**IR ASSIGNMENT 4**

Name: Rahul Shreerang Lele USC ID: 2483165273

**1. Tools used in the assignment**

1. Lucene

Lucene is a high performance, scalable, full-text search library. One adds documents to Lucene which creates an inverted index. Lucene includes flexible text analysis including tokenizers and filters. Lucene does NOT include crawlers or explicit searching of documents. Lucene is 100% Java with no dependencies. Lucene is Offered by the Apache Software Foundation.

2. Apache Solr

Solr is a full text search server based on Lucene. It communicates with Lucene via XML/HTTP, JSON interfaces. Solr has following features - flexible data schema to define types and fields, output includes highlighting matches, configurable advanced caching, index replication, Extensible Open Architecture, Plugins, Web Administration Interface. Solr is written in Java 5.

3. Eclipse IDE to run Jsoup Java Code

Download the Jsoup library jar and import the jar into your java project in eclipse. Jsoup java code is used to extract links from the downloaded web pages.

4. NetworkX library

NetworkX provides data structures for graphs (or networks) along with graph algorithms, generators and drawing tools.

**2. Steps to complete the assignment**

Step 1: Download the HTML folder and UrlToHtml\_NBCNews.csv mapping file for NBC News.

Step 2: Download and install Solr 7.2.1.

Step 3: Start Solr using bin/solr start command.

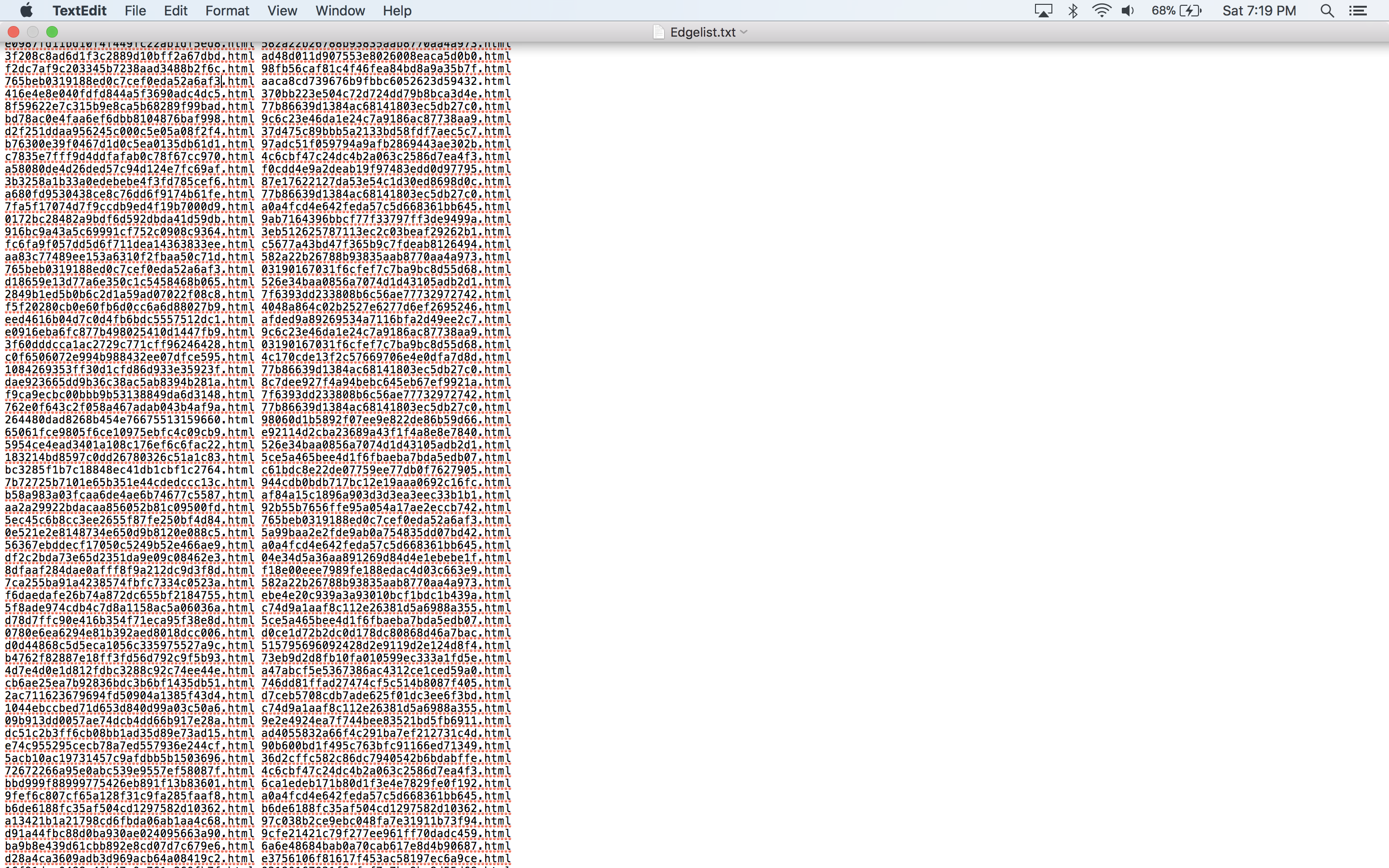
Step 4: Created a new core, myexample, using the command - bin/solr create\_core –c myexample

Step 5: Store the HTML files in a crawl\_data folder. Index the HTML files in this folder using the command –

bin/post –c myexample –filetypes html /Users/rahul/solr-7.2.1/example/exampledocs/crawl\_data/

Step 6: Download the Jsoup library jar and import the jar in eclipse.

Step 7: Write the Java code using Jsoup to extract the outgoing links from the HTML files in the crawl\_data folder. The output of this code is a file Edgelist.txt which contains links between two HTML pages in such a way that if there is an outgoing link from a.html to b.html then there is an edge between the two.



Step 8:

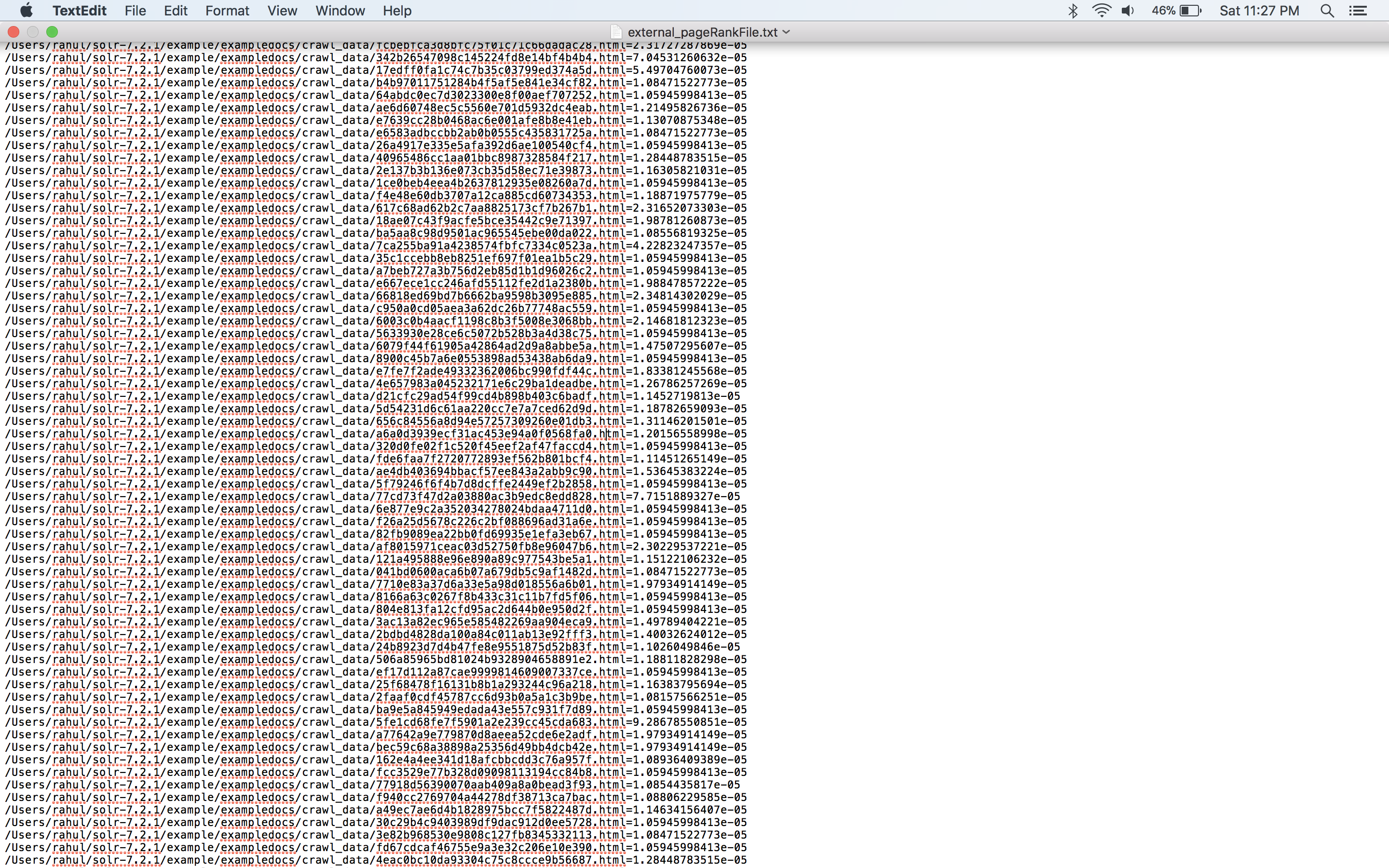
**Note**:

a. Solr uses Lucene to facilitate ranking. Lucene uses a combination of the Vector Space Model and the Boolean model to determine how relevant a given document is to a user's query. The vector space model is based on term frequency. The Boolean model is first used to narrow down the documents that need to be scored based on the use of Boolean logic in the query specification.

b. Solr permits us to change the ranking algorithm. This gives us an opportunity to use the PageRank algorithm and see how the results differ. There are several ways to manipulate the ranking of a document in Solr. Here, we explain the method which uses a field that refers to an External File that stores the PageRank scores. This external file basically contains a mapping from a key field to the field value. In order to create this External file, first carry out the PageRank process on the HTML files in the crawl\_data folder. We use NetworkX library for this.

We write a python code using NetworkX library, the python code takes edges from ‘Edgelist.txt’ file to create a NetworkX graph and returns a dictionary of graph nodes (HTML files) with corresponding pagerank scores. The following parameters are used –

alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None



Place the external\_pageRankFile.txt in the data folder of the core at Users/rahul/solr-7.2.1/server/solr/myexample/data

Step 9: Add the following fields in the managed-schema file.

<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField"/>

<field name="pageRankFile" type="external" stored="false" indexed="false"/>

Step 10: Define the following listeners within the element in the solrconfig.xml file.

<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />

<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />

Step 11: Reload the index, by going to the Solr Dashboard UI ->Core Admin and clicking on the “Reload” button.

Step 12: Run and test the queries in the Solr query view with lucene and with PageRank.

1. For testing the query with lucene, use the following parameters-

fl: id, indent: true, q: \*:\*, wt: json

1. For testing the query with PageRank, use the following parameters-

fl: id, indent: true, q: \*:\*, wt: json, sort: pageRankFile desc

Step 13: Write a PHP code which implements a search box with a submit button and two options for choosing one of the two ranking algorithms: Lucene or PageRank. After the query is submitted, the PHP code should return 10 results. For every result, there should be title, id, description and URL. The title and URL should be clickable. If the description or title is missing, N/A is displayed. If the URL is missing, the url is retrieved from the mapping csv file.

**3. PageRank**

PageRank is a way of measuring the importance of website pages.

**3.1 PageRank Formula:**

"We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also, C(A) is defined as the number of links going out of page A."

The PageRank of a page A is given as follows: PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

PR(A) is PageRank of Page A (one we want to calculate)

PR(T1) is the PageRank of Site T1 pointing to page A

C(T1) is the number of outgoing links of page T1

PR(Tn)/C(Tn) : If page A has a backlink from page “Tn” the share of the vote page A will get is “PR(Tn)/C(Tn)”

d: damping factor, generally d=0.85

**3.2 Why some pages have higher PageRanks than other pages?**

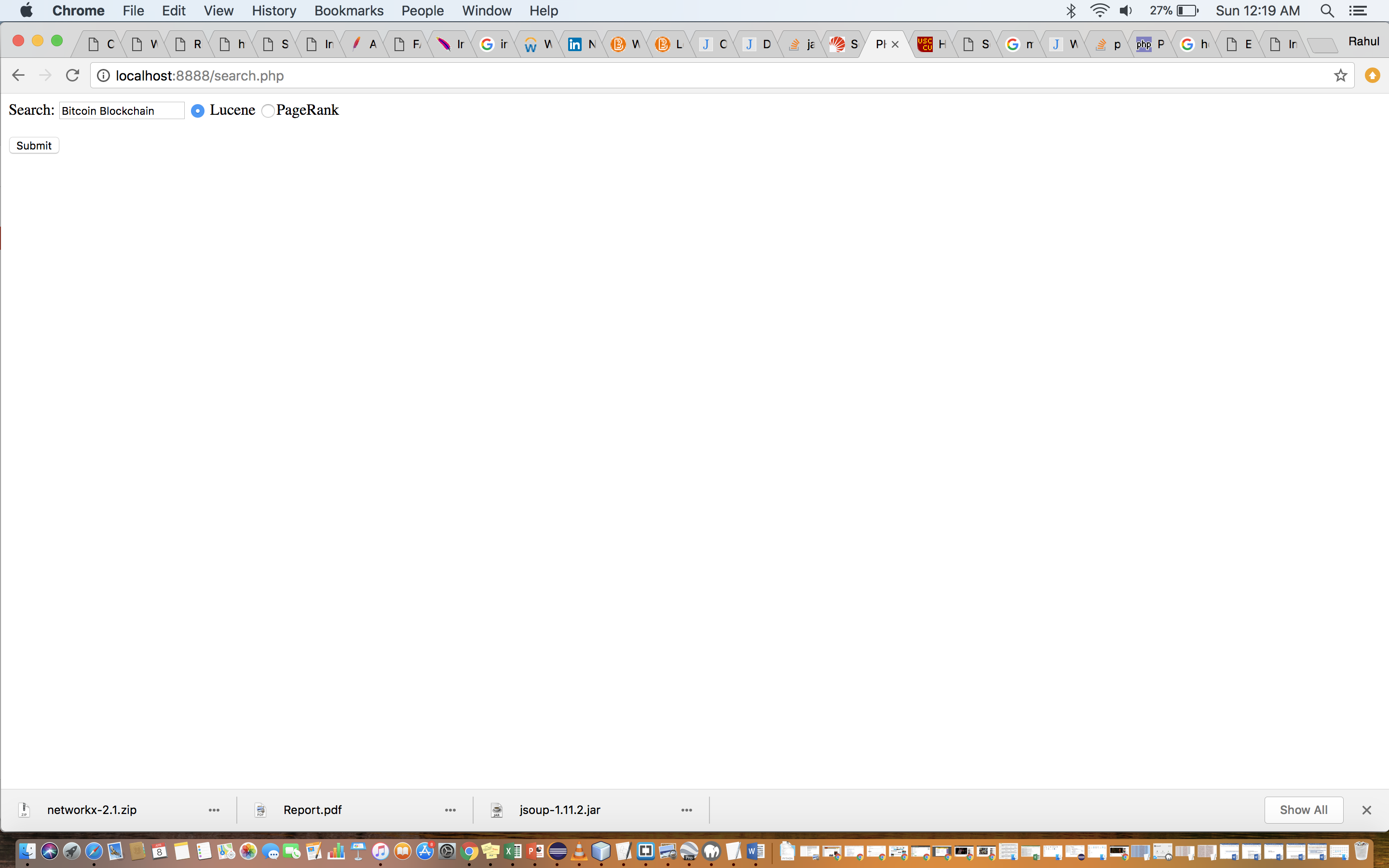
If we compare and observe the results, obtained for Lucene and pagerank for any query, it can be seen that the Lucene results are much more relevant and specific, whereas pagerank results are main pages and less specific. This is because the main pages have more incoming links and so higher pagerank, while more specific pages are deeper than the main pages and have less incoming links so lower pagerank. Pagerank focuses just on the number of incoming links and not relevance. PageRank of a page depends on the pageranks of other pages. If the pages pointing to a page A have a higher pagerank then the page A also has higher pagerank. If there a lot of incoming links from other pages having higher pagerank to page A, it means page A is important and so it will have higher pagerank value. On the other hand, if page B has lesser incoming links or if the pages linking to page B have lower pagerank, it means page B is less important and so it will have lower pagerank.

**4. Flow of the website for query: Bitcoin Blockchain**

1. Default webpage with no query

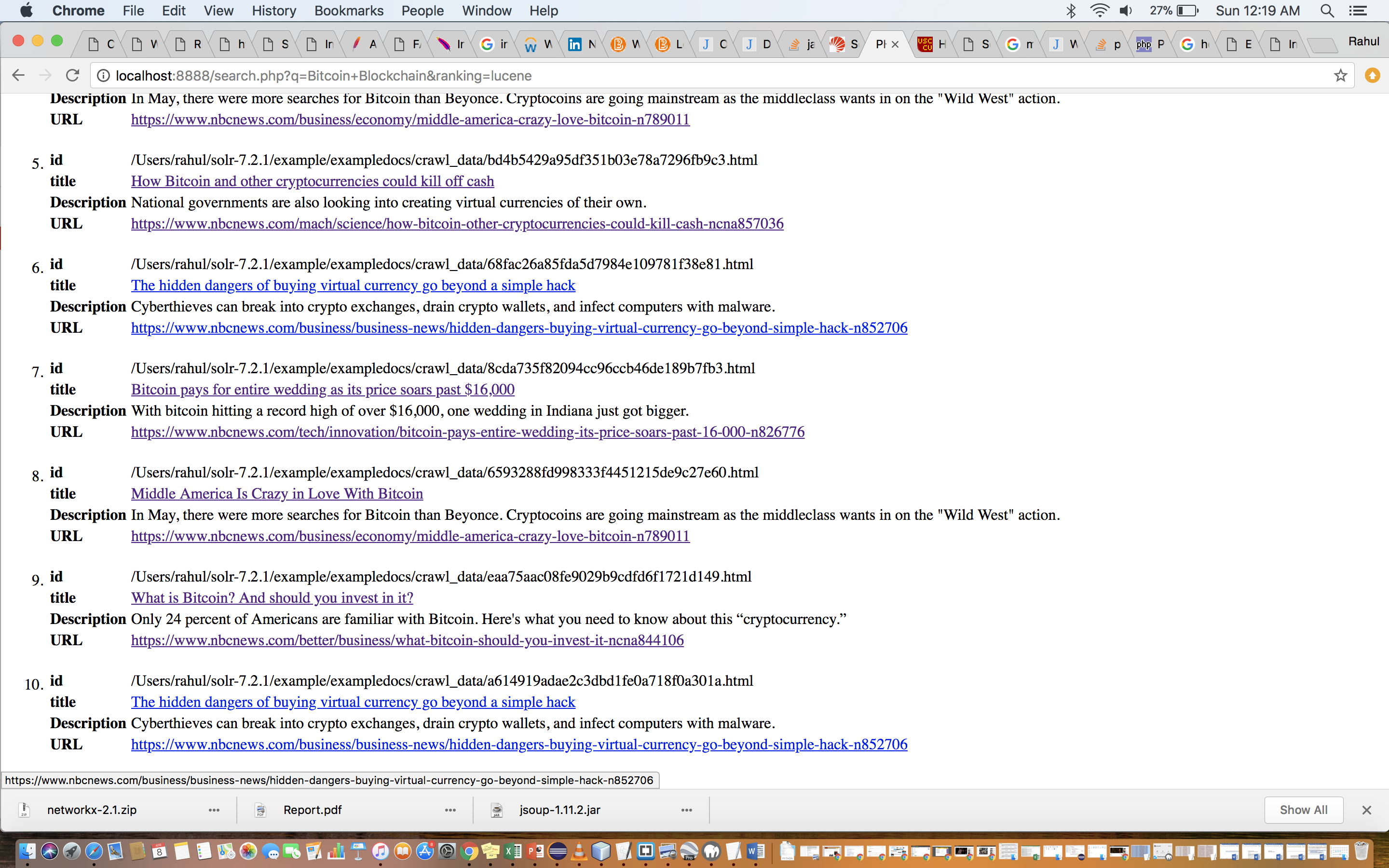


1. Query is entered and the option Lucene is selected.

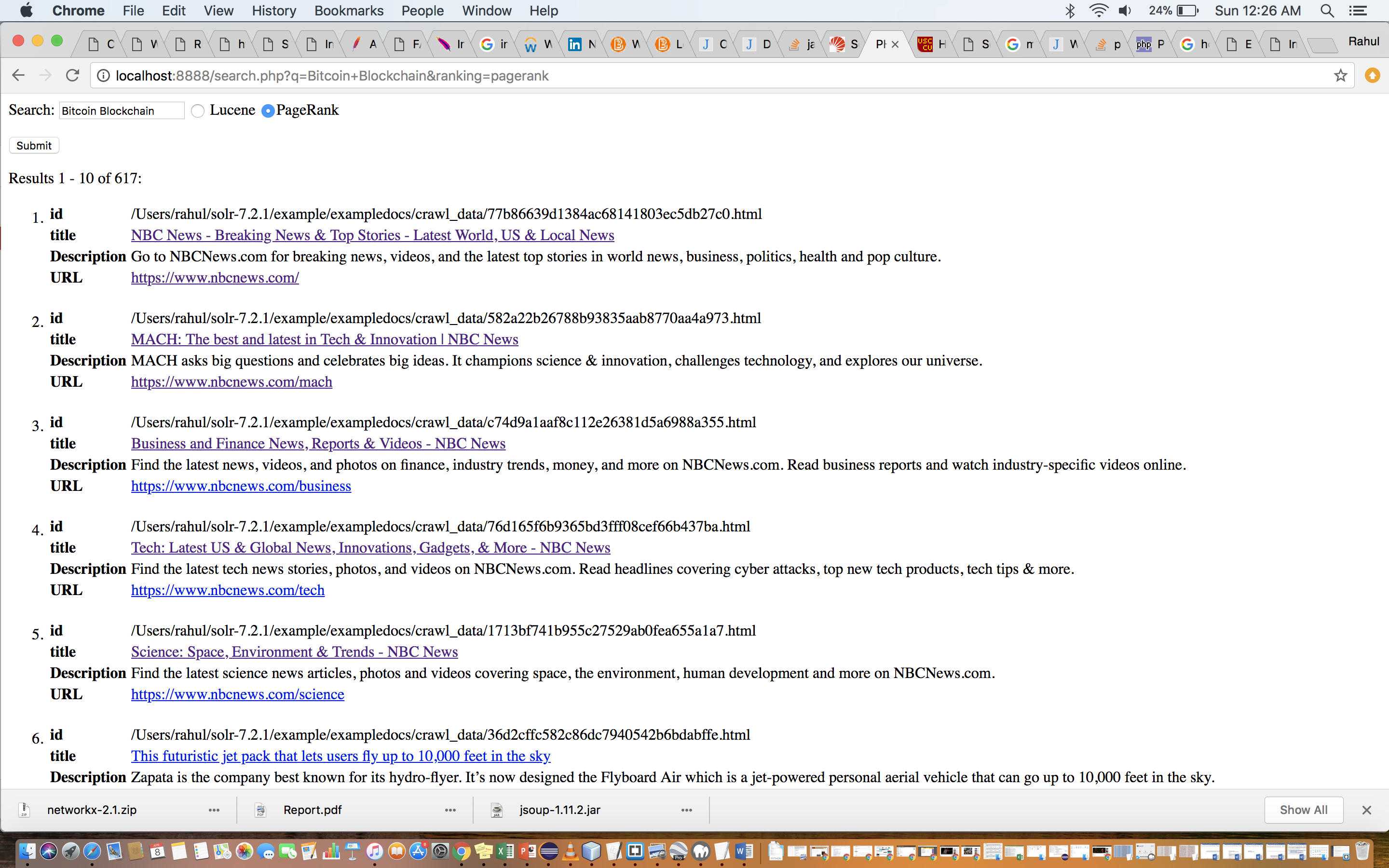


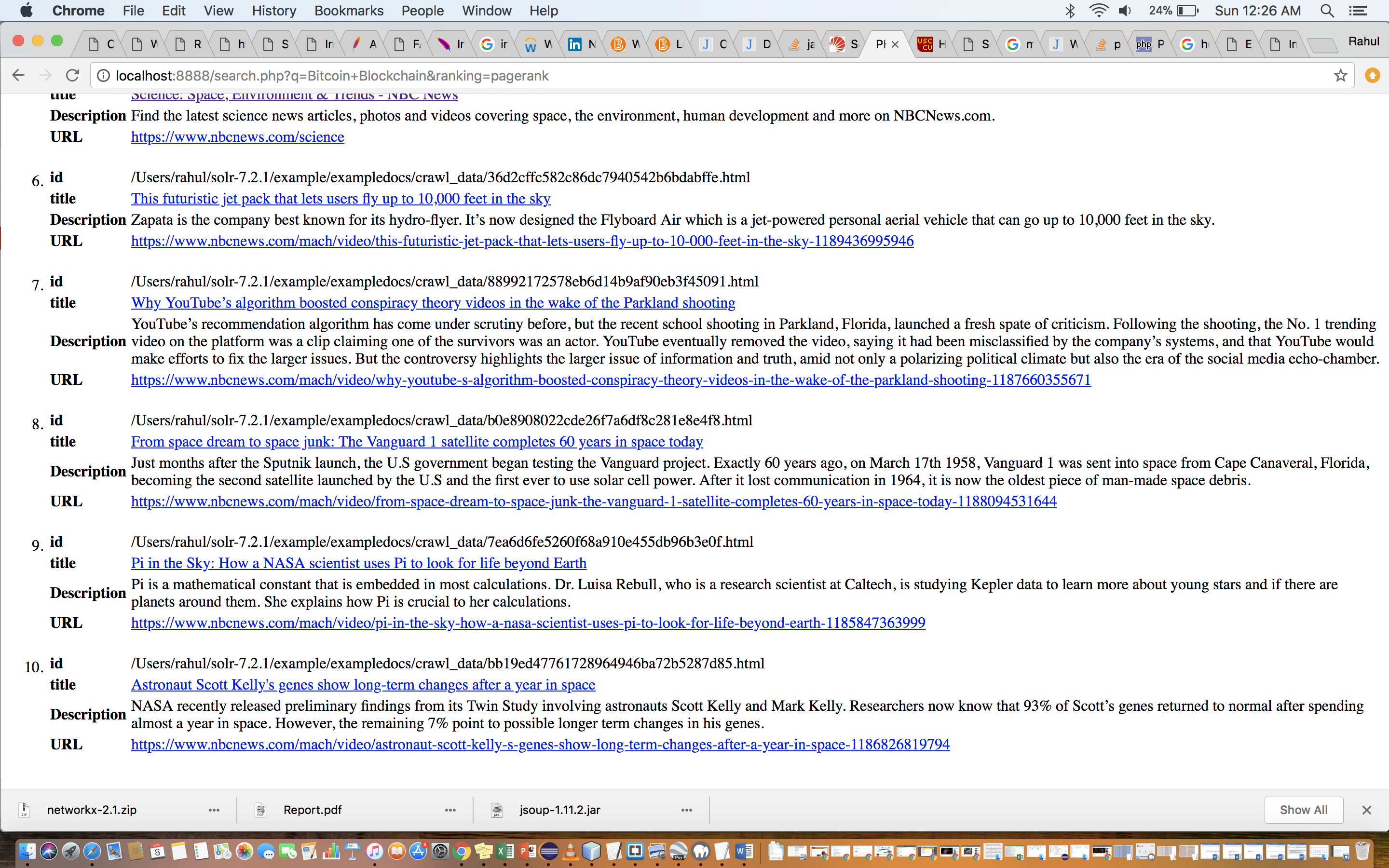
1. Results for the query: Bitcoin Blockchain with Lucene.





1. Results for the query: Bitcoin Blockchain with PageRank.





1. Bitcoin Blockchain

|  |  |
| --- | --- |
| Lucene | Pagerank |
| 1. <https://www.nbcnews.com/mach/science/how-bitcoin-other-cryptocurrencies-could-kill-cash-ncna857036> | 1. <https://www.nbcnews.com/> |
| 1. https://www.nbcnews.com/tech/innovation/bitcoin-pays-entire-wedding-its-price-soars-past-16-000-n826776 | 2. <https://www.nbcnews.com/mach> |
| 1. https://www.nbcnews.com/better/business/what-bitcoin-should-you-invest-it-ncna844106 | 3. <https://www.nbcnews.com/business> |
| 1. <https://www.nbcnews.com/business/economy/middle-america-crazy-love-bitcoin-n789011> | 4. <https://www.nbcnews.com/tech> |
| 1. <https://www.nbcnews.com/mach/science/how-bitcoin-other-cryptocurrencies-could-kill-cash-ncna857036> | 5. <https://www.nbcnews.com/science> |
| 1. <https://www.nbcnews.com/business/business-news/hidden-dangers-buying-virtual-currency-go-beyond-simple-hack-n852706> | 6. <https://www.nbcnews.com/mach/video/this-futuristic-jet-pack-that-lets-users-fly-up-to-10-000-feet-in-the-sky-1189436995946> |
| 1. <https://www.nbcnews.com/tech/innovation/bitcoin-pays-entire-wedding-its-price-soars-past-16-000-n826776> | 7. <https://www.nbcnews.com/mach/video/why-youtube-s-algorithm-boosted-conspiracy-theory-videos-in-the-wake-of-the-parkland-shooting-1187660355671> |
| 1. <https://www.nbcnews.com/business/economy/middle-america-crazy-love-bitcoin-n789011> | 8. <https://www.nbcnews.com/mach/video/from-space-dream-to-space-junk-the-vanguard-1-satellite-completes-60-years-in-space-today-1188094531644> |
| 1. <https://www.nbcnews.com/better/business/what-bitcoin-should-you-invest-it-ncna844106> | 9. <https://www.nbcnews.com/mach/video/pi-in-the-sky-how-a-nasa-scientist-uses-pi-to-look-for-life-beyond-earth-1185847363999> |
| 1. <https://www.nbcnews.com/business/business-news/hidden-dangers-buying-virtual-currency-go-beyond-simple-hack-n852706> | 10. <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 | Pagerank |
| 1. <https://www.nbcnews.com/mach/science/brief-history-stephen-hawking-cambridge-cultural-icon-ncsl856456> | 1. <https://www.nbcnews.com/> |
| 2. <https://www.nbcnews.com/news/world/british-physicist-stephen-hawking-among-world-s-greatest-minds-science-n702301> | 2. <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> | 3. <https://www.nbcnews.com/mach> |
| 4. <https://www.nbcnews.com/science/space/i-m-atheist-stephen-hawking-god-space-travel-n210076> | 4. <https://www.nbcnews.com/think> |
| 5. <https://www.nbcnews.com/science/space/i-m-atheist-stephen-hawking-god-space-travel-n210076> | 5. <https://www.nbcnews.com/health> |
| 6. <https://www.nbcnews.com/news/world/british-physicist-stephen-hawking-among-world-s-greatest-minds-science-n702301> | 6. <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> | 7. <https://www.nbcnews.com/science> |
| 8. <http://www.nbcnews.com/science/science-news/six-years-after-zero-g-flight-stephen-hawking-still-space-f6C9703084> | 8. <https://www.nbcnews.com/news/photo> |
| 9. <https://www.nbcnews.com/mach/science/what-stephen-hawking-taught-us-about-black-holes-ncna857471> | 9. <https://www.nbcnews.com/video> |
| 10. <https://www.nbcnews.com/mach/science/stephen-hawking-says-humans-must-leave-earth-within-600-years-ncna818926> | 10. <https://www.nbcnews.com/news> |

3. Falcon Heavy

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

4. Syrian War

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

5. LA Rams

|  |  |
| --- | --- |
| Lucene | Pagerank |
| 1. <https://www.nbcnews.com/news/sports/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366> | 1. <https://www.nbcnews.com/> |
| 2. <https://www.nbcnews.com/news/sports/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366> | 2. <https://www.nbcnews.com/meet-the-press> |
| 3. <https://www.nbcnews.com/news/sports/sports-broadcasting-legend-dick-enberg-dies-82-n832076> | 3. <https://www.nbcnews.com/news/nbcblk> |
| 4. <https://www.nbcnews.com/slideshow/black-quarterbacks-who-ve-played-super-bowl-n513226> | 4. <https://www.nbcnews.com/news/latino> |
| 5. <https://www.nbcnews.com/news/sports/sports-broadcasting-legend-dick-enberg-dies-82-n832076> | 5. <https://www.nbcnews.com/tech> |
| 6. <https://www.nbcnews.com/pop-culture/music/bernie-casey-pro-football-player-turned-fan-favorite-actor-dies-n803291> | 6. <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> | 7. <https://www.nbcnews.com/news> |
| 8. <https://www.nbcnews.com/meet-the-press/meet-press-transcript-october-11-2015-n442476> | 8. <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> | 9. <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> | 10. <https://www.nbcnews.com/better/video/how-to-lose-weight-and-why-it-s-so-hard-1187543619775> |

6. Black Panther

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

7. Austin Bombings

|  |  |
| --- | --- |
| Lucene | Pagerank |
| 1. <https://www.nbcnews.com/news/us-news/police-identify-teen-killed-austin-package-bombing-investigation-grows-n856371> | 1. <https://www.nbcnews.com/> |
| 2. <https://www.nbcnews.com/news/us-news/deadly-austin-bombings-were-meant-send-message-police-chief-says-n857786> | 2. <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> | 3. <https://www.nbcnews.com/nightly-news> |
| 4. <https://www.nbcnews.com/widget/video-embed/1184733251954> | 4. <https://www.nbcnews.com/dateline> |
| 5. <https://www.nbcnews.com/news/us-news/police-identify-teen-killed-austin-package-bombing-investigation-grows-n856371> | 5. <https://www.nbcnews.com/better> |
| 6. <https://www.nbcnews.com/news/us-news/austin-texas-edge-after-spate-package-bombs-kill-two-n855956> | 6. <https://www.nbcnews.com/mach> |
| 7. <https://www.nbcnews.com/news/us-news/austin-texas-edge-after-spate-package-bombs-kill-two-n855956> | 7. <https://www.nbcnews.com/leftfield> |
| 8. <http://www.nbcnews.com/> | 8. <https://www.nbcnews.com/think> |
| 9. <https://www.nbcnews.com/news/latino/celia-cruz-be-remembered-grammy-lifetime-achievement-award-n516956> | 9. <https://www.nbcnews.com/news/us-news> |
| 10. <https://www.nbcnews.com/news/latino/california-affirmative-action-bill-sparks-fierce-debate-n87426> | 10. <https://www.nbcnews.com/news/nbcblk> |

8. Paris Climate Deal

|  |  |
| --- | --- |
| Lucene | Pagerank |
| 1. <https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/trump-declines-endorsing-paris-climate-change-deal-g7-summit-will-n765406> | 1. <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> | 2. <https://www.nbcnews.com/meet-the-press> |
| 3. <https://www.nbcnews.com/politics/white-house/trump-pulls-u-s-out-paris-climate-agreement-n767066> | 3. <https://www.nbcnews.com/dateline> |
| 4. <https://www.nbcnews.com/politics/white-house/here-s-what-happens-if-trump-leaves-paris-climate-agreement-n766761> | 4. <https://www.nbcnews.com/mach> |
| 5. <https://www.nbcnews.com/politics/politics-news/forget-trump-u-s-storms-ahead-climate-change-never-n801826> | 5. <https://www.nbcnews.com/health> |
| 6. <https://www.nbcnews.com/meet-the-press/meet-press-june-4-2017-n768046> | 6. <https://www.nbcnews.com/investigations> |
| 7. <https://www.nbcnews.com/news/world/analysis-trump-s-america-first-foreign-policy-isolates-u-s-n833046> | 7. <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> | 8. <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> | 9. <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> | 10. <https://www.nbcnews.com/storyline/bill-cosby-scandal> |

**Final Analysis and Observations:**

1. The results by Lucene are more specific and relevant. The results by Pagerank are less specific and majorly irrelevant because pagerank focuses just on the number of incoming links and not relevance.
2. There is only one overlap by url found in the query “Austin Bombings”.

X-axis: queries

Y-axis: overlap count

Note: I have considered the overlap by url. There is only one overlap by url in the query “Austin Bombings”.

Lucene: 8th result: <http://www.nbcnews.com/>

Pagerank: 1st result: <https://www.nbcnews.com/>