**CSCI 572: Information Retrieval and Web Search Engines**

**Homework 4: Comparing Search Engine Ranking Algorithms**

**Name : Rahul Ethiraj**

**USC ID : 3765791028**

Website Indexed: LATimes

Step 1: Solr Installation:

* Installed Ubuntu 16 on my Windows laptop in dual boot mode.
* I downloaded solr 7.5.0 from the mentioned website and unzipped the tar.
* I started the solr using bin/solr start.
* Post which, I created a core using the command: bin/solr create -c myexample, as suggested in the document and named it ‘myexample’.
* Solr was launched in the web browser at port 8983: localhost:8983/solr/#

Difficulties faced:

* I installed java 1.8.0\_45,(Version:8), and solr was starting because of the old java version. I upgraded java 8 to java 8-1.8.0\_141, and then solr was able to come up.

Step 2: Indexing the HTML files:

* Started the solr server using bin/solr start.
* I edited the managed-schema file in the conf folder so that all the text content can be mapped properly.
* This was reflected in Solr admin page in the dropdown.

Step 3: Apache, PHP installation:

* I installed both using the below commands:

sudo apt-get install apache2

sudo apt-get install libapache2-mod-php

* Post which the files are kept in var/www/html and tested for basic configuration.

Step 4: Page Rank algorithm:

* I downloaded Jsoup from the Web and wrote a Java file List.java to find the edgelist for the websites
* Unnecessary files were deleted and only .html files are kept in a separate folder.
* Written the code to construct various data structures such as: map\*, urlFileMap and fileUrlMap to construct the edgelist.
* All the outputs are consolidated in a separate file called edgelist.txt.
* The above program was written with the help of external library: jSoup and contains web page files as vertices and edge between two indicates a link between them.

In summary:

* a) Extract outgoing links
* b) Find the file name for each outgoing link using the provided map\* files
* c) Create edges between files that have links connecting them (You can either use the

Step 5: Python program to compute PageRanks:

Downloaded and installed the NetworkX python library from <https://pypi.python.org/pypi/networkx/>

* G = nx.read\_edgelist("edgeList.txt", create\_using=nx.DiGraph())
* I used the below parameters for conversion:
* **alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None**

Step 6: Modified managed-schema as below to take care of PageRank external file:

* <fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" />
* <field name="pageRankFile" type="external" stored="false" indexed="false" />

Step 7: Same has been done for solrconfig.xml:

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

Step 8: Additional changes:

* I placed the calculated external\_pageRankFile.txt file in data folder of the created core. Post which, solr was reloaded to take care of new changes.

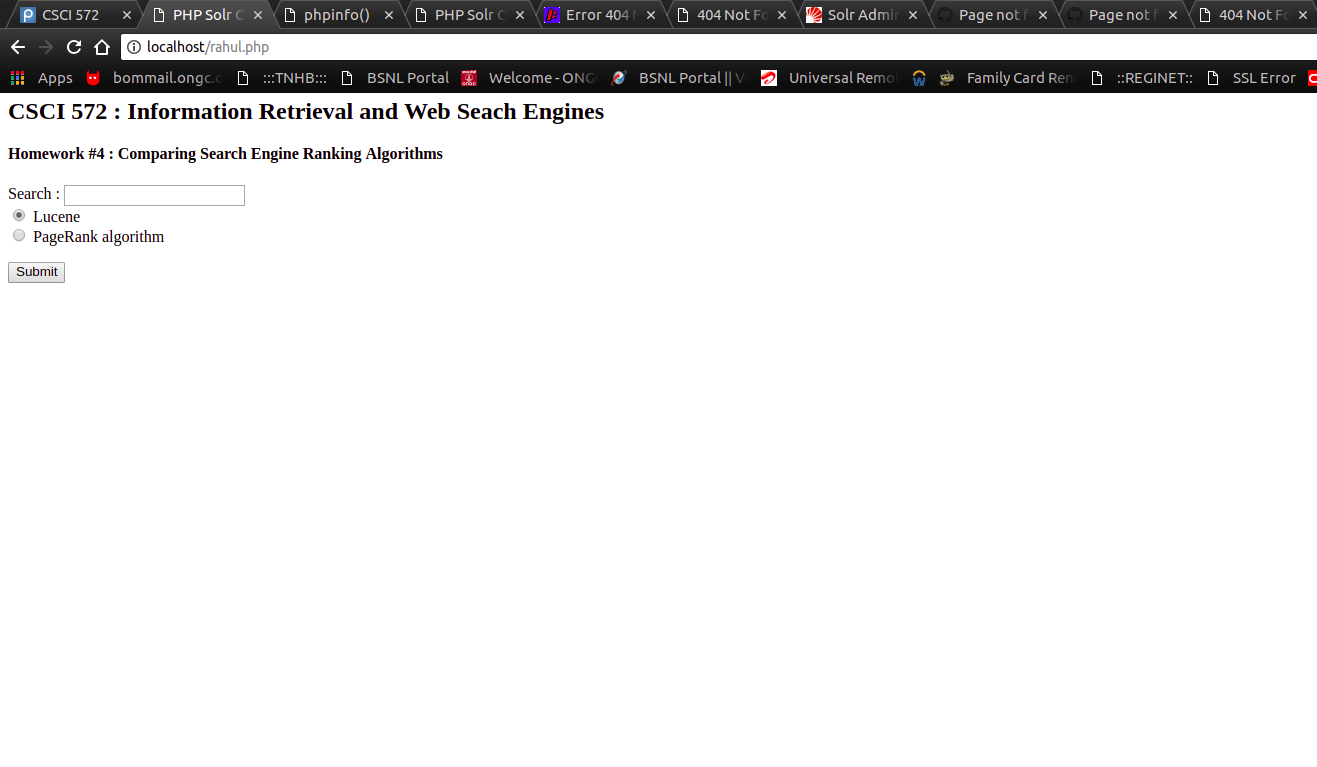
**Explanation regarding why some pages have higher PageRank values than others:**

Some pages have higher PageRank because the algorithm calculates score for each page by computing the number of incoming links to itself. In our case, this data is being fed from external\_edgelist.txt file. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

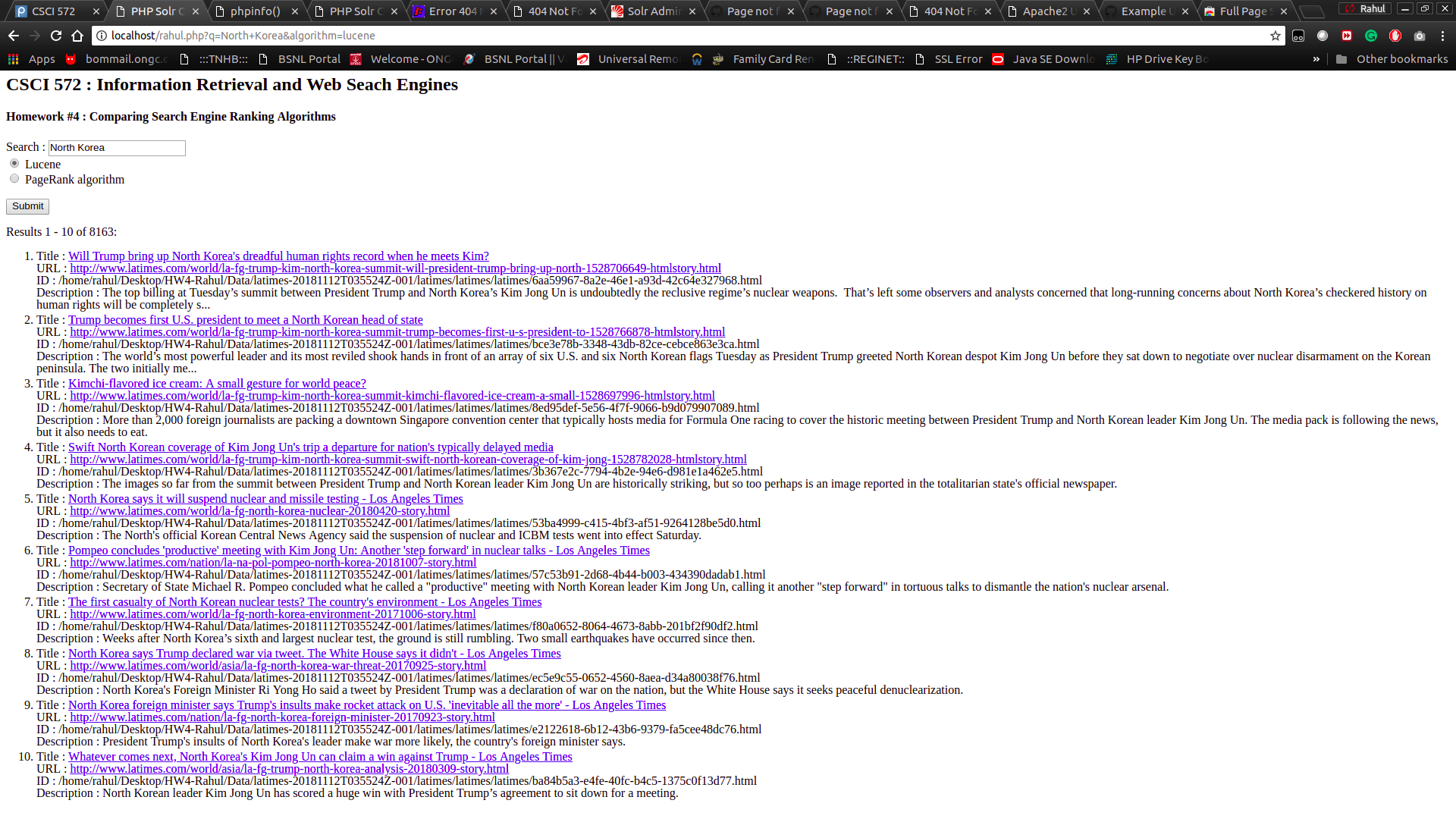
The graph is calculated using jSoup and the algorithm calculated the vector space model based on term frequency. The Boolean model is first used to narrow down the documents that need to be scored based on the use of Boolean logic in the query specification. However, all these weights are directly dependable on the number of incoming links to it.

**Flow for the query “North Korea”:**

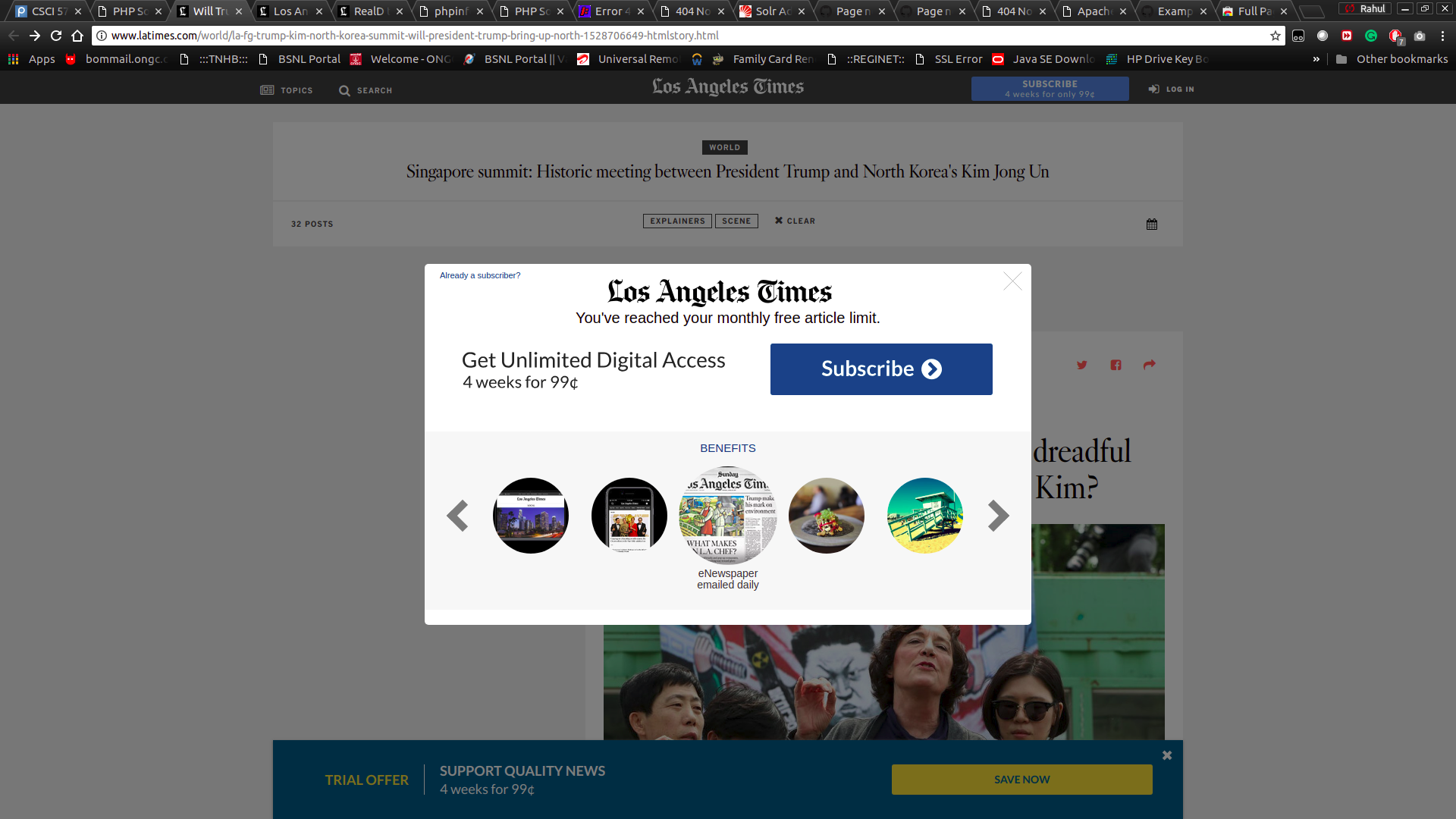
**Initial Page:**

****

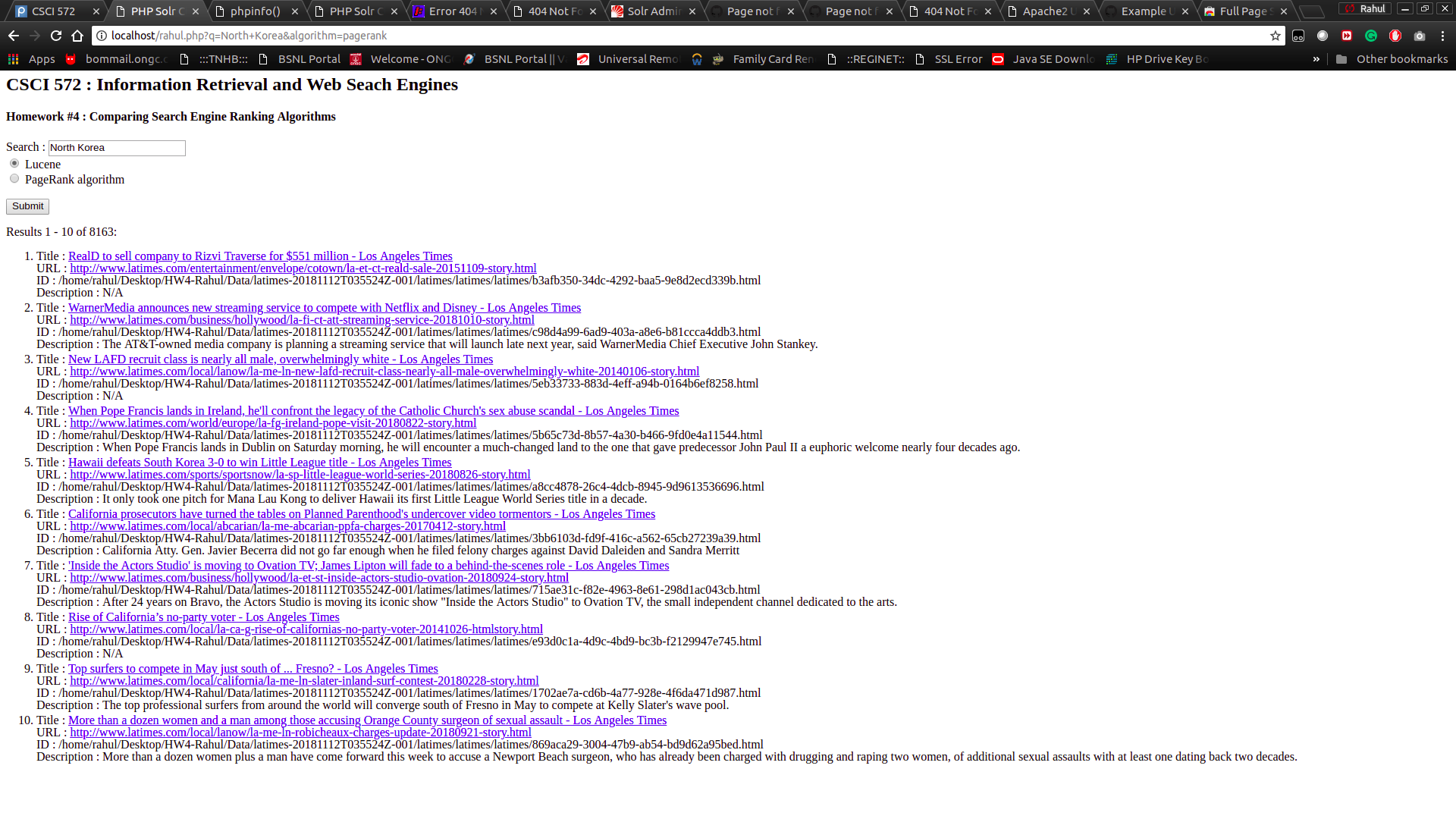
**Results for Lucene:**

****

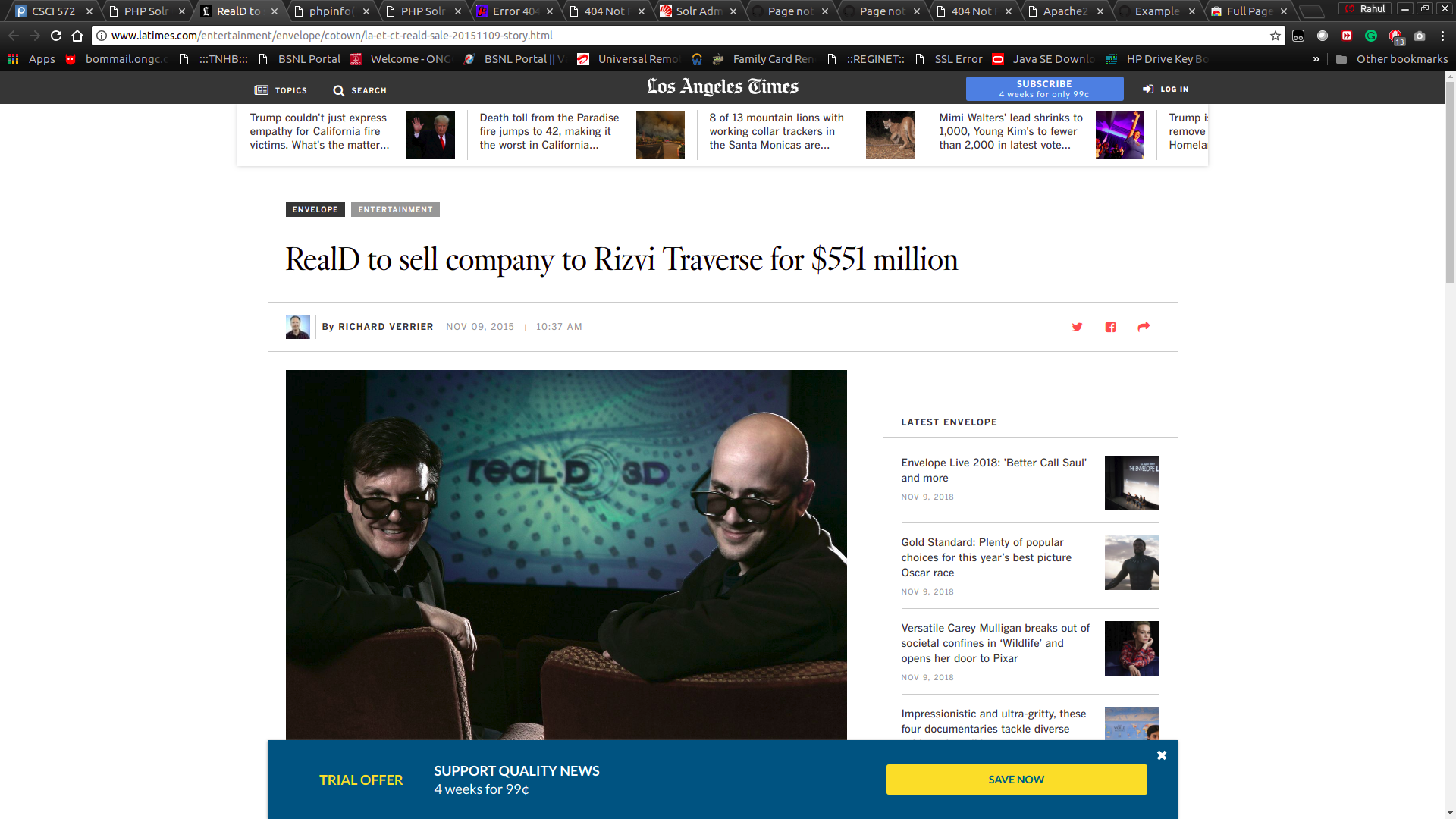
**1st Result of Lucene:**

****

**Results for PageRank:**

****

**1st result of PageRank:**

****

**Query Results**

|  |  |
| --- | --- |
| Donald Trump | |
| **LUCENE** | **PAGERANK** |
| <http://www.latimes.com/nation/politics/trailguide/la-na-trailguide-updates-trump-release-new-doctor-s-letter-1473951652-htmlstory.html> | <http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-reald-sale-20151109-story.html> |
| <http://www.latimes.com/nation/politics/trailguide/la-na-trailguide-updates-donald-trump-open-to-allowing-those-in-1473296491-htmlstory.html> | <http://www.latimes.com/fashion/la-ig-wwd-estee-lauder-sues-deciem-founder-20181011-story.html> |
| <http://www.latimes.com/politics/la-pol-ca-donald-trump-california-races-downticket-20160527-snap-htmlstory.html> | <http://www.latimes.com/sports/ducks/la-sp-ducks-what-we-learned-20181011-story.html> |
| <http://www.latimes.com/opinion/op-ed/la-oe-johnston-trump-cons-and-cheats-20181004-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-adelanto-oig-20181002-story.html> |
| <http://www.latimes.com/sports/olympics/la-sp-olympics-live-updates-the-moment-when-the-donald-trump-and-kim-1518187562-htmlstory.html> | <http://www.latimes.com/health/la-he-safety-tips-for-women-who-run-20180922-story.html> |
| [Whom do you believe, Michael Cohen or Donald Trump? Yes, that's a rhetorical question](http://localhost/Array) | <http://www.latimes.com/local/education/la-me-edu-puc-charter-closes-20180824-story.html> |
| <http://www.latimes.com/espanol/politica/la-es-dos-funerales-y-una-boda-el-rechazo-a-donald-trump-20180828-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-new-lafd-recruit-class-nearly-all-male-overwhelmingly-white-20140106-story.html> |
| <http://www.latimes.com/nation/politics/trailguide/la-na-live-updates-trailguide-1475872277-htmlstory.html> | <http://www.latimes.com/politics/essential/la-pol-ca-essential-politics-may-2018-california-lawmakers-pass-bill-designed-1535781153-htmlstory.html> |
| <http://www.latimes.com/opinion/la-ol-enter-the-fray-donald-trump-might-actually-have-to-1535040294-htmlstory.html> | <http://www.latimes.com/world/europe/la-fg-ireland-pope-visit-20180822-story.html> |
| <http://www.latimes.com/entertainment/tv/la-et-st-donald-trump-60-minutes-20161110-story.html> | <http://www.latimes.com/business/la-fi-himi-jurich-sunrun-20180318-htmlstory.html> |

|  |  |
| --- | --- |
| LA Lakers |  |
| **LUCENE** | **PAGERANK** |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/opinion/op-ed/la-oe-riley-indian-children-adoption-20181012-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-reald-sale-20151109-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-plaschke-20180924-story.html> | <http://www.latimes.com/fashion/la-ig-wwd-estee-lauder-sues-deciem-founder-20181011-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-plaschke-20180924-story.html> | <http://www.latimes.com/entertainment/arts/culture/la-ca-cm-ann-philbin-20180615-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/business/hollywood/la-fi-ct-att-streaming-service-20181010-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/sports/ducks/la-sp-ducks-what-we-learned-20181011-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-media-day-20180924-story.html> | <http://www.latimes.com/world/middleeast/la-fg-israel-temple-mount-20160422-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/opinion/op-ed/la-oe-0111-quartz-lausd-schools-20160111-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-sidebar-20180924-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-adelanto-oig-20181002-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-5-questions-20181013-story.html> | <http://www.latimes.com/health/la-he-safety-tips-for-women-who-run-20180922-story.html> |

|  |  |
| --- | --- |
| Star Wars |  |
| **LUCENE** | **PAGERANK** |
| <http://www.latimes.com/entertainment/movies/la-ca-mn-star-wars-timeline-20180516-story.html> | <http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-reald-sale-20151109-story.html> |
| <http://www.latimes.com/entertainment/movies/la-et-mn-star-wars-last-jedi-characters-20171214-htmlstory.html> | <http://www.latimes.com/business/hollywood/la-fi-ct-att-streaming-service-20181010-story.html> |
| <http://www.latimes.com/travel/themeparks/la-trb-star-wars-land-disneyland-20150817-story.html> | <http://www.latimes.com/entertainment/la-et-tiff-2018-toronto-film-festival-the-weekend-cast-describe-working-with-1536761909-htmlstory.html> |
| <http://www.latimes.com/entertainment/movies/la-et-mn-ron-howard-star-wars-han-solo-20170622-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-new-lafd-recruit-class-nearly-all-male-overwhelmingly-white-20140106-story.html> |
| <http://www.latimes.com/fashion/la-ig-wwd-rag-and-bones-star-wars-20171106-story.html> | <http://www.latimes.com/world/europe/la-fg-ireland-pope-visit-20180822-story.html> |
| <http://www.latimes.com/fashion/la-ig-wwd-rag-and-bones-star-wars-20171106-story.html> | <http://www.latimes.com/local/abcarian/la-me-abcarian-ppfa-charges-20170412-story.html> |
| <http://www.latimes.com/entertainment/herocomplex/la-et-hc-jon-favreau-star-wars-mandalorian-20181003-htmlstory.html> | <http://www.latimes.com/entertainment/la-et-entertainment-news-updates-august-one-mississippi-season-2-1503947289-htmlstory.html> |
| <http://www.latimes.com/entertainment/herocomplex/la-et-hc-comic-con-star-wars-fandom-20180723-htmlstory.html> | <http://www.latimes.com/business/hollywood/la-et-st-inside-actors-studio-ovation-20180924-story.html> |
| <http://www.latimes.com/brandpublishing/travelplus/lasvegasguide/features/la-ss-vegas-the-hot-list-burlesque-star-wars-speakeasies-20180620dto-story.html> | <http://www.latimes.com/local/la-ca-g-rise-of-californias-no-party-voter-20141026-htmlstory.html> |
| <http://www.latimes.com/entertainment/music/la-et-ms-star-wars-secrets-empire-20180116-story.html> | <http://www.latimes.com/projects/la-et-mn-film-critics-diversity/> |

|  |  |
| --- | --- |
| Lebron James |  |
| **LUCENE** | **PAGERANK** |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-lebron-james-diet-20181004-story.html> | <http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-reald-sale-20151109-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-lebron-james-20181013-story.html> | <http://www.latimes.com/business/hollywood/la-fi-ct-att-streaming-service-20181010-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-nuggets-20181002-story.html> | <http://www.latimes.com/entertainment/la-et-tiff-2018-toronto-film-festival-the-weekend-cast-describe-working-with-1536761909-htmlstory.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-nuggets-20181002-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-new-lafd-recruit-class-nearly-all-male-overwhelmingly-white-20140106-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-lebron-rondo-defense-20180930-story.html> | <http://www.latimes.com/sports/sportsnow/la-sp-little-league-world-series-20180826-story.html> |
| <http://www.latimes.com/sports/nba/la-sp-lebron-james-20160620-snap-story.html> | <http://www.latimes.com/business/hollywood/la-et-st-inside-actors-studio-ovation-20180924-story.html> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-lebron-james-20181013-story.html> | <http://www.latimes.com/projects/la-et-mn-film-critics-diversity/> |
| <http://www.latimes.com/sports/lakers/la-sp-lakers-lebron-james-20181013-story.html> | <http://www.latimes.com/business/hollywood/la-fi-ct-azoff-20181008-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/entertainment/movies/la-et-mn-box-office-night-school-smallfoot-20180930-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/sports/nfl/la-sp-nfl-week-6-farmer-20181008-story.html> |

|  |  |
| --- | --- |
| 2018 World Cup |  |
| **LUCENE** | **PAGERANK** |
| <http://www.latimes.com/sports/la-sp-42nd-ryder-cup-pictures-2018-photogallery.html> | <http://www.latimes.com/opinion/op-ed/la-oe-riley-indian-children-adoption-20181012-story.html> |
| <http://www.latimes.com/sports/la-lb-816-45174-la-me-ln-mexico-korea-soccer-viewing-20180622-htmlstory.html> | <http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-reald-sale-20151109-story.html> |
| <http://www.latimes.com/sports/la-lb-816-45174-la-me-ln-mexico-korea-soccer-viewing-20180622-htmlstory.html> | <http://www.latimes.com/fashion/la-ig-wwd-estee-lauder-sues-deciem-founder-20181011-story.html> |
| <http://www.latimes.com/sports/soccer/worldcup/la-sp-mexico-korea-1998-world-cup-2018-world-cup-20180622-story.html> | <http://www.latimes.com/entertainment/arts/culture/la-ca-cm-ann-philbin-20180615-story.html> |
| <http://www.latimes.com/espanol/deportes/la-world-cup-espanol-updates-2018-b-lgica-detiene-en-seco-a-brasil-1530907910-htmlstory.html> | <http://www.latimes.com/business/hollywood/la-fi-ct-att-streaming-service-20181010-story.html> |
| <http://www.latimes.com/espanol/deportes/la-world-cup-espanol-updates-2018-empresario-que-fue-secuestrado-en-m-xico-1530960855-htmlstory.html> | <http://www.latimes.com/sports/ducks/la-sp-ducks-what-we-learned-20181011-story.html> |
| <http://www.latimes.com/espanol/deportes/la-world-cup-espanol-updates-2018-video-brasil-anota-un-autogol-y-b-lgica-1530901739-htmlstory.html> | <http://www.latimes.com/world/middleeast/la-fg-israel-temple-mount-20160422-story.html> |
| <http://www.latimes.com/espanol/deportes/la-world-cup-espanol-updates-2018-video-grave-error-de-muslera-le-regala-1530891077-htmlstory.html> | <http://www.latimes.com/opinion/op-ed/la-oe-0111-quartz-lausd-schools-20160111-story.html> |
| <http://www.latimes.com/espanol/deportes/la-world-cup-espanol-updates-2018-video-francia-acaricia-las-semifinales-1530889005-htmlstory.html> | <http://www.latimes.com/local/lanow/la-me-ln-adelanto-oig-20181002-story.html> |
| <http://www.latimes.com/espanol/deportes/la-world-cup-espanol-updates-2018-aficionados-de-rusia-agradecen-a-1530473387-htmlstory.html> | <http://www.latimes.com/health/la-he-safety-tips-for-women-who-run-20180922-story.html> |

|  |  |
| --- | --- |
| North Korea |  |
| **LUCENE** | **PAGERANK** |
| <http://www.latimes.com/world/la-fg-trump-kim-north-korea-summit-will-president-trump-bring-up-north-1528706649-htmlstory.html> | <http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-reald-sale-20151109-story.html> |
| <http://www.latimes.com/world/la-fg-trump-kim-north-korea-summit-trump-becomes-first-u-s-president-to-1528766878-htmlstory.html> | <http://www.latimes.com/business/hollywood/la-fi-ct-att-streaming-service-20181010-story.html> |
| <http://www.latimes.com/world/la-fg-trump-kim-north-korea-summit-kimchi-flavored-ice-cream-a-small-1528697996-htmlstory.html> | <http://www.latimes.com/local/lanow/la-me-ln-new-lafd-recruit-class-nearly-all-male-overwhelmingly-white-20140106-story.html> |
| <http://www.latimes.com/world/la-fg-trump-kim-north-korea-summit-swift-north-korean-coverage-of-kim-jong-1528782028-htmlstory.html> | <http://www.latimes.com/world/europe/la-fg-ireland-pope-visit-20180822-story.html> |
| <http://www.latimes.com/world/la-fg-north-korea-nuclear-20180420-story.html> | <http://www.latimes.com/sports/sportsnow/la-sp-little-league-world-series-20180826-story.html> |
| <http://www.latimes.com/nation/la-na-pol-pompeo-north-korea-20181007-story.html> | <http://www.latimes.com/local/abcarian/la-me-abcarian-ppfa-charges-20170412-story.html> |
| <http://www.latimes.com/world/la-fg-north-korea-environment-20171006-story.html> | <http://www.latimes.com/business/hollywood/la-et-st-inside-actors-studio-ovation-20180924-story.html> |
| <http://www.latimes.com/world/asia/la-fg-north-korea-war-threat-20170925-story.html> | <http://www.latimes.com/local/la-ca-g-rise-of-californias-no-party-voter-20141026-htmlstory.html> |
| <http://www.latimes.com/nation/la-fg-north-korea-foreign-minister-20170923-story.html> | <http://www.latimes.com/local/california/la-me-ln-slater-inland-surf-contest-20180228-story.html> |
| <http://www.latimes.com/world/asia/la-fg-trump-north-korea-analysis-20180309-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-robicheaux-charges-update-20180921-story.html> |

|  |  |
| --- | --- |
| Hurricane Florence |  |
| **LUCENE** | **PAGERANK** |
| <http://www.latimes.com/nation/la-na-hurricane-florence-20180913-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-new-lafd-recruit-class-nearly-all-male-overwhelmingly-white-20140106-story.html> |
| <http://www.latimes.com/nation/la-na-hurricane-florence-20180913-story.html> | <http://www.latimes.com/world/europe/la-fg-ireland-pope-visit-20180822-story.html> |
| <http://www.latimes.com/nation/la-na-hurricane-florence-carolinas-virginia-20180912-story.html> | <http://www.latimes.com/local/abcarian/la-me-abcarian-ppfa-charges-20170412-story.html> |
| <http://www.latimes.com/nation/la-na-hurricane-florence-live-updates-hurricane-florence-meanders-through-1536975893-htmlstory.html> | <http://www.latimes.com/entertainment/la-et-entertainment-news-updates-august-one-mississippi-season-2-1503947289-htmlstory.html> |
| <http://www.latimes.com/nation/la-na-hurricane-florence-20180914-story.html> | <http://www.latimes.com/local/la-ca-g-rise-of-californias-no-party-voter-20141026-htmlstory.html> |
| <http://www.latimes.com/nation/la-lb-838-47191-la-na-pol-trump-hurricane-florence-20180915-htmlstory.html> | <http://www.latimes.com/local/california/la-me-ln-slater-inland-surf-contest-20180228-story.html> |
| <http://www.latimes.com/nation/la-na-hurricane-florence-live-updates-htmlstory.html> | <http://www.latimes.com/local/lanow/la-me-ln-robicheaux-charges-update-20180921-story.html> |
| <http://www.latimes.com/nation/la-na-hurricane-florence-20180915-story.html> | <http://www.latimes.com/local/abcarian/la-me-abcarian-winchester-mystery-20170510-story.html> |
| <http://www.latimes.com/nation/la-lb-838-47190-la-na-hurricane-florence-20180915-htmlstory.html> | <http://www.latimes.com/world/la-fg-england-meghan-markle-pregnancy-20181015-story.html> |
| <http://www.latimes.com/nation/la-lb-838-47192-la-na-florence-new-bern-20180915-htmlstory.html> | <http://www.latimes.com/business/lazarus/la-fi-lazarus-att-time-warner-price-hike-20180710-story.html> |

|  |  |
| --- | --- |
| Paul Allen |  |
| **LUCENE** | **PAGERANK** |
| <http://www.latimes.com/local/obituaries/la-fi-paul-allen-pictures-20181015-photogallery.html> | <http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-reald-sale-20151109-story.html> |
| <http://www.latimes.com/local/obituaries/la-fi-tn-paul-allen-obituary-20181015-story.html> | <http://www.latimes.com/fashion/la-ig-wwd-estee-lauder-sues-deciem-founder-20181011-story.html> |
| <http://www.latimes.com/local/obituaries/la-fi-tn-paul-allen-obituary-20181015-story.html> | <http://www.latimes.com/business/hollywood/la-fi-ct-att-streaming-service-20181010-story.html> |
| <http://www.latimes.com/business/realestate/hot-property/la-fi-hotprop-paul-allen-beverly-hills-20180710-story.html> | <http://www.latimes.com/sports/ducks/la-sp-ducks-what-we-learned-20181011-story.html><http://www.latimes.com/business/la-fi-himi-jurich-sunrun-20180318-htmlstory.html> |
| <http://www.latimes.com/business/realestate/hot-property/la-fi-hotprop-paul-allen-beverly-hills-20180710-story.html> | <http://www.latimes.com/local/lanow/la-me-ln-adelanto-oig-20181002-story.html> |
| <http://www.latimes.com/local/obituaries/la-fi-tn-paul-allen-obituary-20181015-story.html> | <http://www.latimes.com/health/la-he-safety-tips-for-women-who-run-20180922-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/local/education/la-me-edu-puc-charter-closes-20180824-story.html> |
| [http://www.latimes.com](http://www.latimes.com/) | <http://www.latimes.com/world/europe/la-fg-ireland-pope-visit-20180822-story.html> |
| <http://www.latimes.com/local/obituaries/la-me-maureen-paul-turlish-20180802-story.html> |  |
| <http://www.latimes.com/local/obituaries/la-na-paul-laxalt-20180806-story.html> | <http://www.latimes.com/business/hollywood/la-et-st-inside-actors-studio-ovation-20180924-story.html> |

**Overlaps:**

**No overlaps seen for all queries.**