CSCI 572 Assignment 4: Indexing the Web Using Solr

Name: Prakhar Sethi
USC ID: 6386461387
Website Indexed: NBC_NEWS

Step 1: Solr Installation

- Downloaded the Solr by following the instructions given in the Assignment PDF IndexingwithTikaV3.pdf. Started the Solr using bin/solr start.
- Then I created a core myexample in Solr by using bin/solr create -c myexample. In the managed schema file of the core I made the changes as suggested in PDF.
- I indexed the NBC_NEWS html pages that are downloaded from the google drive.
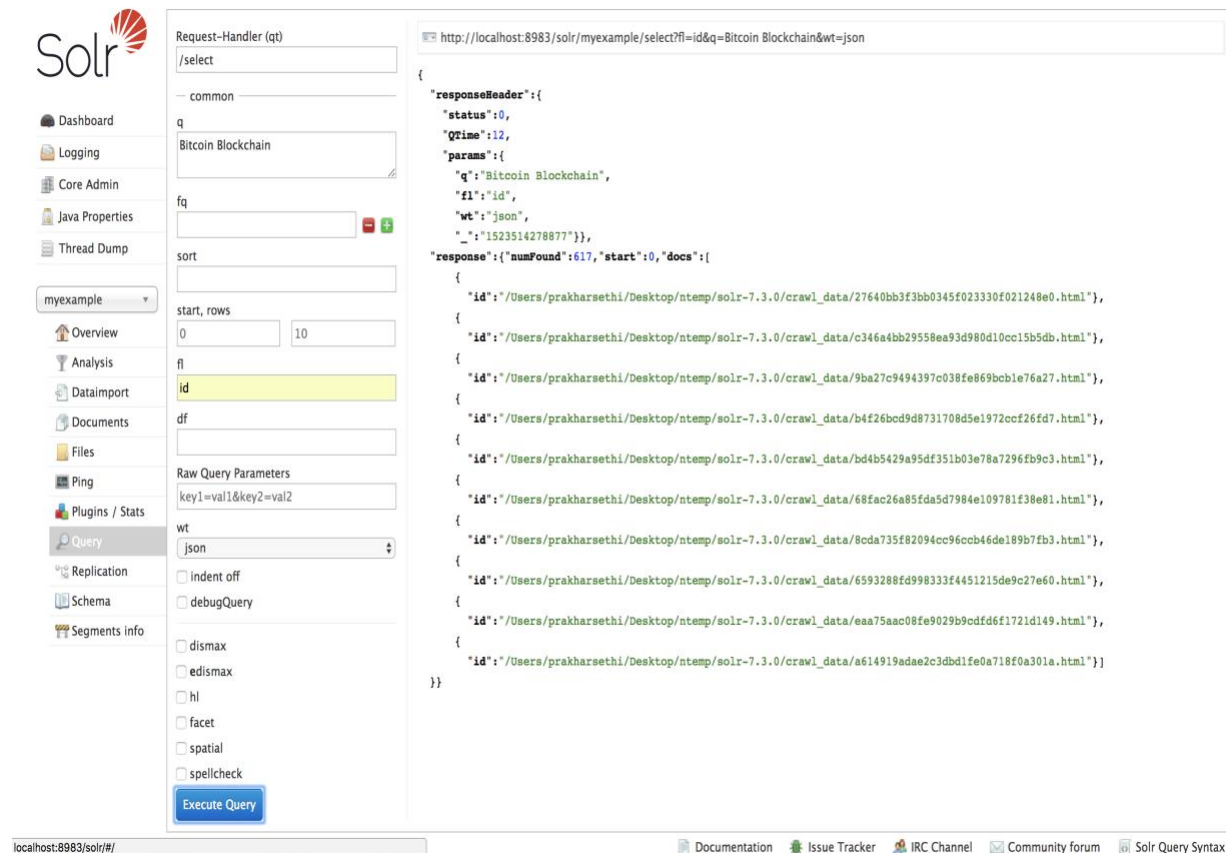- This indexing on the solr that can be viewed on the port 8983 by tying localhost:8983/solr/# on the web browser.

Step 2: Extracting Edge List using JAVA and PageRank Algorithm

- I downloaded the Jsoup.jar library using the link given in the assignment PDF and wrote the code in java called Extractor.java to extract the EdgeList.txt on the downloaded HTML data.
- Then I used Networkx graph in the pagerank algorithm that returns the pages with the corresponding page rank score. After running the python code using Networkx and following parameters:
  alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None

  generated a external_PageRankFile.txt.
- This txt file contains the page rank score of the corresponding web pages. I copied this file in the data folder of the core myexample.
- In the conf folder of the core made changes in the managed schema and solrconfig.xml file as given in the PDF. I went to the solr dashboard and reload my core after making all changes

Step 3: Verifying Results on Solr

- When I executed the Bitcoin Blockchain query on Solr admin by putting fl=id and wt=json I got the following results:

- Similarly, When I put sort=pageRankFile desc then I got the sorted result in descending order on the of their PageRank score. The output is shown below:

Solr

Request-Handler (qt)
/select

— common —

q
Bitcoin Blockchain

fq

sort
pageRankFile desc

start, rows
0   10

fl
id

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

☐ indent off
☐ debugQuery

☐ dismax
☐ edismax
☐ hl
☐ facet
☐ spatial
☐ spellcheck

Execute Query

Dashboard
Logging
Core Admin
Java Properties
Thread Dump

myexample ▾

Overview
Analysis
Dataimport
Documents
Files
Ping
Plugins / Stats
Query
Replication
Schema
Segments info

http://localhost:8983/solr/myexample/select?fl=id&q=Bitcoin Blockchain&sort=pageRankFile desc&wt=json

{
  "responseHeader":{
    "status":0,
    "QTime":8,
    "params":{
      "q":"Bitcoin Blockchain",
      "fl":"id",
      "sort":"pageRankFile desc",
      "wt":"json",
      "_":"1523514278877"}},
  "response":{"numFound":617,"start":0,"docs":[
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/77b86639d1384ac68141803ec5db27c0.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/582a22b26788b93835aab8770aa4a973.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/c74d9a1aaf8c112e26381d5a6988a355.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/76d165f6b9365bd3fff08cef66b437ba.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/1713bf741b955c27529ab0fea655a1a7.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/36d2cffc582c86dc7940542b6bdabffe.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/88992172578eb6d14b9af90eb3f45091.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/b0e8908022cde26f7a6df8c281e8e4f8.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/7ea6d6fe5260f68a910e455db96b3e0f.html"},
      {
        "id":"/Users/prakharsethi/Desktop/ntemp/solr-7.3.0/crawl_data/bb19ed47761728964946ba72b5287d85.html"}]
}}

Documentation   Issue Tracker   IRC Channel   Community forum   Solr Query Syntax

Hence, Lucene and PageRank algorithms are running.

**Question: Explanation regarding why some pages have higher PageRank values than others?**

Answer: Some pages have higher PageRank values than others as there are more number of pages pointing to that page. PageRank score is basically a vote by other pages on web that how important a page is. A link to page is counted as a vote of support. In this assignment, we created the EdgeList file and with that EdgeList a graph is created to calculate the indegree of a node/link. If a link has a lot of links pointing to it that means the indegree or PageRank score of that link will be high

When the indegree also called PagerRank score of a link is calculated then the results are displayed in a descending order that is the top result has the highest PageRank score or indegree. This is basic idea PageRank algorithm use to calculate score of a particular and this is why some pages have higher PageRank values.

Many other methods are used to calculate PageRank but the answer to this question is limited to one paragraph, so the voting criteria is the basic idea.
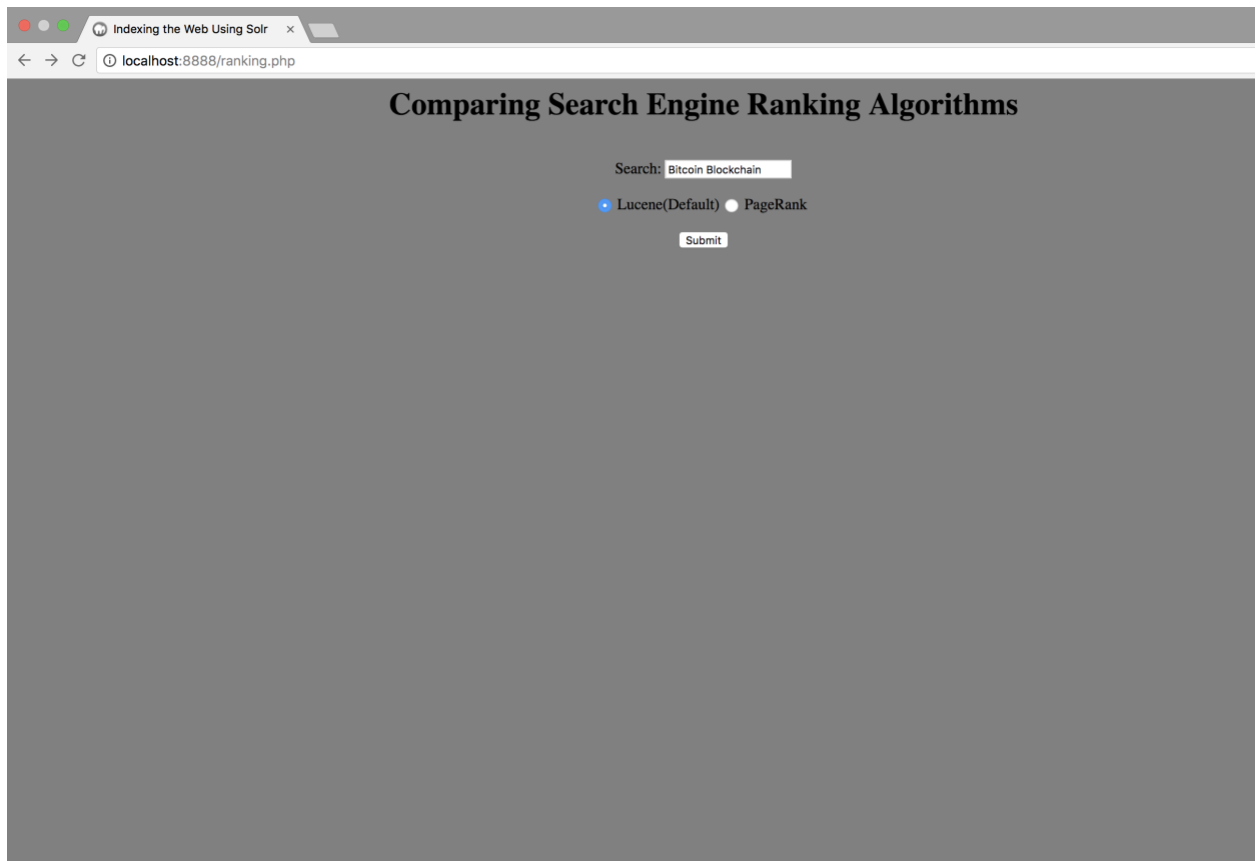
Flow of Bitcoin Blockchain Query:

1. Using HTML in ranking.php I came up with UI given in the figure below. The given page comes up when launch localhost on the browser. It contains the search bar to type the query along with two searching algorithm choice Lucene and Pagerank along with submit button to display results.
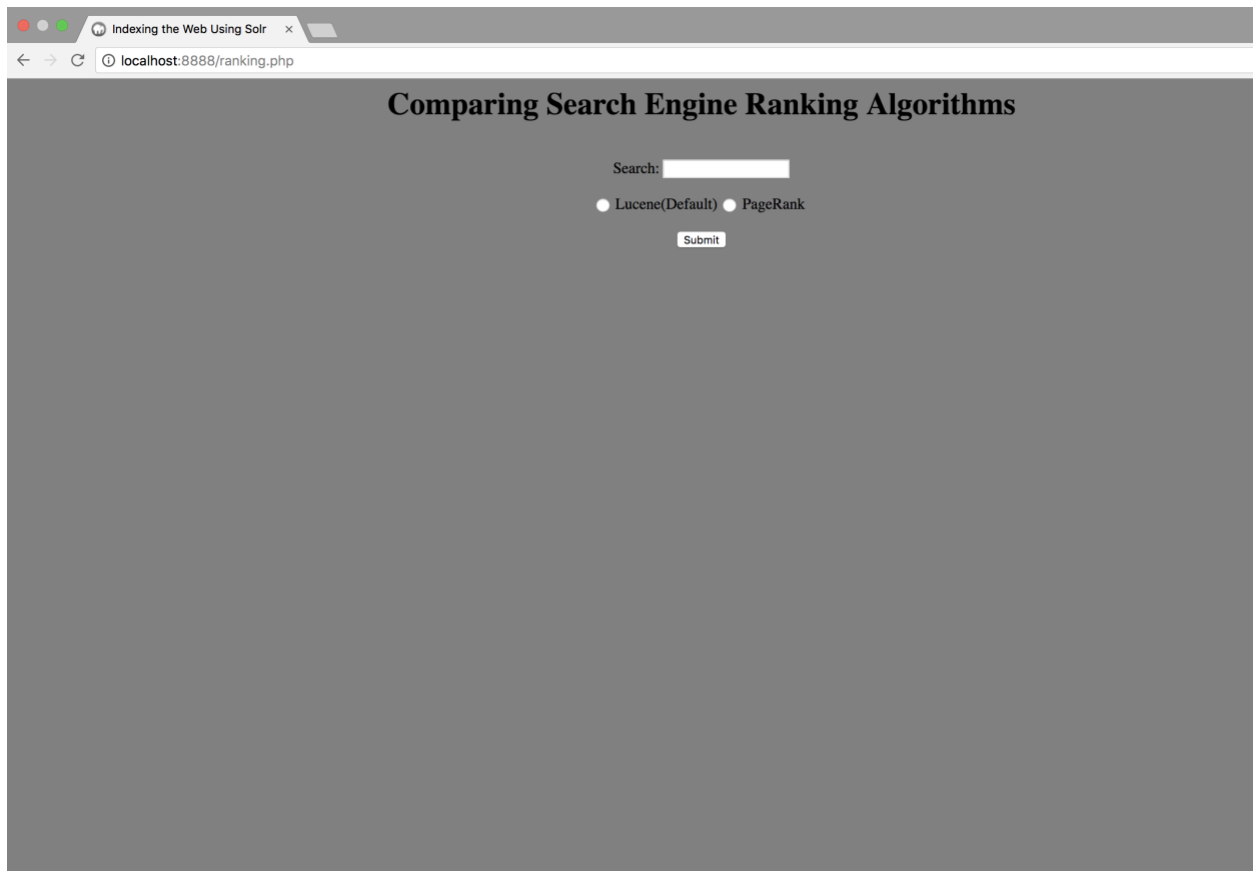
Screenshot of initial page where you enter the query in input box.

2. Then I entered a query "**Bitcoin Blockchain**" and selected the algorithm to apply on it and then submit.

Screenshot of initial page where you enter the query in input box.

3. First, I selected Lucene(Default) algorithm and it displayed the following results for Bitcoin Blockchain query.

Screenshot of page which shows the results for Lucene(Default).

4. Then, I selected Pagerank Google's algorithm and it displayed the following results for Bitcoin Blockchain query.

Screenshot of page which shows the results for PageRank

5. Screenshot of the actual webpage that opens when clicking on one of the result's link of Lucene(Default) Algorithm:

6. Screenshot of the actual webpage that opens when clicking on one of the result's link of Pagerank Algorithm:
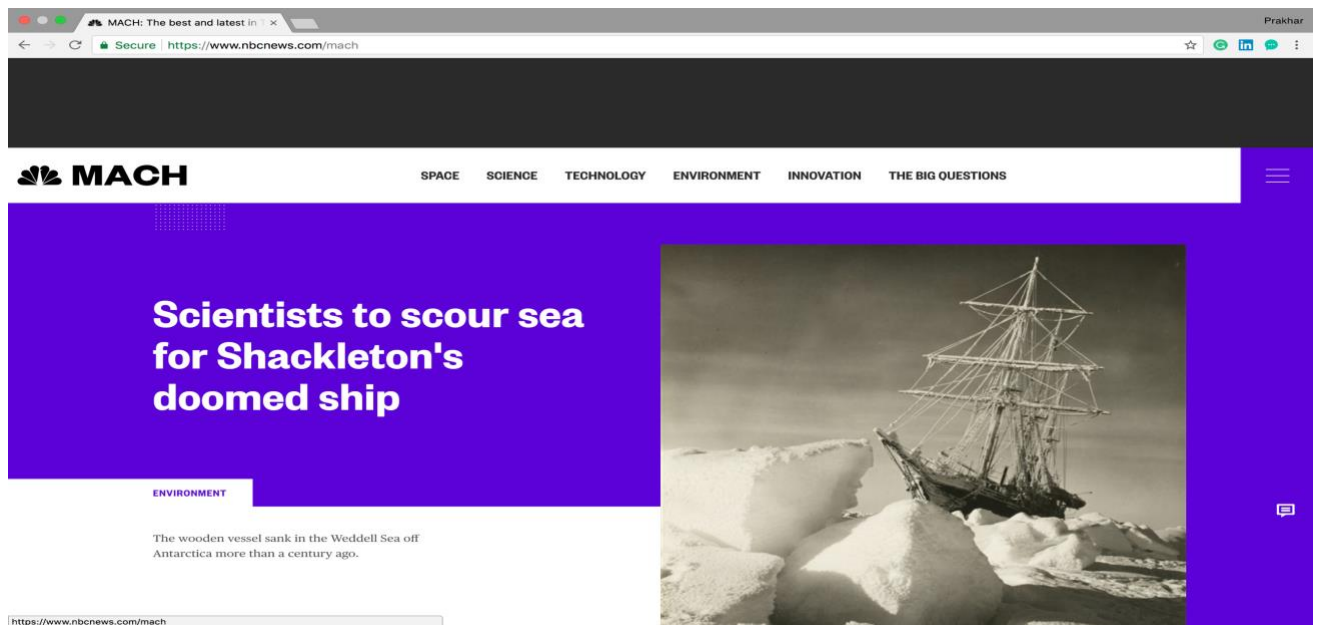
Next pages contain table containing list of URLs of the top ten results produced by both ranking methods for each query.

	1. Bitcoin Blockchain	
	Lucene(Default)	PageRank
1	https://www.nbcnews.com/mach/science/how-bitcoin-other-cryptocurrencies-could-kill-cash-ncna857036	https://www.nbcnews.com/
2	https://www.nbcnews.com/tech/innovation/bitcoin-pays-entire-wedding-its-price-soars-past-16-000-n826776	https://www.nbcnews.com/mach
3	https://www.nbcnews.com/better/business/what-bitcoin-should-you-invest-it-ncna844106	https://www.nbcnews.com/business
4	https://www.nbcnews.com/business/economy/middle-america-crazy-love-bitcoin-n789011	https://www.nbcnews.com/tech
5	https://www.nbcnews.com/mach/science/how-bitcoin-other-cryptocurrencies-could-kill-cash-ncna857036	https://www.nbcnews.com/science
6	https://www.nbcnews.com/business/business-news/hidden-dangers-buying-virtual-currency-go-beyond-simple-hack-n852706	https://www.nbcnews.com/mach/video/this-futuristic-jet-pack-that-lets-users-fly-up-to-10-000-feet-in-the-sky-1189436995946
7	https://www.nbcnews.com/tech/innovation/bitcoin-pays-entire-wedding-its-price-soars-past-16-000-n826776	https://www.nbcnews.com/mach/video/why-youtube-s-algorithm-boosted-conspiracy-theory-videos-in-the-wake-of-the-parkland-shooting-1187660355671
8	https://www.nbcnews.com/business/economy/middle-america-crazy-love-bitcoin-n789011	https://www.nbcnews.com/mach/video/from-space-dream-to-space-junk-the-vanguard-1-satellite-completes-60-years-in-space-today-1188094531644
9	https://www.nbcnews.com/better/business/what-bitcoin-should-you-invest-it-ncna844106	https://www.nbcnews.com/mach/video/pi-in-the-sky-how-a-nasa-scientist-uses-pi-to-look-for-life-beyond-earth-1185847363999
10	https://www.nbcnews.com/business/business-news/hidden-dangers-buying-virtual-currency-go-beyond-simple-hack-n852706	https://www.nbcnews.com/mach/video/astronaut-scott-kelly-s-genes-show-long-term-changes-after-a-year-in-space-1186826819794

	2. Stephen Hawking	
	Lucene(Default)	PageRank
1	https://www.nbcnews.com/mach/science/brief-history-stephen-hawking-cambridge-cultural-icon-ncsl856456	https://www.nbcnews.com/
2	https://www.nbcnews.com/news/world/british-physicist-stephen-hawking-among-world-s-greatest-minds-science-n702301	https://www.nbcnews.com/meet-the-press
3	https://www.nbcnews.com/think/opinion/stephen-hawking-s-life-was-made-possible-access-great-health-ncna856971	https://www.nbcnews.com/mach
4	https://www.nbcnews.com/science/space/i-m-atheist-stephen-hawking-god-space-travel-n210076	https://www.nbcnews.com/think
5	https://www.nbcnews.com/science/space/i-m-atheist-stephen-hawking-god-space-travel-n210076	https://www.nbcnews.com/health
6	https://www.nbcnews.com/news/world/british-physicist-stephen-hawking-among-world-s-greatest-minds-science-n702301	https://www.nbcnews.com/news/us-news
7	https://www.nbcnews.com/mach/science/stephen-hawking-says-humans-must-leave-earth-within-600-years-ncna818926	https://www.nbcnews.com/science
8	http://www.nbcnews.com/science/science-news/six-years-after-zero-g-flight-stephen-hawking-still-space-f6C9703084	https://www.nbcnews.com/news/photo
9	https://www.nbcnews.com/mach/science/what-stephen-hawking-taught-us-about-black-holes-ncna857471	https://www.nbcnews.com/video
10	https://www.nbcnews.com/mach/science/stephen-hawking-says-humans-must-leave-earth-within-600-years-ncna818926	https://www.nbcnews.com/news

	3. Falcon Heavy	
	Lucene(Default)	PageRank
1	https://www.nbcnews.com/mach/science/spacex-s-falcon-heavy-launch-gives-big-boost-elon-musk-ncna845256	https://www.nbcnews.com/
2	https://www.nbcnews.com/mach/science/watch-spacex-s-falcon-heavy-rocket-make-its-maiden-flight-ncna845096	https://www.nbcnews.com/meet-the-press
3	https://www.nbcnews.com/mach/science/watch-spacex-s-falcon-heavy-rocket-make-its-maiden-flight-ncna845096	https://www.nbcnews.com/nightly-news
4	https://www.nbcnews.com/mach/science/spacex-s-falcon-heavy-launch-gives-big-boost-elon-musk-ncna845256	https://www.nbcnews.com/news/world
5	https://www.nbcnews.com/mach/science/spacex-s-falcon-feat-may-be-just-prelude-main-event-ncna846446	https://www.nbcnews.com/news/latino
6	https://www.nbcnews.com/mach/science/here-s-what-will-happen-tesla-spacex-shot-space-ncna845566	https://www.nbcnews.com/science
7	https://www.nbcnews.com/mach/science/month-space-pictures-cosmic-penguin-playing-mars-ncsl852081	https://www.nbcnews.com/news/weather
8	https://www.nbcnews.com/mach/science/spacex-s-falcon-feat-may-be-just-prelude-main-event-ncna846446	https://www.nbcnews.com/news/photo
9	https://www.nbcnews.com/science/space	https://www.nbcnews.com/video
10	https://www.nbcnews.com/mach/science/how-spacex-aims-bring-high-speed-internet-whole-world-ncna853766	https://www.nbcnews.com/news

	4. Syrian War	
	Lucene(Default)	PageRank
1	https://www.nbcnews.com/storyline/syrias-children	https://www.nbcnews.com/
2	https://www.nbcnews.com/storyline/syrias-children/syrian-civil-war-photo-photographer-s-rescue-goes-viral-n748146	https://www.nbcnews.com/meet-the-press
3	https://www.nbcnews.com/storyline/aleppos-children	https://www.nbcnews.com/nightly-news
4	https://www.nbcnews.com/news/mideast/syria-s-civil-war-has-been-raging-7-years-what-n849851	https://www.nbcnews.com/mach
5	https://www.nbcnews.com/news/mideast/syria-s-civil-war-has-been-raging-7-years-what-n849851	https://www.nbcnews.com/politics
6	http://www.nbcnews.com/news/other/barrel-bombs-just-another-symbol-brutality-syrian-civil-war-f2D11765640	https://www.nbcnews.com/think
7	https://www.nbcnews.com/slideshow/children-among-injured-besieged-damascus-suburb-n834526	https://www.nbcnews.com/feature/nbc-out
8	https://www.nbcnews.com/slideshow/children-among-injured-besieged-damascus-suburb-n834526	https://www.nbcnews.com/business
9	https://www.nbcnews.com/storyline/syrias-suffering-families	https://www.nbcnews.com/news/world
10	https://www.nbcnews.com/news/mideast/syria-s-civil-war-has-been-raging-7-years-what-n849851	https://www.nbcnews.com/news/latino

	Lucene(Default)	PageRank
	5. LA Rams	
1	https://www.nbcnews.com/news/sports/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366	https://www.nbcnews.com/
2	https://www.nbcnews.com/news/sports/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366	https://www.nbcnews.com/meet-the-press
3	https://www.nbcnews.com/news/sports/sports-broadcasting-legend-dick-enberg-dies-82-n832076	https://www.nbcnews.com/news/nbcblk
4	https://www.nbcnews.com/slideshow/black-quarterbacks-who-ve-played-super-bowl-n513226	https://www.nbcnews.com/news/latino
5	https://www.nbcnews.com/news/sports/sports-broadcasting-legend-dick-enberg-dies-82-n832076	https://www.nbcnews.com/tech
6	https://www.nbcnews.com/pop-culture/music/bernie-casey-pro-football-player-turned-fan-favorite-actor-dies-n803291	https://www.nbcnews.com/news/us-news/nbc-affiliates-n19981
7	https://www.nbcnews.com/pop-culture/music/bernie-casey-pro-football-player-turned-fan-favorite-actor-dies-n803291	https://www.nbcnews.com/news
8	https://www.nbcnews.com/meet-the-press/meet-press-transcript-october-11-2015-n442476	https://www.nbcnews.com/better/video/how-mika-brzezinski-found-her-value-and-got-paid-what-she-s-worth-1189457987619
9	https://www.nbcnews.com/meet-the-press/meet-press-transcript-october-11-2015-n442476	https://www.nbcnews.com/better/video/how-not-to-choke-in-sports-and-other-pressure-charged-situations-1186761795616
10	https://www.nbcnews.com/politics/2016-election/la	https://www.nbcnews.com/better/video/how-to-lose-weight-and-why-it-s-so-hard-1187543619775

6. Black Panther

	Lucene(Default)	PageRank
1	https://www.nbcnews.com/news/nbcblk/black-panther-s-wakanda-sheds-light-black-excellence-n844801	https://www.nbcnews.com/
2	https://www.nbcnews.com/news/nbcblk/black-panther-sets-fandango-record-top-pre-selling-marvel-movie-n836631	https://www.nbcnews.com/nightly-news
3	https://www.nbcnews.com/news/nbcblk/black-panther-novelist-journalism-sci-fi-bring-hidden-light-n848011	https://www.nbcnews.com/mach
4	https://www.nbcnews.com/tech/internet/plot-destroy-black-panther-ratings-thwarted-facebook-n844226	https://www.nbcnews.com/think
5	https://www.nbcnews.com/pop-culture/movies/black-panther-blows-away-box-office-192-million-weekend-n849246	https://www.nbcnews.com/health
6	https://www.nbcnews.com/news/nbcblk/black-panther-black-comic-creators-spent-years-breaking-down-barriers-n848461	https://www.nbcnews.com/feature/nbc-out
7	https://www.nbcnews.com/news/nbcblk/black-panther-actress-janeshia-adams-ginyard-talks-living-stuntwoman-life-n847646	https://www.nbcnews.com/pop-culture
8	https://www.nbcnews.com/think/opinion/defeat-killmonger-black-panther-holds-lessons-israel-ncna850831	https://www.nbcnews.com/news/nbcblk
9	https://www.nbcnews.com/news/nbcblk/black-panther-novelist-journalism-sci-fi-bring-hidden-light-n848011	https://www.nbcnews.com/news/asian-america
10	https://www.nbcnews.com/news/nbcblk/stars-black-panther-discuss-hope-marvel-film-impact-n847706	https://www.nbcnews.com/tech

7. Austin Bombings

	Lucene(Default)	PageRank
1	https://www.nbcnews.com/news/us-news/police-identify-teen-killed-austin-package-bombing-investigation-grows-n856371	https://www.nbcnews.com/
2	https://www.nbcnews.com/news/us-news/deadly-austin-bombings-were-meant-send-message-police-chief-says-n857786	https://www.nbcnews.com/meet-the-press
3	https://www.nbcnews.com/news/us-news/deadly-austin-bombings-were-meant-send-message-police-chief-says-n857786	https://www.nbcnews.com/nightly-news
4	https://www.nbcnews.com/widget/video-embed/1184733251954	https://www.nbcnews.com/dateline
5	https://www.nbcnews.com/news/us-news/police-identify-teen-killed-austin-package-bombing-investigation-grows-n856371	https://www.nbcnews.com/better
6	https://www.nbcnews.com/news/us-news/austin-texas-edge-after-spate-package-bombs-kill-two-n855956	https://www.nbcnews.com/mach
7	https://www.nbcnews.com/news/us-news/austin-texas-edge-after-spate-package-bombs-kill-two-n855956	https://www.nbcnews.com/leftfield
8	http://www.nbcnews.com/	https://www.nbcnews.com/think
9	https://www.nbcnews.com/news/latino/1st-latino-poet-laureate-juan-felipe-herrera-migrant-workers-son-n372816	https://www.nbcnews.com/news/us-news
10	https://www.nbcnews.com/news/latino/california-affirmative-action-bill-sparks-fierce-debate-n87426	https://www.nbcnews.com/news/nbcblk

8. Paris Climate Deal

	Lucene(Default)	PageRank
1	https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/trump-declines-endorsing-paris-climate-change-deal-g7-summit-will-n765406	https://www.nbcnews.com/
2	https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/trump-declines-endorsing-paris-climate-change-deal-g7-summit-will-n765406	https://www.nbcnews.com/meet-the-press
3	https://www.nbcnews.com/politics/white-house/trump-pulls-u-s-out-paris-climate-agreement-n767066	https://www.nbcnews.com/dateline
4	https://www.nbcnews.com/politics/white-house/here-s-what-happens-if-trump-leaves-paris-climate-agreement-n766761	https://www.nbcnews.com/mach
5	https://www.nbcnews.com/politics/politics-news/forget-trump-u-s-storms-ahead-climate-change-never-n801826	https://www.nbcnews.com/health
6	https://www.nbcnews.com/meet-the-press/meet-press-june-4-2017-n768046	https://www.nbcnews.com/investigations
7	https://www.nbcnews.com/news/world/analysis-trump-s-america-first-foreign-policy-isolates-u-s-n833046	https://www.nbcnews.com/science
8	https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/trump-end-first-foreign-trip-pledges-unwavering-support-u-s-n765446	https://www.nbcnews.com/news/weather
9	https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/g-7-summit-leaders-brace-clash-trump-trade-climate-n764971	https://www.nbcnews.com/video
10	https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/trump-end-first-foreign-trip-pledges-unwavering-support-u-s-n765446	https://www.nbcnews.com/storyline/bill-cosby-scandal

Overlap Table:

Query Name	Number of Overlaps
Bitcoin Blockchain	0
Stephen Hawking	0
Falcon Heavy	0
Syrian War	0
LA Rams	0
Black Panther	0
Austin Bombings	1
Paris Climate Deal	0

An overlap graph showing the amount of overlap between the two ranking strategies for all queries.