an authority on a topic, with authority being incoming links and hub being outgoing, while PageRank calculates the single importance based on the rank of the incoming links' page.
HITS calculate on query while PageRank calculate on crawl, making PageRank more effective and usable on todays web size.

HITS on Graph

Authority	**Hub**	**Normalisation**
$A(p) = \sum_{v \in S \mid v \to p} H(v)$	$H(p) = \sum_{u \varepsilon S \mid u -> p} A(u)$	$a = \frac{a}{\sqrt{\sum_{i=1}^{n} i^2}}$

Initial values

	a	**b**	**c**	**d**
A	1	1	1	1
H	1	1	1	1

First iteration:

$A(a) = H(b) = 1$

$A(b) = H(a) = 1$

$A(c) = H(a) + H(b) = 2$

$A(d) = H(a) + H(c) = 2$

$A(a) = Norm(a) = \frac{1}{\sqrt{1^2 + 1^2 + 2^2 + 2^2}} = \frac{1}{3.16}$

$A(b) = Norm(b) = \frac{1}{\sqrt{1^2 + 1^2 + 2^2 + 2^2}} = \frac{1}{3.16}$

$A(c) = Norm(c) = \frac{2}{\sqrt{1^2 + 1^2 + 2^2 + 2^2}} = \frac{2}{3.16}$

$A(d) = Norm(d) = \frac{2}{\sqrt{1^2 + 1^2 + 2^2 + 2^2}} = \frac{2}{3.16}$

$H(a) = A(b) + A(c) + A(d) = \frac{1}{3.16} + \frac{2}{3.16} + \frac{2}{3.16} = 1.58$

$H(b) = A(a) + A(c) = \frac{1}{3.16} + \frac{2}{3.16} = 0.95$

$H(c) = A(d) = \frac{2}{3.16}$

$H(d) = 0$

$H(a) = Norm(a) = \frac{1.58}{\sqrt{1.58^2 + 0.95^2 + 0.63^2}} = 0.81$

$H(b) = Norm(b) = \frac{0.95}{\sqrt{1.58^2 + 0.95^2 + 0.63^2}} = 0.49$

$H(c) = Norm(c) = \frac{0.63}{\sqrt{1.58^2 + 0.95^2 + 0.63^2}} = 0.32$

$H(d) = Norm(d) = \frac{0}{\sqrt{1.58^2 + 0.95^2 + 0.63^2}} = 0$

Second iteration:

$A(a) = H(b) = 0.49$

$A(b) = H(a) = 0.81$

$A(c) = H(a) + H(b) = 1.3$

$$A(d) = H(a) + H(c) = 1.13$$

$$A(a) = Norm(a) = \frac{0.49}{\sqrt{0.49^2 + 0.81^2 + 1.3^2 + 1.13^2}} = 0.25$$

$$A(b) = Norm(b) = \frac{0.81}{\sqrt{0.49^2 + 0.81^2 + 1.3^2 + 1.13^2}} = 0.41$$

$$A(c) = Norm(c) = \frac{1.3}{\sqrt{0.49^2 + 0.81^2 + 1.3^2 + 1.13^2}} = 0.66$$

$$A(d) = Norm(d) = \frac{1.13}{\sqrt{0.49^2 + 0.81^2 + 1.3^2 + 1.13^2}} = 0.58$$

$$H(a) = A(b) + A(c) + A(d) = 0.41 + 0.66 + 0.58 = 1.65$$

$$H(b) = A(a) + A(c) = 0.25 + 0.66 = 0.91$$

$$H(c) = A(d) = 0.58$$

$$H(d) = 0$$

$$H(a) = Norm(a) = \frac{1.65}{\sqrt{1.65^2 + 0.91^2 + 0.58^2}} = 0.84$$

$$H(b) = Norm(b) = \frac{0.91}{\sqrt{1.65^2 + 0.91^2 + 0.58^2}} = 0.46$$

$$H(c) = Norm(c) = \frac{0.58}{\sqrt{1.65^2 + 0.91^2 + 0.58^2}} = 0.29$$

$$H(d) = Norm(d) = \frac{0}{\sqrt{1.65^2 + 0.91^2 + 0.58^2}} = 0$$

Third iteration:

$$A(a) = H(b) = 0.46$$

$$A(b) = H(a) = 0.84$$

$$A(c) = H(a) + H(b) = 1.3$$

$$A(d) = H(a) + H(c) = 1.13$$

$$A(a) = Norm(a) = \frac{0.46}{\sqrt{0.46^2 + 0.84^2 + 1.3^2 + 1.13^2}} = 0.23$$

$$A(b) = Norm(b) = \frac{0.84}{\sqrt{0.46^2 + 0.84^2 + 1.3^2 + 1.13^2}} = 0.43$$

$$A(c) = Norm(c) = \frac{1.3}{\sqrt{0.46^2 + 0.84^2 + 1.3^2 + 1.13^2}} = 0.66$$

$$A(d) = Norm(d) = \frac{1.13}{\sqrt{0.46^2 + 0.84^2 + 1.3^2 + 1.13^2}} = 0.57$$

$$H(a) = A(b) + A(c) + A(d) = 0.43 + 0.66 + 0.57 = 1.66$$

$$H(b) = A(a) + A(c) = 0.23 + 0.66 = 0.89$$

$$H(c) = A(d) = 0.57$$

$$H(d) = 0$$

$$H(a) = Norm(a) = \frac{1.66}{\sqrt{1.66^2 + 0.89^2 + 0.57^2}} = 0.84$$

$$H(b) = Norm(b) = \frac{0.89}{\sqrt{1.66^2 + 0.89^2 + 0.57^2}} = 0.45$$

$$H(c) = Norm(c) = \frac{0.57}{\sqrt{1.66^2 + 0.89^2 + 0.57^2}} = 0.29$$

$$H(d) = Norm(d) = \frac{0}{\sqrt{1.66^2 + 0.89^2 + 0.57^2}} = 0$$

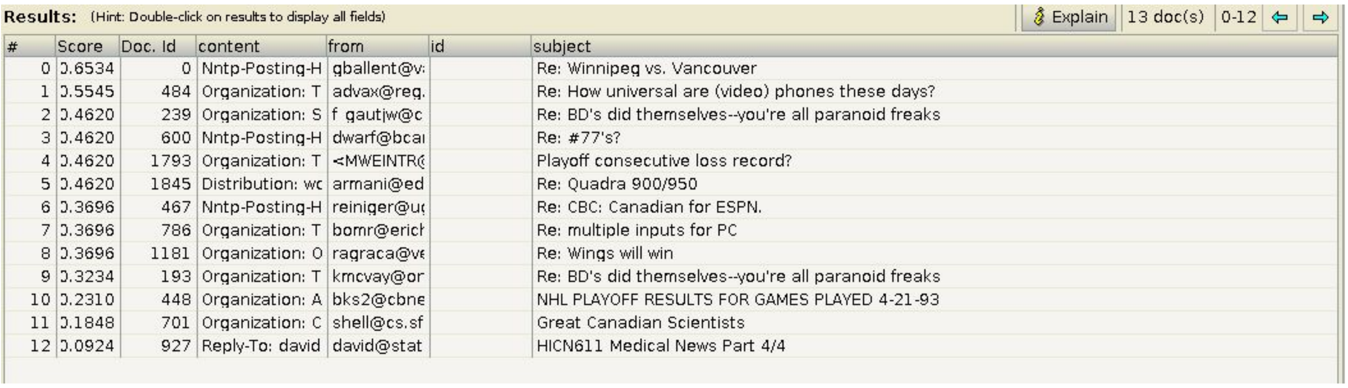
Task 2: Structured Indexing and Retrieval in Lucene

Subtask A:

```
doc.add(new StringField("id", newsDocument.getId(),Store.NO));
doc.add(new TextField("from",newsDocument.getFrom(), Store.YES));
doc.add(new TextField("subject",newsDocument.getSubject(), Store.YES));
doc.add(new TextField("content",newsDocument.getContent(), Store.YES));
```

Subtask B:

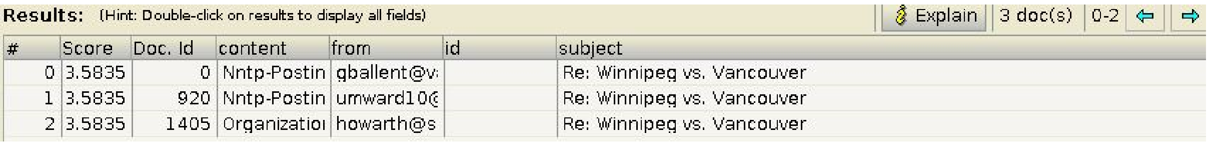
On search content:



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

#	Score	Doc. Id	content	from	id	subject
0	0.6534	0	Nntp-Posting-H	gballent@v		Re: Winnipeg vs. Vancouver
1	0.5545	484	Organization: T	advax@reg.		Re: How universal are (video) phones these days?
2	0.4620	239	Organization: S	f gautjw@c		Re: BD's did themselves--you're all paranoid freaks
3	0.4620	600	Nntp-Posting-H	dwarf@bca		Re: #77's?
4	0.4620	1793	Organization: T	<MWEINTR		Playoff consecutive loss record?
5	0.4620	1845	Distribution: wc	armani@ed		Re: Quadra 900/950
6	0.3696	467	Nntp-Posting-H	reiniger@u		Re: CBC: Canadian for ESPN.
7	0.3696	786	Organization: T	bomr@eric		Re: multiple inputs for PC
8	0.3696	1181	Organization: O	ragraca@ve		Re: Wings will win
9	0.3234	193	Organization: T	kmcvay@or		Re: BD's did themselves--you're all paranoid freaks
10	0.2310	448	Organization: A	bks2@cbne		NHL PLAYOFF RESULTS FOR GAMES PLAYED 4-21-93
11	0.1848	701	Organization: C	shell@cs.sf		Great Canadian Scientists
12	0.0924	927	Reply-To: david	david@stat		HICN611 Medical News Part 4/4

On search subject:



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

#	Score	Doc. Id	content	from	id	subject
0	3.5835	0	Nntp-Postin	gballent@v		Re: Winnipeg vs. Vancouver
1	3.5835	920	Nntp-Postin	umward10		Re: Winnipeg vs. Vancouver
2	3.5835	1405	Organizatio	howarth@s		Re: Winnipeg vs. Vancouver

From and ID returned no hits.

First, Lucene indexes the documents with the code from Subtask A.
Then, Luke parses our query [field]:Vancouver and returns any documents matching the query and shows their respective tf-idf scores.