

Report
IR Homework - 4
Student: Lohith Paripati (paripati@usc.edu)
Dataset: NBCNews

Steps involved in building search engine -

1. Downloaded Solr 6.5.0 from [Apache Website](#)
2. Extracted the downloaded file to the Documents directory
3. Opened terminal
4. Changed directory to Solr 6.5.0 setup in documents folder
5. In terminal,
 - a. entered *bin/solr start* command to start solr
 - b. entered *bin/solr create -c LohithSearchEngineV4* command to create a core named LohithSearchEngineV4
6. Went to created *LohithSearchEngineV4* folder directory in solr, navigated to conf folder and opened managed-schema.xml file. Edited it to such a way that values are as below. (True/False values should match) and then Saved the file.

```
<field name="id" type="string" indexed="true" stored="true" required="true" multiValued="false" />
<field name="_version_" type="long" indexed="true" stored="true"/>
<field name="_root_" type="string" indexed="true" stored="false"/>
<field name="_text_" type="text_general" indexed="true" stored="false" multiValued="true"/>
<copyField source="*" dest="_text_" />
```

7. Downloaded NBCnews data given [here](#) (Note my id ends with 86 so I get NBCnews, same as HW2)
8. Extracted the folder and renamed html files folder as *crawl_data_NBCNews*
9. In terminal,

I entered below command -

```
bin/post -c myexample -filetypes html /users/Documents/crawl_data_NBCNews/
```

This has made solr index all the crawled data (downloaded - data set provided)
10. After indexing completes, terminal has shown me the below result
11. Opened Solr at <http://localhost:8983/solr/#/LohithSearchEngineV4> and it looks like below on selecting *myexample* cluster

The screenshot shows the Solr Admin UI for the LohithSearchEngineV4 instance. The left sidebar contains navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, and a dropdown menu for LohithSearchEn... which includes Overview, Analysis, DataImport, Documents, Files, Ping, Plugins / Stats, Query, and Replication. The main content area is divided into three sections: Statistics, Instance, and Replication (Master). The Statistics section shows: Last Modified: about an hour ago, Num Docs: 19360, Max Doc: 19360, Heap Memory: -1, Usage: Deleted Docs: 0, Version: 223, Segment Count: 15, Optimized: ✓, and Current: ✓. The Instance section shows: CWD: /Users/Lohith/Documents/solr-6.5.0/server, Instance: /Users/Lohith/Documents/solr-6.5.0/server/solr/LohithSearchEngineV4, Data: /Users/Lohith/Documents/solr-6.5.0/server/solr/LohithSearchEngineV4/data, Index: /Users/Lohith/Documents/solr-6.5.0/server/solr/LohithSearchEngineV4/data/index, and Impl: org.apache.solr.core.NRTCachingDirectoryFactory. The Replication (Master) section shows a table with columns: Version, Gen, and Size. The table has two rows: Master (Searching) with Version 1491875966040, Gen 37, and Size 770.1 MB; and Master (Replicable) with Version -, Gen -, and Size -. The Healthcheck section shows: Ping request handler is not configured with a healthcheck file. At the bottom, there are links for Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

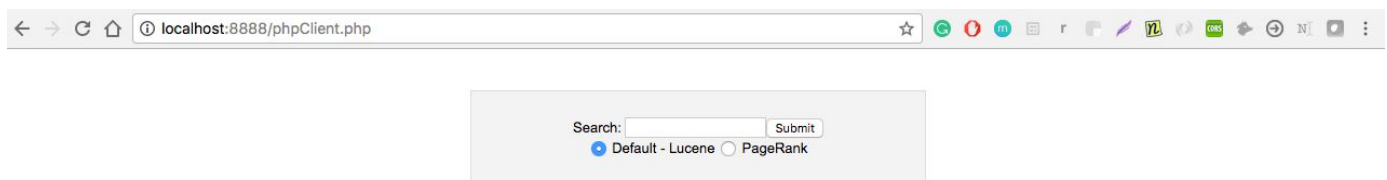
Note: 19362 files are indexed as per logs and in solr UI it shows as 19360. There is a problem with two files in provided data set and Instructor said to ignore on Piazza

12. Coded *LohithSearchEngineV4.php* file to simulate the UI of search engine and act as interface to communicate with Solr server for sending queries and showing back the retrieved results. *The file is attached as part of this submission.*

13. Installed MAMP

14. In MAMP>htdocs stored the Apache (Solr-Php client folder) and *phpClient.php* file.

15. Now, executed given 8 queries at my *http://localhost:8888/LohithSearchEngineV4.php* search engine



Attached in submission are files listing all top ten results for below mentioned 8 queries using both algorithms (1) Default - Lucene (2) PageRank. *File Names in submission for same are: **Lucene Top10 Results for 8 queries.txt**, and **PageRank Top10 Results for 8 queries.txt (2nd Deliverable)***
Queries are: Brexit, NASDAQ, NBA, Snapchat, Illegal Immigration, Donald Trump, Russia, NASA

Steps for Implementing PageRank into Search Engine -

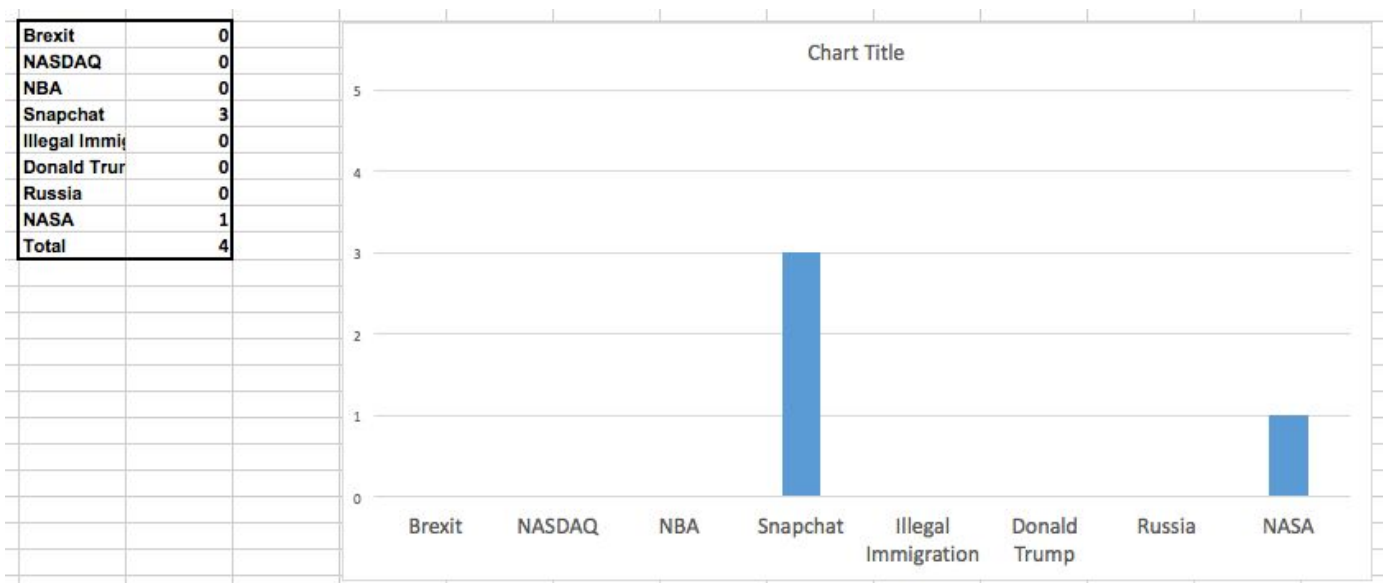
1. Downloaded jsoup jar library available [here](#).

- Wrote java code to generate edges between nodes (news articles in dataset) with help of jsoup jar in Eclipse project and generated *edgeList* file
- Downloaded NetworkX available [here](#).
- Imported NetworkX to my python interpreter in PyCharm IDE and wrote python code to generate PageRank values for links within the docs. Loaded the webgraph to networkx using `G = nx.read_edgelist("edgeList.txt", create_using=nx.DiGraph())`. Also note the parameters - `alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None`
- Copied the generated *external_PageRankFile* file into data folder of the core
- Added the following field (below lines) in the managed-schema file so it refers to the scores in *external_PageRankFile* file generated at previous step.


```
<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" />
<field name="PageRankFile" type="external" stored="false" indexed="false" />
```
- Then added eventlisteners (below lines) to solrconfig.xml file within `<query>` element.


```
<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
```
- Then reloaded the index, by going to the Solr Dashboard UI ->Core Admin and clicking on the "Reload" button.

3rd Deliverable: Overlap



For complete details about the above graph, please check *3.Overlap Details.xls* file in this submission

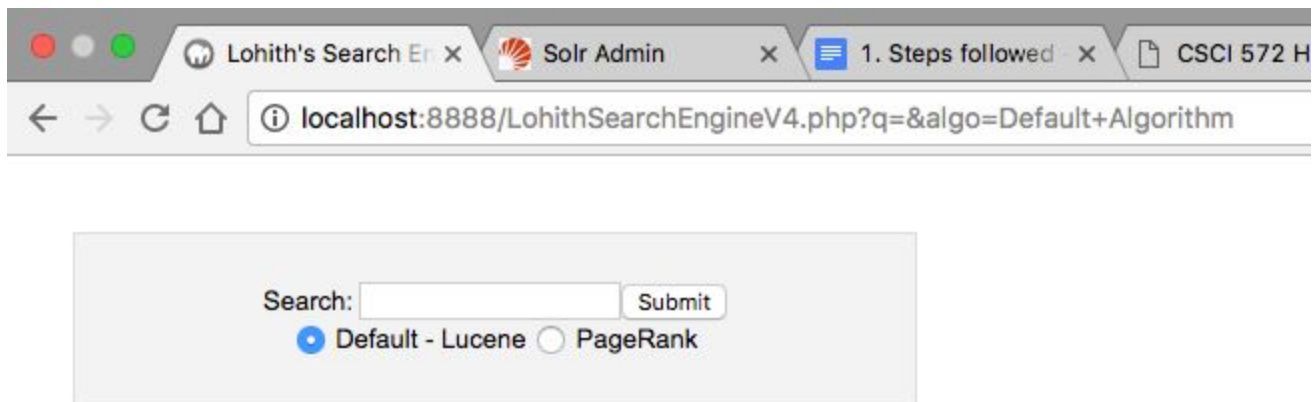
4th Question: Explanation regarding why some pages have higher PageRank values

PageRank: PageRank one of the algorithms used to determine a relevance of a web page. The key idea in PageRank is it ranks the page by its linkage i.e., how many pages point to a page and how important these pages are. The PageRank of a page depends on the PageRank metric of all pages that link to it.

As the importance of page is dependent of the number of incoming links, it ignores the fact if they come from important pages which is a disadvantage by simplified pagerank. There are 2 approaches for identifying the importance of page, Manual page and Robust surfer model. Manual page is identifying a good page and then assigning the pagerank for it whereas in Robust surfer model , it starts from any page and follows its outgoing links randomly. Requirements: Pageranks are based on limiting distribution and as the graph is strongly connected, any other node can be reached as long as it does not have spider trap and dead ends

5th grading requirement: Flow in UI

Launch Screen:



Search:

☒ Default - Lucene ☐ PageRank

<Continued below...>

Lucene Query

Search:
☒ Default ☐ Lucene ☐ PageRank

Results 1 - 10 of 233:

1. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/d7f88191-453c-4cc1-8cad-25c9f2541bd0.html
URL: <http://www.nbcnews.com/feature/college-game-plan/colleges-use-snapchat-attract-prospective-students-n570701>
Title : Colleges Use Snapchat to Attract Prospective Students - NBC News
Description : From reaching prospective students to fostering school spirit, universities are using Snapchat as a marketing tool.
2. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/af7031ce-74be-4b88-81f4-84056b26206b.html
URL: <http://www.nbcnews.com/tech/tech-news/snapchat-launches-video-sunglasses-becomes-snap-inc-n653846>
Title : Snapchat Launches Video Sunglasses and Becomes Snap Inc. - NBC News
Description : Snapchat will launch Spectacles, \$130 video sunglasses that can record and upload 10-second clips to social sites.
3. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/a10ab0ed-c53d-4b4f-8508-cfbd1577ad32.html
URL: <http://www.nbcnews.com/tech/tech-news/will-wall-street-love-snapchat-much-millennials-n715401>
Title : Snap Goes Public: Wall Street Prepares for Tech's Biggest IPO Since Facebook - NBC News
Description : Snap, the company behind Snapchat and Spectacles, filed its initial public offering today.
4. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/a4c6131e-c962-4d9b-8808-649ca9030de7.html
URL: <http://www.nbcnews.com/feature/college-game-plan/colleges-use-snapchat-attract-prospective-students-n570701>
Title : Colleges Use Snapchat to Attract Prospective Students - NBC News
Description : From reaching prospective students to fostering school spirit, universities are using Snapchat as a marketing tool.
5. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/ac902771-644d-4027-93ce-39fdb5ff8a4b.html
URL: <http://www.nbcnews.com/tech/tech-news/will-wall-street-love-snapchat-much-millennials-n715401>
Title : Snap Goes Public: Wall Street Prepares for Tech's Biggest IPO Since Facebook - NBC News
Description : Snap, the company behind Snapchat and Spectacles, filed its initial public offering today.
6. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/a431a5d1-ec3c-4e7d-9f29-faa727fa76d8.html
URL: <http://www.nbcnews.com/news/nbcblk/michelle-obama-joins-snapchat-michelleobama-n596581>
Title : Michelle Obama Joins Snapchat as 'MichelleObama' - NBC News
Description : Michelle Obama joined Snapchat on Tuesday to promote her upcoming trip to Liberia, Morocco and Spain to encourage education for girls.
7. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/e0c17f48-c22b-4399-996a-faab4ea3c087.html
URL: <http://www.nbcnews.com/news/latino/mit-snapchat-discover-team-reach-multicultural-latinos-n695386>
Title : mit, Snapchat Discover Team Up to Reach Multicultural, Latinos - NBC News
Description : If you have Snapchat on your smartphone you can now view mit articles, videos, animations and photos on the Discover feature daily.
8. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/e63b1d46-2084-4634-8b59-76c7487ab9a0.html
URL: <http://www.nbcnews.com/news/latino/mit-snapchat-discover-team-reach-multicultural-latinos-n695386>
Title : mit, Snapchat Discover Team Up to Reach Multicultural, Latinos - NBC News
Description : If you have Snapchat on your smartphone you can now view mit articles, videos, animations and photos on the Discover feature daily.
9. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/e7b0bcc8-260b-438d-8950-41ec392e8905.html
URL: <http://www.nbcnews.com/news/nbcblk/michelle-obama-joins-snapchat-michelleobama-n596581>
Title : Michelle Obama Joins Snapchat as 'MichelleObama' - NBC News
Description : Michelle Obama joined Snapchat on Tuesday to promote her upcoming trip to Liberia, Morocco and Spain to encourage education for girls.
10. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/85f874d4-1527-49d9-9386-ac088edd6177.html
URL: <http://www.nbcnews.com/tech/social-media/snapchat-speed-filter-led-georgia-car-crash-lawsuit-alleges-n563616>
Title : Snapchat 'Speed Filter' Led to Georgia Car Crash, Lawsuit Alleges - NBC News
Description : Lawyers say the speed tracker led to a car crash that resulted in a man suffering permanent brain damage.

Clicking on 1st Lucene Result



FEATURE > COLLEGE GAME PLAN

COLLEGE GAME PLAN
MAY 10 2016, 11:37 AM ET

Colleges Use Snapchat to Attract Prospective Students

by SCOTT STUMP

SHARE

- Share
- Tweet
- Email
- Print

When Alex Cosentino visited the campus of the University of South Carolina earlier this year, the 18-year-old looked beyond the brochures and tour guides for a glimpse of what the school is really like. His source? The social media platform Snapchat.

"When you go on Snapchat, you see what students are actively doing that day," Cosentino said. "When I visited South Carolina, the Clintons happened to be in town, so you saw [stories] of all these students trying to meet them."

Special section: [Get tips and advice about college at College Game Plan](#)

Checking Snapchat has become part of the daily routine for students like Cosentino, a senior at Matawan (N.J.) High School: 64% of Internet users between 18 and 24 now use Snapchat, up from just 24 percent in 2013, according to media analytics company comscore. That makes Snapchat

PageRank Query

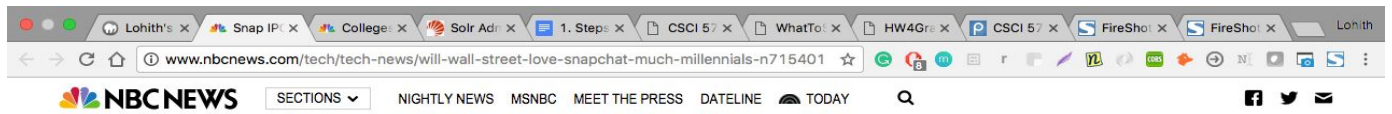
Search:

☐ Default - Lucene ☒ PageRank

Results 1 - 10 of 233:

1. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/ac902771-644d-4027-93ee-39fdb5ff8a4b.html
URL: <http://www.nbcnews.com/tech/tech-news/will-wall-street-love-snapchat-much-millennials-n715401>
Title : Snap Goes Public: Wall Street Prepares for Tech's Biggest IPO Since Facebook - NBC News
Description : Snap, the company behind Snapchat and Spectacles, filed its initial public offering today.
2. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/ad2254df-00ce-48e7-b33a-d052422547a3.html
URL: <http://www.nbcnews.com/tech/security/did-kim-kardashian-west-s-flashy-social-posts-make-her-n658631>
Title : Did Kim Kardashian West's Flashy Social Posts Make Her a Target? - NBC News
Description : The brazen heist has sent shockwaves through social media, where the reality star has relished giving fans an inside look at her glitzy lifestyle.
3. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/ad390fd5-bf35-47c8-a9a1-6675489f9e8a.html
URL: <http://www.nbcnews.com/tech/tech-news/buried-work-emails-here-s-another-reason-envy-french-n702346>
Title : New Law Could Let French Workers Ignore After-Hours Email - NBC News
Description : In the new year, many French workers could have the "right to disconnect" from work email during their off hours.
4. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/af7031ce-74be-4b88-81f4-84056b26206b.html
URL: <http://www.nbcnews.com/tech/tech-news/snapchat-launches-video-sunglasses-becomes-snap-inc-n653846>
Title : Snapchat Launches Video Sunglasses and Becomes Snap Inc. - NBC News
Description : Snapchat will launch Spectacles, \$130 video sunglasses that can record and upload 10-second clips to social sites.
5. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/af90cd0d-868c-4969-b8a6-0dd59cc57706.html
URL: <http://www.nbcnews.com/tech/tech-news/why-keep-your-data-locked-app-when-you-can-wear-n717806>
Title : Why Keep Your Data Locked in an App When You Can Wear It? - NBC News
Description : Wearing your heart on your sleeve is so 2000s. Now you can wear your digital profile on your sleeve.
6. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/83195e02-d5f9-4473-b2e7-d591b2cb8d84.html
URL: <http://www.nbcnews.com/tech/tech-news/amazon-prime-now-has-54-million-u-s-members-report-n505216>
Title : Amazon Prime Now Has 54 Million U.S. Members, Report Says - NBC News
Description : The company said late last year that it had added 3 million Prime members around the world during the third week of December 2015 alone.
7. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/84beb25d-bb51-478e-adfc-27f5a60522c6.html
URL: <http://www.nbcnews.com/tech/social-media/donald-trump-tweets-wrong-ivanka-who-then-tells-him-use-n707666>
Title : Donald Trump Tweets at Wrong Ivanka, Who Then Tells Him to Use 'More Care on Twitter' - NBC News
Description : The apparent flub gave the other Ivanka identified as an English council worker named Ivanka Majic an unexpected platform.
8. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/85f874d4-1527-49d9-9386-ac088edd6177.html
URL: <http://www.nbcnews.com/tech/social-media/snapchat-speed-filter-led-georgia-car-crash-lawsuit-alleges-n563616>
Title : Snapchat 'Speed Filter' Led to Georgia Car Crash, Lawsuit Alleges - NBC News
Description : Lawyers say the speed tracker led to a car crash that resulted in a man suffering permanent brain damage.
9. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/85fc5126-1517-437a-9671-64faef99dc2.html
URL: <http://www.nbcnews.com/tech/tech-news/anti-semitic-stunt-end-youtube-star-pewdiepie-n720696>
Title : Is Anti-Semitic Stunt the End for YouTube Star PewDiePie? - NBC News
Description : YouTube just pulled the plug on the pranky vlogger's second season show after he posted a series of videos showing anti-Semitic imagery.
10. ID: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/87ac86fe-dc60-4150-b98f-20b0ac7d3d52.html
URL: <http://www.nbcnews.com/tech/tech-news/snap-using-vending-machines-its-new-camera-glasses-n682696>
Title : Snap Is Using Vending Machines for Its New Camera Glasses - NBC News
Description : The hipsteresque frames can wirelessly upload short videos to Snapchat, and light up to let others know when recording is taking place.

Clicking on 1st PageRank Result



TECH > TECH NEWS

GADGETS INTERNET SECURITY INNOVATION MOBILE

TECH

MAR 1 2017, 5:51 PM ET

Snap IPO Set to Make Snapchat Co-Founders Into Overnight Billionaires

by ALYSSA NEWCOMB

SHARE

- Share
- Tweet
- Email
- Print

Snap Inc., the parent of social media sensation Snapchat, has priced shares of its initial public offering at \$17, putting its valuation at \$23.6 billion. The company is scheduled to begin trading on Thursday at the New York Stock Exchange, under the ticker name SNAP.

Investors are eager to "snap up" shares in the company, hoping to have a piece of what some analysts have said could become the biggest tech haul since Facebook.



From the Web

Sponsored Links



Congress Gives California Homeowners Who Owe Less Than \$300-625k A Once-In-A-Lifetime Mortgage Bailout