CSCI 572 Information Retrieval and Search Engines Assignment 4

Name: Parinita Badre USC ID: 8804379764

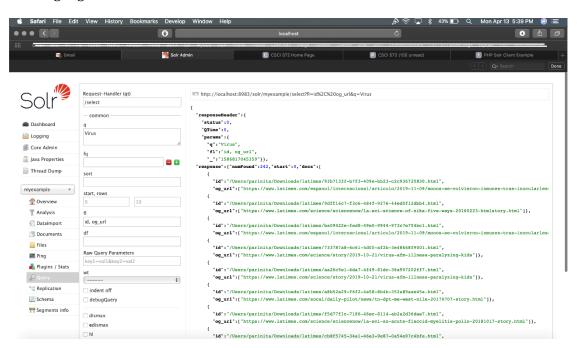
Steps:

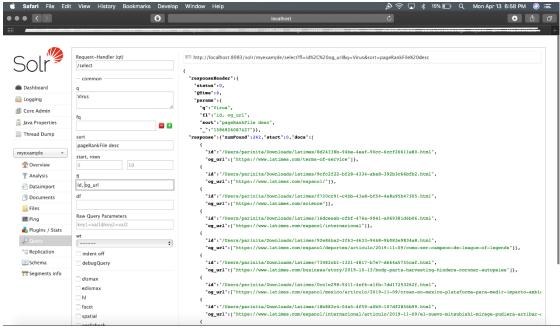
1. Solr Setup

- a. Installed solr-7.7.0
- b. Downloaded the crawl data provided by course producers
- c. Created a core in solr by going to the /solr-7.7.0 directory and running bin/solr create -c myexample
- d. Opened the managed-schema file present in /conf directory and uncommented <copyField source="*" dest="_text_"/> line.
- e. Saved changes and came back to the home directory of solr.
- f. Indexed html files in crawl folder(latimes) by using bin/post -c myexample -filetypes html path_to_crawl_folder/
- g. Checked if the indexing was successful by going to the Solr admin panel at http://localhost:8983/solr/ and selecting 'myexample' from the dropdown
- h. Went to the query section of 'myexample' and submitted the default query (*:*)
- i. The output proved that the indexing was working correctly.
- j. Configured Solr to default querying by changing <!-- <str name="df">text</str>--> in the requestHandler section of solrconfig.xml to <str name="df">_text_</str>

2. Edge list and Page Rank

- a. Wrote a Java program to find edge lists from the pages in latimes folder by using JSoup library
- b. Wrote a Python program using NetworkX library to compute page rank by supplying the EdgeList file as input
- c. Stored the output of the python code in external_pageRankFile.txt
- d. Provided the following parameters to compute the page rank: alpha=0.85, personalization=None, max_iter =30, tol=1e-06, nstart=None, weight='weight', dangling=None





Page Rank results on Solr for query "Virus"

3. Configuring files

a. Stored external_pageRankFile.txt in solr-7.7.0/server/solr/myexample/directory

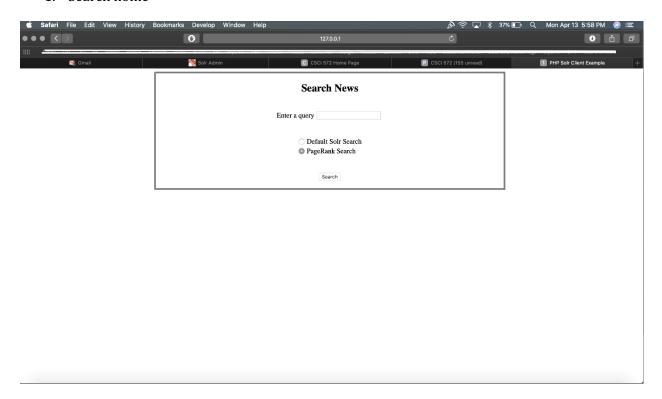
- b. Added the following to managed-schema:
 <fieldType name="external" keyField="id" defVal="0
 class="solr.ExternalFileField"/><field name="pageRankFile" type="external stored="false" indexed="false"
- c. Defined the following within the <query> tags in solrconfig.xml file listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader">
- d. Reloaded the core by clicking on the "Reload" button and verified the results by querying again

4. PHP page

- a. Installed PHP and Apache
- b. Cloned the given solr git repository
- c. Wrote a php code using the sample code provided in the assignment description by including 2 buttons for default solr search and PageRank search index.php
- d. Hosted the webpage and verified the results of the search by comparing them with the results obtained on solr admin panel

PHP Page:

1. Search home



2. Default results:



Total Results: 242 Displaying top 10 results

1. Title: Monos se volvieron inmunes tras inocularles virus del Ébola con mutación - Los Angeles Times

Link: https://www.latimes.com/espano/linteracional/articulo/2019-11-09/mones-se-volvieron-immunes-tras-inocularles-virus-del-ebola-con-mutacion? amp=true ID: //Jsers/parinita/Downloads/latimes/93b7133f-b7f3-409e-bb23-c2c936725830.html

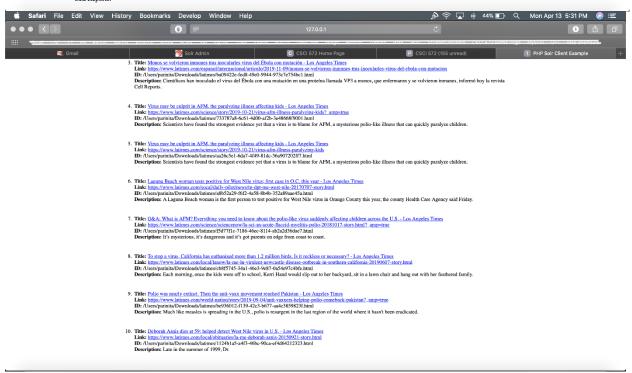
Description: Clentificos han inoculado el virus del Ebola con una mutación en una proteína llamada VP3 a monos, que enfermaron y se volvieron inmunes, informó hoy la revista

Title: Five ways scientists are going after the Zika virus - Los Angeles Times
 Link: https://www.latimes.com/science/sciencenow/la-sci-science-of-zika-five-ways-20160223-htmlstory.html
 ID: /Users/parinita/Downloads/latimes/9dff16c7-f3c6-484f-9276-44ed0f12dbb4.html

Description: To determine whether the Zika virus is indeed responsible for a surge in birth defects, public health officials are drawing on expertise from many disciplines.

3. Title: Monos se volvieron inmunes tras inocularles virus del Ébola con mutación - Los Angeles Times
Link: https://www.latimes.com/espanol/internacional/articulo/2019-11-09/monos-se-volvieron-inmunes-tras-inocularles-virus-del-ebola-con-mutacion
ID: /Users/parinita/Downloads/latimes/ba09422e-fed8-49e0-9944-973c7e754bc1.html

Description: Científicos han inoculado el virus del Ébola con una mutación en una proteína llamada VP3 a monos, que enfermaron y se volvieron inmunes, informó hoy la revista Cell Reports.



3. PageRank results



Total Results: 242 Displaying top 10 results

1. Title: Terms of Service - Los Angeles Times

Link: https://www.latimes.com/terms-of-service
ID: /Users/parinita/Downloads/latimes/8d24338b-94be-4eaf-90cc-6ccf26611e80.html

Description: N/A

2. Title: En Español - Los Angeles Times

Link: https://www.latimes.com/espanol/
ID: /Users/parinita/Downloads/latimes/9cfc2f22-bf28-4334-aba8-392b3c66bfb2.html
Description: N/A

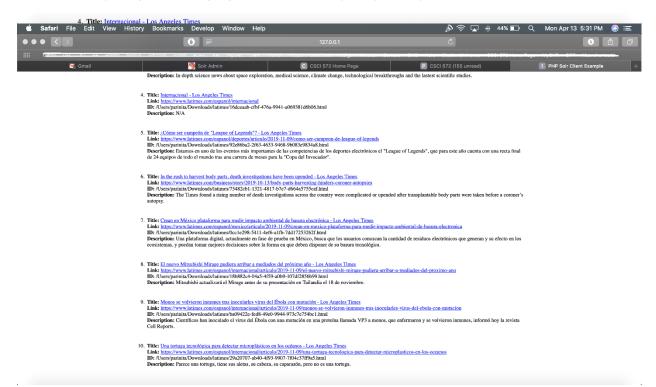
3. Title: Science - Los Angeles Times

Thick, Schemes—Loss Auguste ...tus.

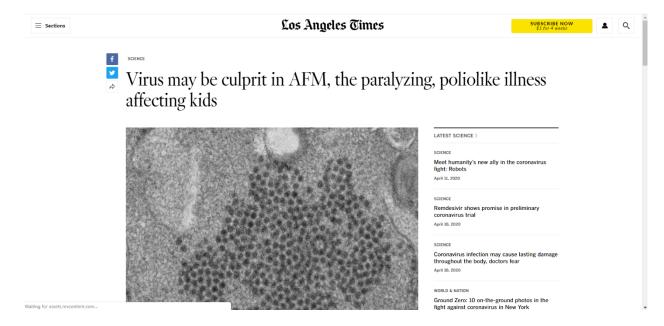
Link: https://www.latimes.com/science

ID: /Users/parinita/Downloads/latimes/f720cc91-c4bb-43a8-bf54-4e8a95b47305.html

Description: In-depth science news about space exploration, medical science, climate change, technological breakthroughs and the lastest scientific studies.



4. Specific URL in the result



Why do some pages have higher page ranks than others?

The PageRank algorithm evaluates the quality and quantity of links to a webpage. This evaluation helps it to determine a relative score of the page's importance. In order to measure this importance, the algorithm counts the number of inlinks and outlinks of the webpage. Inlinks are the number of webpages pointing to this page and outlinks is the number of webpages this page points to. If the number of inlinks is more, the rank of a page is higher. The measure of inlinks and outlinks is determined by calculating EdgeList and then the PageRank is then calculated using the python code in our case.

Comparison of top 10 results of 9 different queries on Lucene (Default) and PageRank:

Cannes	
Default Search Results	PageRank Search Results
https://www.latimes.com/topic/cannes-film-festival	https://www.latimes.com/sitemap
https://www.latimes.com/entertainment/movies/la-et-mn-cannes-wrap-critics-conversation-20190526-story.html	https://www.latimes.com/espanol/entretenim iento/articulo/2019-10-04/por-fin-llega-joker-aqui-un-resumen-de-todo-el-drama-que-ha-provocado
https://www.latimes.com/entertainment/movies/moviesnow/la-et-mn-cannes-2015-winners-20150524-story.html	https://www.latimes.com/entertainment- arts/movies/story/2019-11- 08/documentaries-apollo-11-the-cave-sea-of- shadows-honeyland
https://www.latimes.com/entertainment/movies/la-et-mn-cannes-film-festival-award-winners-20190525-story.html	https://www.latimes.com/lifestyle/gallery/chl oe-sevignys-five-favorite-frocks
https://www.latimes.com/entertainment/movies/la-et-mn-cannes-parasite-atlantics-awards-20190526-story.html	https://www.latimes.com/topic/film-festivals

https://www.latimes.com/entertainment/movies/la-et-mn-cannes-once-upon-a-time-in-hollywood-20190522-story.html	https://www.latimes.com/topic/cannes-film-festival
https://www.latimes.com/entertainment/movies/la-ca-mn-cannes-genre-parasite-wild-goose-lake-tarantino-20190522-story.html	https://www.latimes.com/people/genaro- molina
https://www.latimes.com/entertainment/movies/la-et-mn-cannes-mektoub-my-love-intermezzo-abdellatif-kechiche-20190523-story.html	https://www.latimes.com/entertainment- arts/music/story/2019-11-06/whitney- houston-lesbian-relationship-robyn-crawford
https://www.latimes.com/entertainment/movies/la-et-mn-cannes-quentin-tarantino-once-upon-a-time-in-hollywood-20190521-story.html	https://www.latimes.com/entertainment- arts/movies/story/2019-10-31/apple-disney- hbomax-streaming-movies
https://www.latimes.com/entertainment/movies/la-et-mn-cannes-robert-eggers-the-lighthouse-robert-pattinson-20190524-story.html	https://www.latimes.com/opinion/story/201 9-09-17/patt-morrison-dana-thomas-fast- fashion-environment

Congress	
Default Search Results	PageRank Search Results
https://www.latimes.com/opinion/enterthefra y/la-ol-new-york-rhode-island-congress- marijuana-20190116-story.html	https://www.latimes.com/opinion
https://www.latimes.com/politics/story/2019 -10-28/jerry-brown-to-testify-to-congress-rebutting-trumps-criticism-of-california	https://www.latimes.com/espanol/politica
https://www.latimes.com/politics/la-na-pol- congress-sexual-harassment-20181212- story.html	https://www.latimes.com/politics/story/2019 -11-08/michael-bloomberg-files-papers-for- 2020-democratic-presidential-primary
https://www.latimes.com/politics/la-pol-ca-richest-california-lawmakers-20180305-story.html	https://www.latimes.com/opinion/story/201 9-11-09/facebook-twitter-political-ads-lies
https://www.latimes.com/politics/story/2019 -09-09/congress-gun-control-government-shutdown	https://www.latimes.com/politics/story/2019 -11-08/elizabeth-warren-used-bare-knuckle- tactics-to-take-down-an-obama-nominee
https://www.latimes.com/opinion/story/201 9-10-16/democratic-debate-trump-congress- executive-action	https://www.latimes.com/opinion/story/201 9-11-09/devin-nunes-impeachment-witness- list
https://www.latimes.com/politics/story/2019 -10-14/trump-ukraine-aid-congress- impeachment	https://www.latimes.com/environment/story/2019-11-08/eastern-sierra-towns-recreation-tourism-forest-service
https://www.latimes.com/politics/la-na-pol- congress-bioweapons-detection-system- 20190414-story.html	https://www.latimes.com/politics/story/2019 -11-08/steve-bannon-roger-stone-trump- campaign-wikileaks
https://www.latimes.com/politics/la-na-pol- congress-mueller-pressure-impeachment- 20190529-story.html	https://www.latimes.com/opinion/story/201 9-11-08/impeachment-trump-republicans- toomey
https://www.latimes.com/politics/la-pol-ca-richest-in-congress-darrell-issa-story.html	https://www.latimes.com/politics/story/2019 -11-08/alexander-vindman-fiona-hill- impeachment-transcripts

Democrats	
Default Search Results	PageRank Search Results
https://www.latimes.com/politics/story/2019- 11-07/house-democrats-subpoena-mick- mulvaney-in-impeachment-probe	https://www.latimes.com/california
https://www.latimes.com/opinion/story/2019 -11-05/impeachment-donald-trump- democrats-political-cost-elections	https://www.latimes.com/opinion
https://www.latimes.com/politics/story/2019- 11-03/iowa-democrats-candidate-beat-trump	https://www.latimes.com/politics
https://www.latimes.com/politics/story/2019-10-17/democrats-quick-impeachment-timing-complicated	https://www.latimes.com/politics/story/201 9-11-08/michael-bloomberg-files-papers-for- 2020-democratic-presidential-primary
https://www.latimes.com/politics/story/2019- 10-08/impeachment-trump-democrats- whistleblower-identity-testimony	https://www.latimes.com/politics/story/201 9-11-08/democratic-primary-debate-moves- from-ucla-to-loyola-marymount-university
https://www.latimes.com/politics/story/2019-09-24/democrats-and-pelosi-appear-close-to-tipping-point-on-impeachment	https://www.latimes.com/politics/story/201 9-11-08/elizabeth-warren-used-bare- knuckle-tactics-to-take-down-an-obama- nominee
https://www.latimes.com/politics/story/2019- 09-29/trump-allies-and-democrats-reveal-the- deep-divisions-over-impeachment-inquiry	https://www.latimes.com/opinion/story/201 9-11-09/elizabeth-warren-trump
https://www.latimes.com/politics/story/2019-08-01/assignment-post-debate-health-care-analysis-new-story	https://www.latimes.com/opinion/story/201 9-11-09/devin-nunes-impeachment-witness- list
https://www.latimes.com/nation/ct-democrats-economic-plan-20170824-story.html	https://www.latimes.com/politics/story/201 9-11-08/steve-bannon-roger-stone-trump- campaign-wikileaks
https://www.latimes.com/california/story/201 9-10-21/california-independent-voters-can- cast-ballots-for-democrats-not-trump-march- 2020-primary	https://www.latimes.com/opinion/story/201 9-11-08/impeachment-trump-republicans- toomey

Patriot Movement	
Default Search Results	PageRank Search Results
https://www.latimes.com/archives/la-xpm-	
2012-oct-23-la-me-russell-means-20121023-	https://www.latimes.com/environment
story.html	
https://www.latimes.com/staff/megan-garvey	https://www.latimes.com/science
https://www.latimes.com/archives/la-xpm-	https://www.latimes.com/california/story/20
2001-jun-10-mn-8792-story.html	19-11-09/prop-187-anniversary-
2001-jun-10-min-07 72-story.num	commemoration
https://www.latimes.com/entertainment-	https://www.latimes.com/entertainment-
arts/books/story/2019-10-25/shadowlands-	arts/story/2019-11-09/classical-music-things-
anthony-mccann-oregon-standoff	to-do-in-la-this-week-nov-10-17-magic-flute-
anthony meetin oregon standon	la-opera
	https://www.latimes.com/entertainment-
https://www.latimes.com/sports/highschool/	arts/story/2019-11-09/theater-things-to-do-
story/2019-11-06/girls-tennis-southern-	in-la-this-week-nov-10-17-dr-seusss-how-the-
section-playoff-results-and-updated-pairings	grinch-stole-christmas-old-globe-key-largo-
	andy-garcia-geffen-playhouse

https://www.latimes.com/politics/story/2019 -09-20/bernie-sanders-muslim-voters-2020	https://www.latimes.com/food/story/2019- 11-07/kung-pao-chicken-history-recipe-gong- bao
https://www.latimes.com/sports/nba/la-sp-nba-best-game-ever-20181222-story.html	https://www.latimes.com/obituaries/story/20 19-11-04/james-stern-obit-black-activist-who- led-neo-nazi-group-dies-amid-bid-to-destroy-it
https://www.latimes.com/sports/highschool/story/2019-11-06/girls-tennis-southern-section-playoff-results-and-updated-pairings	https://www.latimes.com/topic/column-one
https://www.latimes.com/entertainment/movies/la-et-mn-july-4-box-office-history-20180629-story.html	https://www.latimes.com/topic/mexico- americas
https://www.latimes.com/entertainment/movies/la-et-mn-july-4-box-office-history-20180629-story.html	https://www.latimes.com/topic/museums

Republicans	
Default Search Results	PageRank Search Results
https://www.latimes.com/opinion/story/2019 -10-23/house-republicans-storm-hearing-impeachment	https://www.latimes.com/
https://www.latimes.com/politics/story/2019-10-30/democrats-hoped-theyd-win-over-republicans-on-impeachment-but-its-not-looking-that-way-so-far	https://www.latimes.com/opinion
https://www.latimes.com/opinion/story/2019 -11-08/impeachment-trump-republicans- toomey	https://www.latimes.com/opinion/la-letter- to-the-editor-htmlstory.html
https://www.latimes.com/politics/story/2019-10-04/vulnerable-senate-republicans-impeachment	https://www.latimes.com/opinion/story/201 9-11-09/facebook-twitter-political-ads-lies
https://www.latimes.com/california/story/201 9-10-16/california-republicans-democrats- wildfires-homelessness	https://www.latimes.com/opinion/story/201 9-11-08/berlin-wall-30th-anniversary-cold- war-donald-trump
https://www.latimes.com/politics/story/2019- 10-23/impeachment-deposition-room- stormed-by-republicans	https://www.latimes.com/politics/story/201 9-11-08/elizabeth-warren-used-bare- knuckle-tactics-to-take-down-an-obama- nominee
https://www.latimes.com/opinion/story/2019 -09-17/california-republicans-trumpfundraising-president	https://www.latimes.com/opinion/story/201 9-11-09/elizabeth-warren-trump
https://www.latimes.com/politics/story/2019-10-28/katie-hill-resignation-trump-obstacle-for-republicans	https://www.latimes.com/opinion/story/201 9-11-09/devin-nunes-impeachment-witness- list
https://www.latimes.com/opinion/story/2019 -11-06/trump-republicans-democrats- predictions-polling-2020	https://www.latimes.com/opinion/story/201 9-11-09/sat-uc-college-admissions
https://www.latimes.com/politics/story/2019- 11-06/republicans-election-trump-candidate- loses-kentucky	https://www.latimes.com/opinion/story/201 9-11-08/impeachment-trump-republicans- toomey

Senate	
Default Search Results	PageRank Search Results
https://www.latimes.com/politics/story/201 9-11-07/gop-is-already-thinking-about-how- to-turn-a-senate-impeachment-trial-to- trumps-advantage	https://www.latimes.com/
https://www.latimes.com/politics/story/201 9-10-30/california-donors-spend-millions-on- 2020-senate-races-across-the-country	https://www.latimes.com/opinion
https://www.latimes.com/politics/story/201 9-10-04/vulnerable-senate-republicans- impeachment	https://www.latimes.com/california/story/201 9-11-09/prop-187-anniversary- commemoration
https://www.latimes.com/politics/la-na-pol-kavanaugh-hearing-20180927-story.html	https://www.latimes.com/opinion/la-letter-to-the-editor-htmlstory.html
https://www.latimes.com/nation/la-pol- scotus-confirmation-votes-over-the-years- 20181005-htmlstory.html	https://www.latimes.com/opinion/story/2019 -11-09/facebook-twitter-political-ads-lies
https://www.latimes.com/politics/story/201 9-08-14/2020-senate-control-presidential- race	https://www.latimes.com/opinion/story/2019 -11-08/berlin-wall-30th-anniversary-cold-wardonald-trump
https://www.latimes.com/politics/story/201 9-08-14/john-hickenlooper-quits- presidential-race-for-senate-run	https://www.latimes.com/politics/story/2019- 11-08/elizabeth-warren-used-bare-knuckle- tactics-to-take-down-an-obama-nominee
https://www.latimes.com/politics/story/201 9-10-30/california-donors-spend-millions-on- 2020-senate-races-across-the-country	https://www.latimes.com/opinion/story/2019 -11-09/elizabeth-warren-trump
https://www.latimes.com/politics/la-na-pol-william-barr-senate-confirm-attorney-general-20190214-story.html	https://www.latimes.com/opinion/story/2019 -11-09/devin-nunes-impeachment-witness-list
https://www.latimes.com/opinion/story/201 9-10-15/wmcaleen-homeland-security- trump-nielsen-senate-acting-secretary	https://www.latimes.com/opinion/story/2019 -11-09/sat-uc-college-admissions

Olympics 2020	
Default Search Results	PageRank Search Results
https://www.latimes.com/sports/olympics/st ory/2019-08-15/2020-tokyo-olympics- searching-for-answers-amid-heat-wave	https://www.latimes.com/
https://www.latimes.com/sports/olympics	https://www.latimes.com/sitemap
https://www.latimes.com/people/david-wharton	https://www.latimes.com/sports
https://www.latimes.com/people/david-wharton	https://www.latimes.com/opinion
https://www.latimes.com/sports/olympics/st ory/2019-10-16/2020-tokyo-olympics- marathon-sapporo-avoid-heat	https://www.latimes.com/business
https://www.latimes.com/sports/olympics/st ory/2019-07-31/deadly-heat-wave-2020- summer-olympics-tokyo	https://www.latimes.com/entertainment-arts
https://www.latimes.com/sports/story/2019- 11-05/ioc-anti-doping-2020-tokyo-olympics	https://www.latimes.com/world-nation

https://www.latimes.com/sports/olympics/st ory/2019-10-10/naomi-osaka-chooses-japan- over-u-s-2020-tokyo-olympics	https://www.latimes.com/environment
https://www.latimes.com/sports/story/2019-07-23/tokyos-rough-road-to-2020-summer-olympics	https://www.latimes.com/newsroom- directory
https://www.latimes.com/people/david- wharton	https://www.latimes.com/politics

Stock	
Default Search Results	PageRank Search Results
https://www.latimes.com/business/story/201 9-11-06/uber-lock-up-period-ends-with- falling-stock-protests	https://www.latimes.com/
https://www.latimes.com/business/la-fi-lyft-stock-20190401-story.html	https://www.latimes.com/business
https://www.latimes.com/business/la-fi-hy-tesla-stock-20160518-snap-story.html	https://www.latimes.com/business/realestate
https://www.latimes.com/business/story/201 9-11-06/uber-lock-up-period-ends-with- falling-stock-protests	https://www.latimes.com/business/realestate/story/2019-11-09/hulu-looking-foralaska-landry-bender
https://www.latimes.com/business/story/201 9-11-09/stock-managers-who-played-defense- in-2019-are-left-scrambling-to-make-up- ground	https://www.latimes.com/business/story/201 9-11-08/trump-china-tariffs
https://www.latimes.com/business/la-fi-uber-ipo-stock-trading-price-20190510-story.html	https://www.latimes.com/business/realestate/story/2019-11-09/hot-property-verybrady-sale-malibu
https://www.latimes.com/business/story/201 9-09-09/at-t-stock-surges-elliott-reveals-3-2- billion-stake	https://www.latimes.com/politics/story/2019 -11-08/elizabeth-warren-used-bare-knuckle- tactics-to-take-down-an-obama-nominee
https://www.latimes.com/business/la-fi- stock-market-status-update-20180209- story.html	https://www.latimes.com/business/real- estate/story/2019-11-09/hot-property- newsletter-rifleman-leads-the-posse-on-a- little-mosey-down-memory-lane
https://www.latimes.com/business/la-fi- stock-market-correction-20180212-story.html	https://www.latimes.com/business/story/201 9-11-09/stock-managers-who-played-defense- in-2019-are-left-scrambling-to-make-up- ground
https://www.latimes.com/business/story/201 9-11-09/stock-managers-who-played-defense- in-2019-are-left-scrambling-to-make-up- ground	https://www.latimes.com/business/story/201 9-11-08/gap-ceo-was-done-in-by-fashion- missteps-and-fading-brands

Virus	
Default Search Results	PageRank Search Results
https://www.latimes.com/espanol/internacio nal/articulo/2019-11-09/monos-se-volvieron- inmunes-tras-inocularles-virus-del-ebola-con- mutacion	https://www.latimes.com/terms-of-service

https://www.latimes.com/science/sciencenow/la-sci-science-of-zika-five-ways-20160223-htmlstory.html	https://www.latimes.com/espanol/
https://www.latimes.com/espanol/internacio nal/articulo/2019-11-09/monos-se-volvieron- inmunes-tras-inocularles-virus-del-ebola-con- mutacion	https://www.latimes.com/science
https://www.latimes.com/science/story/2019 -10-21/virus-afm-illness-paralyzing-kids	https://www.latimes.com/espanol/internacional
https://www.latimes.com/science/story/2019 -10-21/virus-afm-illness-paralyzing-kids	https://www.latimes.com/espanol/deportes/a rticulo/2019-11-09/como-ser-campeon-de- league-of-legends
https://www.latimes.com/socal/daily- pilot/news/tn-dpt-me-west-nile-20170707- story.html	https://www.latimes.com/business/story/201 9-10-13/body-parts-harvesting-hinders- coroner-autopsies
https://www.latimes.com/science/sciencenow/la-sci-sn-acute-flaccid-myelitis-polio-20181017-story.html	https://www.latimes.com/espanol/mexico/art iculo/2019-11-09/crean-en-mexico-plataforma-para-medir-impacto-ambiental-debasura-electronica
https://www.latimes.com/local/lanow/la-me-ln-virulent-newcastle-disease-outbreak-in-southern-california-20190607-story.html	https://www.latimes.com/espanol/internacio nal/articulo/2019-11-09/el-nuevo-mitsubishi- mirage-pudiera-arribar-a-mediados-del- proximo-ano
https://www.latimes.com/world- nation/story/2019-09-04/anti-vaxxers- helping-polio-comeback-pakistan	https://www.latimes.com/espanol/internacio nal/articulo/2019-11-09/monos-se-volvieron- inmunes-tras-inocularles-virus-del-ebola-con- mutacion
https://www.latimes.com/local/obituaries/la- me-deborah-asnis-20150921-story.html	https://www.latimes.com/espanol/internacio nal/articulo/2019-11-09/una-tortuga- tecnologica-para-detectar-microplasticos-en- los-oceanos

Overlap Chart:

