Repo

Yiming Chen

ID:9600260082

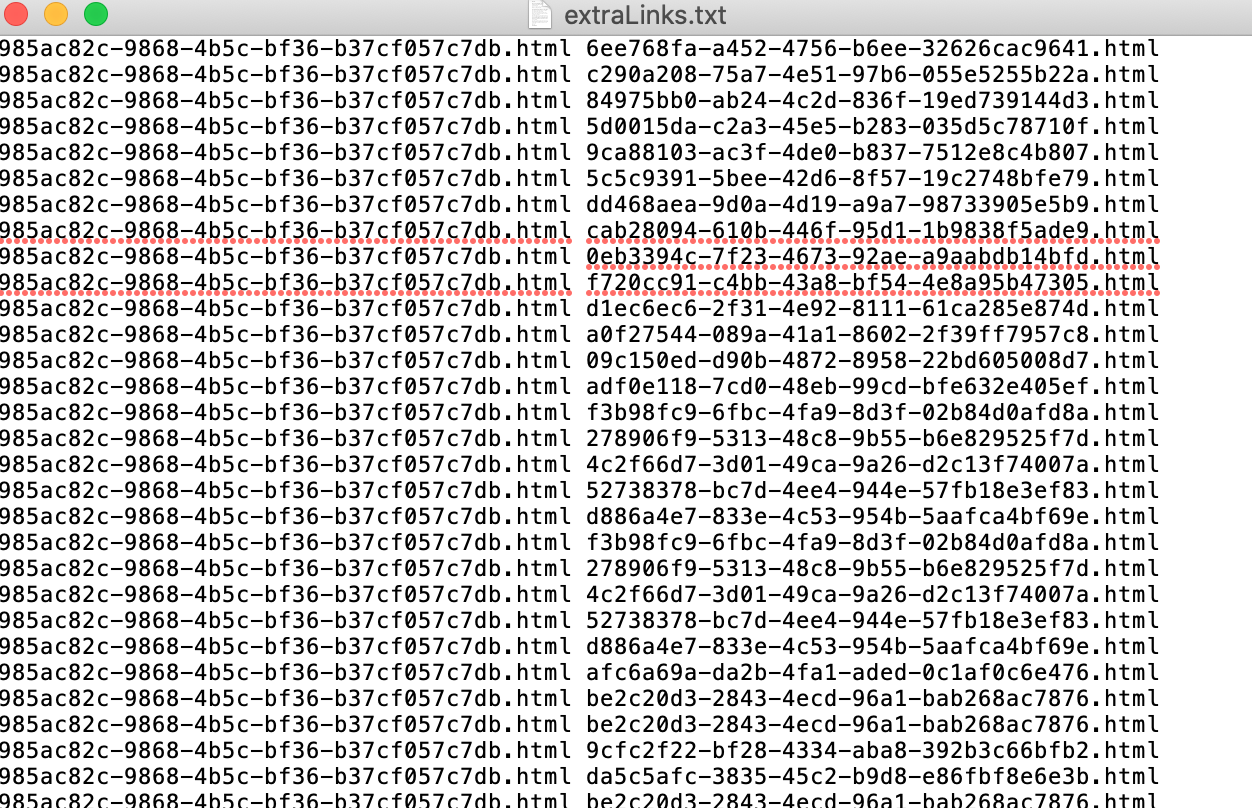
Email: ychen983@usc.edu

For this homework, there are mainly two part, the first one is compute the page rank, and second using solr to submit data and show in front end.

PageRank

1.1 CSV document: we need to put it as input and using hash map to store the key: URL address, value: document name. One we extra links from webpages and we can map it to document name. regrade as link.

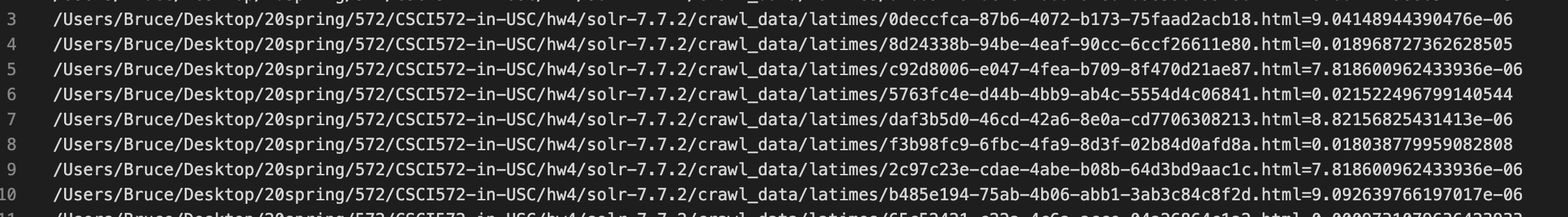
1.2 Using Jsoup to extract links from webpages. My program is to us Jsoup to parse each document and extract the links it contains. And also map it to actual document name as “a b”. we can think there are edge lists for graph.



1.3 Using NetworkX to calculate the page rank for this. Just easily apply

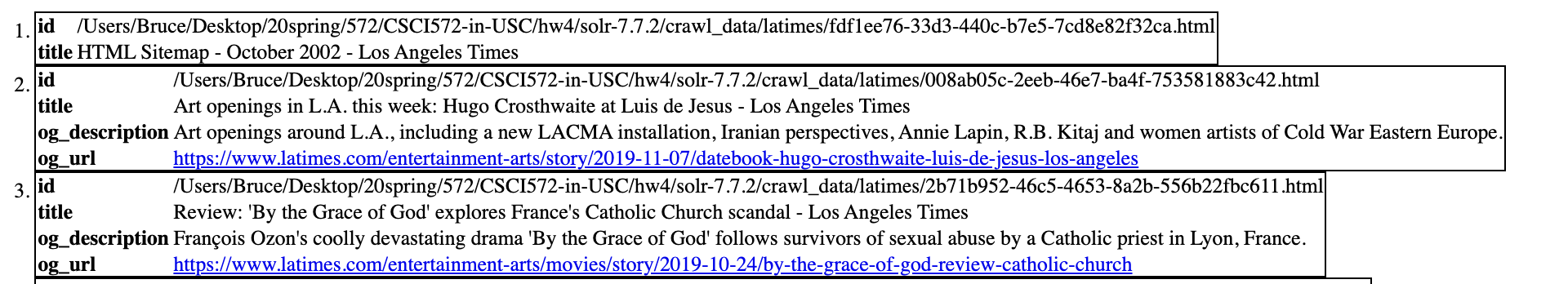
pr = nx.pagerank(G, alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None)

to get the graph. And Next step is output the results of each node (document). And set the format as <id> = <pagerank>



Solr

2.1 using example php code to build the front end. Which could make a query and send the request to the Solr system, and fetch the results. So what we need to do is filter the data, and present it in a clear way. We just need id, title, og\_description and og\_url. SO for my method is to show the fields which is those name, therefor we get only four fields only for each result.



2.2 install Solr and start solr at localhost port 8983.

2.3 new core. We need to create a new core -> myexample, and and enable

<copyField source="\*" dest="\_text\_"/> In managed-schema.  
then run the index by lucene.  


2.4 apply page rank to solr. Motify two documents managed-schemaandsolrconfig.xml

for using external\_pageRankFile to sort the results



2.5 using $additionalParameters = array(

'sort' => 'pageRankFile desc'

);

$results = $solr->search($query, 0, $limit, $additionalParameters);

To fetch data which using the sort to get the results follow the pageRankFile results.

Result

|  |  |
| --- | --- |
| Query | Overlap |
| Cannes | 1 |
| Congress | 0 |
| Democrats | 0 |
| Patriot Movement | 0 |
| Republicans | 1 |
| Senate | 0 |
| Olympics 2020 | 0 |
| Stock | 1 |
| Virus | 1 |

t**he Links table and screenshots are attached in the end.**

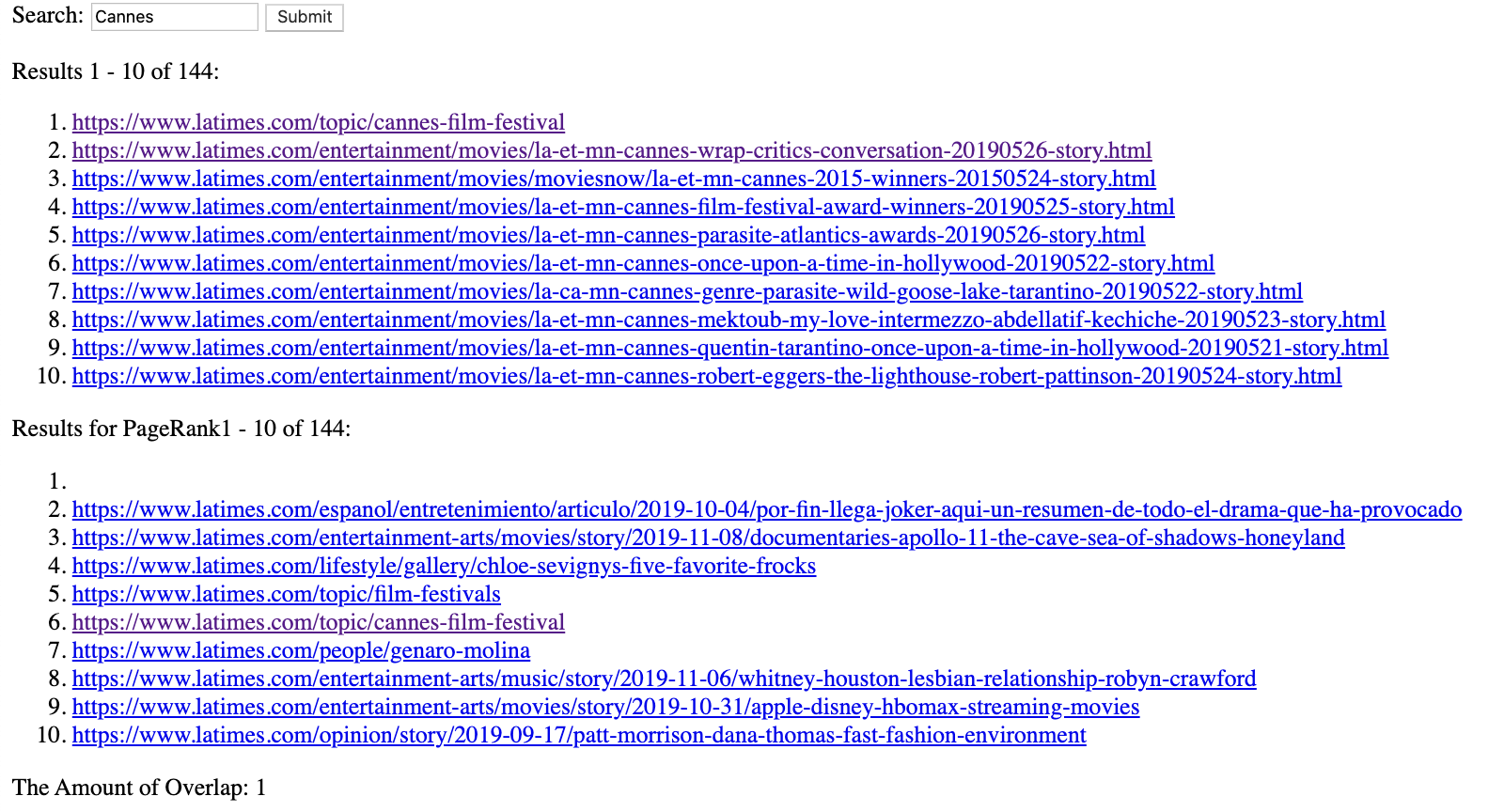
I just post it as screenshot rather than table, I also attach it after the each image. (for image, I use another code, that maybe have null og\_url, but I use another code for it to present id, og\_url, and title so to ensure the url)

About Page Rank Values

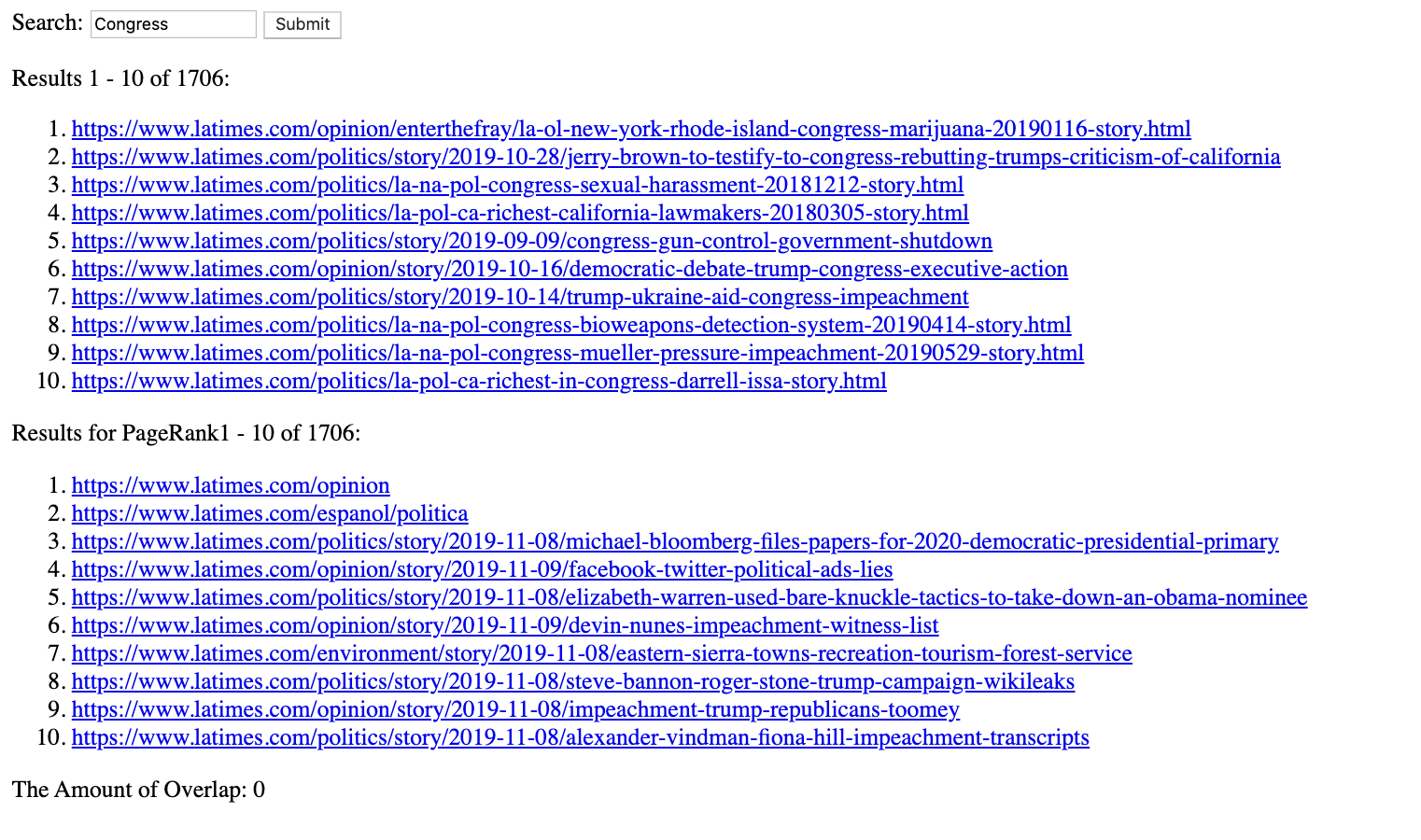
PageRank use the weight only focus on how many pages point out and point to this page. So the higher page weight because more pages point to it. And for Solr algorithm, it use tf-idf value to calculate the weight for Boolean and vector model.

There are different algorithm to weight pages, the more pages point to it will get higher weight and another is base on word. That’s the reason why there are different page rank values.

The search results and overlapping



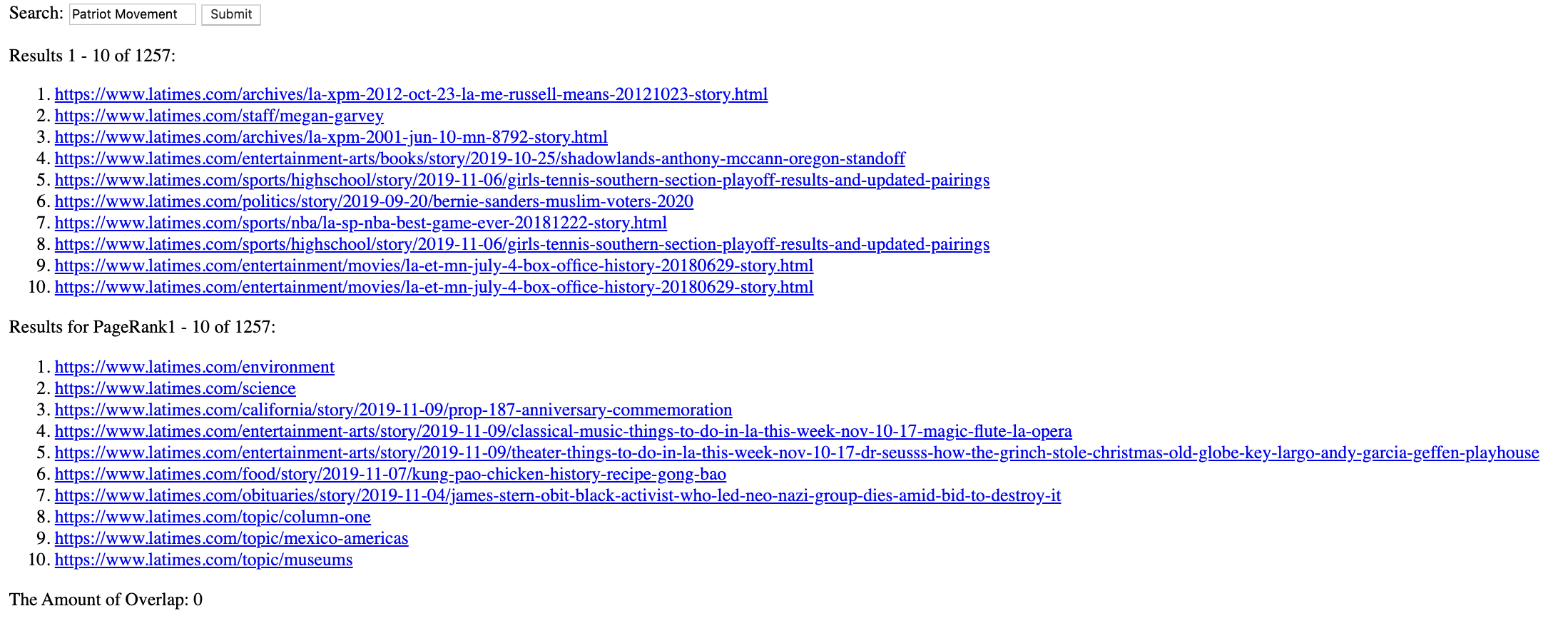
|  |  |
| --- | --- |
| https://www.latimes.com/topic/cannes-film-festival  https://www.latimes.com/entertainment/movies/la-et-mn-cannes-wrap-critics-conversation-20190526-story.html  https://www.latimes.com/entertainment/movies/moviesnow/la-et-mn-cannes-2015-winners-20150524-story.html  https://www.latimes.com/entertainment/movies/la-et-mn-cannes-film-festival-award-winners-20190525-story.html  https://www.latimes.com/entertainment/movies/la-et-mn-cannes-parasite-atlantics-awards-20190526-story.html  https://www.latimes.com/entertainment/movies/la-et-mn-cannes-once-upon-a-time-in-hollywood-20190522-story.html  https://www.latimes.com/entertainment/movies/la-ca-mn-cannes-genre-parasite-wild-goose-lake-tarantino-20190522-story.html  https://www.latimes.com/entertainment/movies/la-et-mn-cannes-mektoub-my-love-intermezzo-abdellatif-kechiche-20190523-story.html  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/movies/la-et-mn-cannes-robert-eggers-the-lighthouse-robert-pattinson-20190524-story.html | <https://www.latimes.com/sitemap>  https://www.latimes.com/espanol/entretenimiento/articulo/2019-10-04/por-fin-llega-joker-aqui-un-resumen-de-todo-el-drama-que-ha-provocado  https://www.latimes.com/topic/ film-festival  https://www.latimes.com/entertainment-arts/movies/story/2019-11-08/documentaries-apollo-11-the-cave-sea-of-shadows-honeyland  https://www.latimes.com/lifestyle/gallery/chloe-sevignys-five-favorite-frocks  https://www.latimes.com/topic/cannes-film-festival  https://www.latimes.com/people/genaro-molina  https://www.latimes.com/entertainment-arts/music/story/2019-11-06/whitney-houston-lesbian-relationship-robyn-crawford  https://www.latimes.com/entertainment-arts/movies/story/2019-10-31/apple-disney-hbomax-streaming-movies  https://www.latimes.com/opinion/story/2019-09-17/patt-morrison-dana-thomas-fast-fashion-environment |



|  |  |
| --- | --- |
| 1. <https://www.latimes.com/opinion/enterthefray/la-ol-new-york-rhode-island-congress-marijuana-20190116-story.html> 2. <https://www.latimes.com/politics/story/2019-10-28/jerry-brown-to-testify-to-congress-rebutting-trumps-criticism-of-california> 3. <https://www.latimes.com/politics/la-na-pol-congress-sexual-harassment-20181212-story.html> 4. <https://www.latimes.com/politics/la-pol-ca-richest-california-lawmakers-20180305-story.html> 5. <https://www.latimes.com/politics/story/2019-09-09/congress-gun-control-government-shutdown> 6. <https://www.latimes.com/opinion/story/2019-10-16/democratic-debate-trump-congress-executive-action> 7. <https://www.latimes.com/politics/story/2019-10-14/trump-ukraine-aid-congress-impeachment> 8. <https://www.latimes.com/politics/la-na-pol-congress-bioweapons-detection-system-20190414-story.html> 9. <https://www.latimes.com/politics/la-na-pol-congress-mueller-pressure-impeachment-20190529-story.html> 10. <https://www.latimes.com/politics/la-pol-ca-richest-in-congress-darrell-issa-story.html> | 1. <https://www.latimes.com/opinion> 2. <https://www.latimes.com/espanol/politica> 3. <https://www.latimes.com/politics/story/2019-11-08/michael-bloomberg-files-papers-for-2020-democratic-presidential-primary> 4. <https://www.latimes.com/opinion/story/2019-11-09/facebook-twitter-political-ads-lies> 5. <https://www.latimes.com/politics/story/2019-11-08/elizabeth-warren-used-bare-knuckle-tactics-to-take-down-an-obama-nominee> 6. <https://www.latimes.com/opinion/story/2019-11-09/devin-nunes-impeachment-witness-list> 7. <https://www.latimes.com/environment/story/2019-11-08/eastern-sierra-towns-recreation-tourism-forest-service> 8. <https://www.latimes.com/politics/story/2019-11-08/steve-bannon-roger-stone-trump-campaign-wikileaks> 9. <https://www.latimes.com/opinion/story/2019-11-08/impeachment-trump-republicans-toomey> 10. <https://www.latimes.com/politics/story/2019-11-08/alexander-vindman-fiona-hill-impeachment-transcripts> |



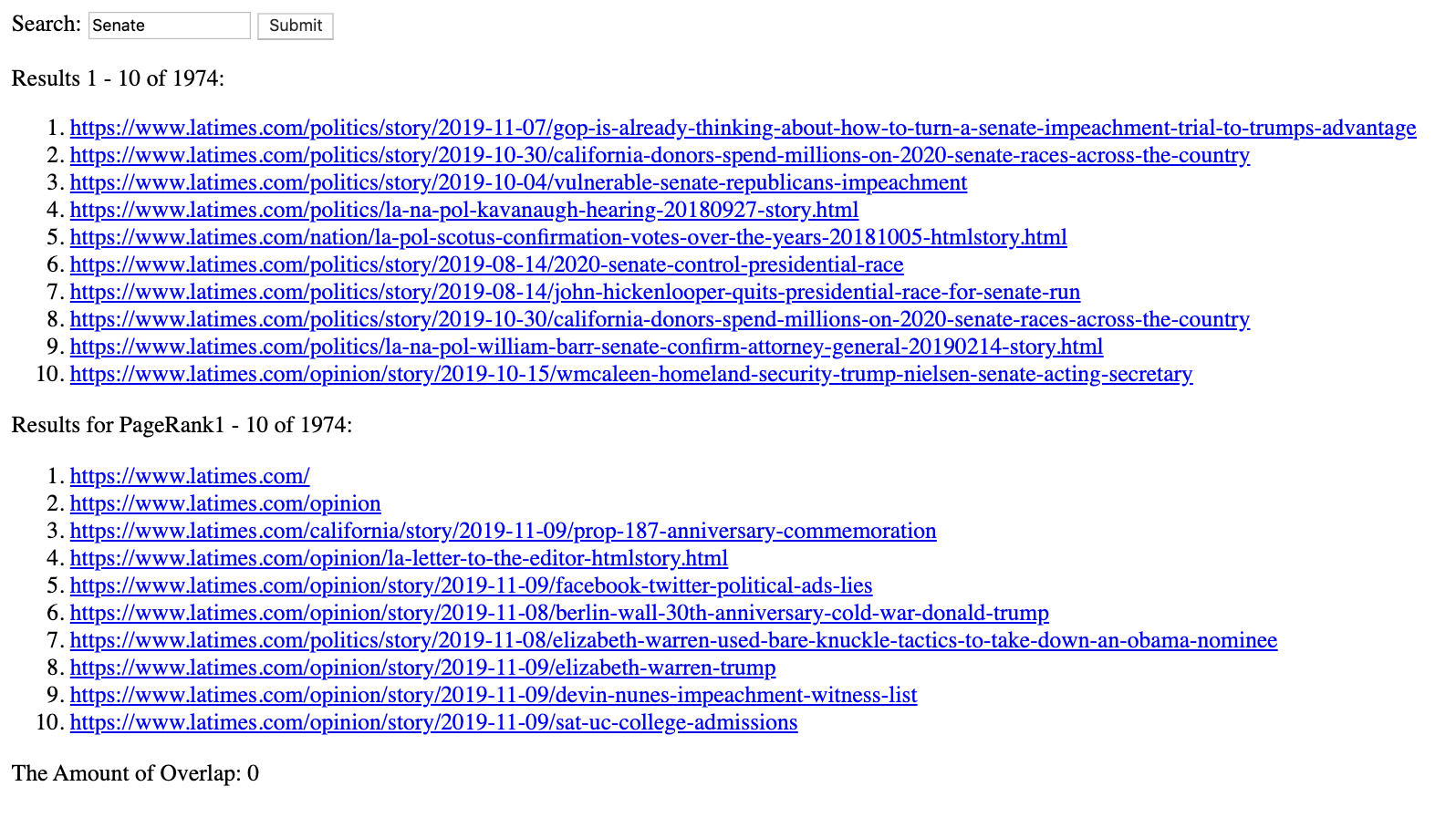
|  |  |
| --- | --- |
| 1. <https://www.latimes.com/politics/story/2019-11-07/house-democrats-subpoena-mick-mulvaney-in-impeachment-probe> 2. <https://www.latimes.com/opinion/story/2019-11-05/impeachment-donald-trump-democrats-political-cost-elections> 3. <https://www.latimes.com/politics/story/2019-11-03/iowa-democrats-candidate-beat-trump> 4. <https://www.latimes.com/politics/story/2019-10-17/democrats-quick-impeachment-timing-complicated> 5. <https://www.latimes.com/politics/story/2019-10-08/impeachment-trump-democrats-whistleblower-identity-testimony> 6. <https://www.latimes.com/politics/story/2019-09-24/democrats-and-pelosi-appear-close-to-tipping-point-on-impeachment> 7. <https://www.latimes.com/politics/story/2019-09-29/trump-allies-and-democrats-reveal-the-deep-divisions-over-impeachment-inquiry> 8. <https://www.latimes.com/politics/story/2019-08-01/assignment-post-debate-health-care-analysis-new-story> 9. <https://www.latimes.com/nation/ct-democrats-economic-plan-20170824-story.html> 10. <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> | 1. <https://www.latimes.com/california> 2. <https://www.latimes.com/opinion> 3. <https://www.latimes.com/politics> 4. <https://www.latimes.com/politics/story/2019-11-08/michael-bloomberg-files-papers-for-2020-democratic-presidential-primary> 5. <https://www.latimes.com/politics/story/2019-11-08/democratic-primary-debate-moves-from-ucla-to-loyola-marymount-university> 6. <https://www.latimes.com/politics/story/2019-11-08/elizabeth-warren-used-bare-knuckle-tactics-to-take-down-an-obama-nominee> 7. <https://www.latimes.com/opinion/story/2019-11-09/elizabeth-warren-trump> 8. <https://www.latimes.com/opinion/story/2019-11-09/devin-nunes-impeachment-witness-list> 9. <https://www.latimes.com/politics/story/2019-11-08/steve-bannon-roger-stone-trump-campaign-wikileaks> 10. <https://www.latimes.com/opinion/story/2019-11-08/impeachment-trump-republicans-toomey> |



|  |  |
| --- | --- |
| 1. <https://www.latimes.com/archives/la-xpm-2012-oct-23-la-me-russell-means-20121023-story.html> 2. <https://www.latimes.com/staff/megan-garvey> 3. <https://www.latimes.com/archives/la-xpm-2001-jun-10-mn-8792-story.html> 4. <https://www.latimes.com/entertainment-arts/books/story/2019-10-25/shadowlands-anthony-mccann-oregon-standoff> 5. <https://www.latimes.com/sports/highschool/story/2019-11-06/girls-tennis-southern-section-playoff-results-and-updated-pairings> 6. <https://www.latimes.com/politics/story/2019-09-20/bernie-sanders-muslim-voters-2020> 7. <https://www.latimes.com/sports/nba/la-sp-nba-best-game-ever-20181222-story.html> 8. <https://www.latimes.com/sports/highschool/story/2019-11-06/girls-tennis-southern-section-playoff-results-and-updated-pairings> 9. <https://www.latimes.com/entertainment/movies/la-et-mn-july-4-box-office-history-20180629-story.html> 10. <https://www.latimes.com/entertainment/movies/la-et-mn-july-4-box-office-history-20180629-story.html> | 1. <https://www.latimes.com/environment> 2. <https://www.latimes.com/science> 3. <https://www.latimes.com/california/story/2019-11-09/prop-187-anniversary-commemoration> 4. <https://www.latimes.com/entertainment-arts/story/2019-11-09/classical-music-things-to-do-in-la-this-week-nov-10-17-magic-flute-la-opera> 5. <https://www.latimes.com/entertainment-arts/story/2019-11-09/theater-things-to-do-in-la-this-week-nov-10-17-dr-seusss-how-the-grinch-stole-christmas-old-globe-key-largo-andy-garcia-geffen-playhouse> 6. <https://www.latimes.com/food/story/2019-11-07/kung-pao-chicken-history-recipe-gong-bao> 7. <https://www.latimes.com/obituaries/story/2019-11-04/james-stern-obit-black-activist-who-led-neo-nazi-group-dies-amid-bid-to-destroy-it> 8. <https://www.latimes.com/topic/column-one> 9. <https://www.latimes.com/topic/mexico-americas> 10. <https://www.latimes.com/topic/museums> |



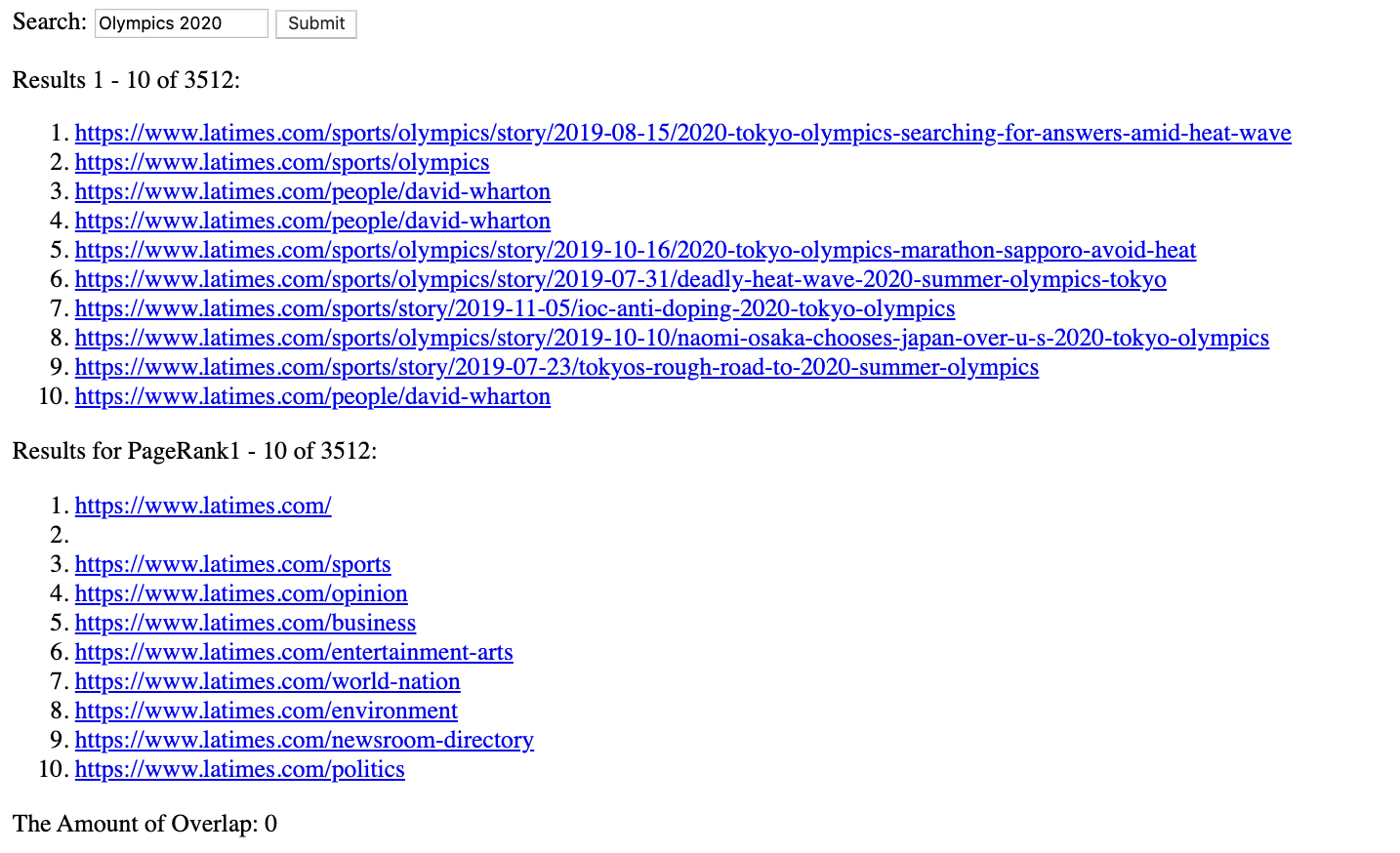
|  |  |
| --- | --- |
| 1. <https://www.latimes.com/opinion/story/2019-10-23/house-republicans-storm-hearing-impeachment> 2. <https://www.latimes.com/politics/story/2019-10-30/democrats-hoped-theyd-win-over-republicans-on-impeachment-but-its-not-looking-that-way-so-far> 3. <https://www.latimes.com/opinion/story/2019-11-08/impeachment-trump-republicans-toomey> 4. <https://www.latimes.com/politics/story/2019-10-04/vulnerable-senate-republicans-impeachment> 5. <https://www.latimes.com/california/story/2019-10-16/california-republicans-democrats-wildfires-homelessness> 6. <https://www.latimes.com/politics/story/2019-10-23/impeachment-deposition-room-stormed-by-republicans> 7. <https://www.latimes.com/opinion/story/2019-09-17/california-republicans-trump-fundraising-president> 8. <https://www.latimes.com/politics/story/2019-10-28/katie-hill-resignation-trump-obstacle-for-republicans> 9. <https://www.latimes.com/opinion/story/2019-11-06/trump-republicans-democrats-predictions-polling-2020> 10. <https://www.latimes.com/politics/story/2019-11-06/republicans-election-trump-candidate-loses-kentucky> | 1. <https://www.latimes.com/> 2. <https://www.latimes.com/opinion> 3. <https://www.latimes.com/opinion/la-letter-to-the-editor-htmlstory.html> 4. <https://www.latimes.com/opinion/story/2019-11-09/facebook-twitter-political-ads-lies> 5. <https://www.latimes.com/opinion/story/2019-11-08/berlin-wall-30th-anniversary-cold-war-donald-trump> 6. <https://www.latimes.com/politics/story/2019-11-08/elizabeth-warren-used-bare-knuckle-tactics-to-take-down-an-obama-nominee> 7. <https://www.latimes.com/opinion/story/2019-11-09/elizabeth-warren-trump> 8. <https://www.latimes.com/opinion/story/2019-11-09/devin-nunes-impeachment-witness-list> 9. <https://www.latimes.com/opinion/story/2019-11-09/sat-uc-college-admissions> 10. <https://www.latimes.com/opinion/story/2019-11-08/impeachment-trump-republicans-toomey> |



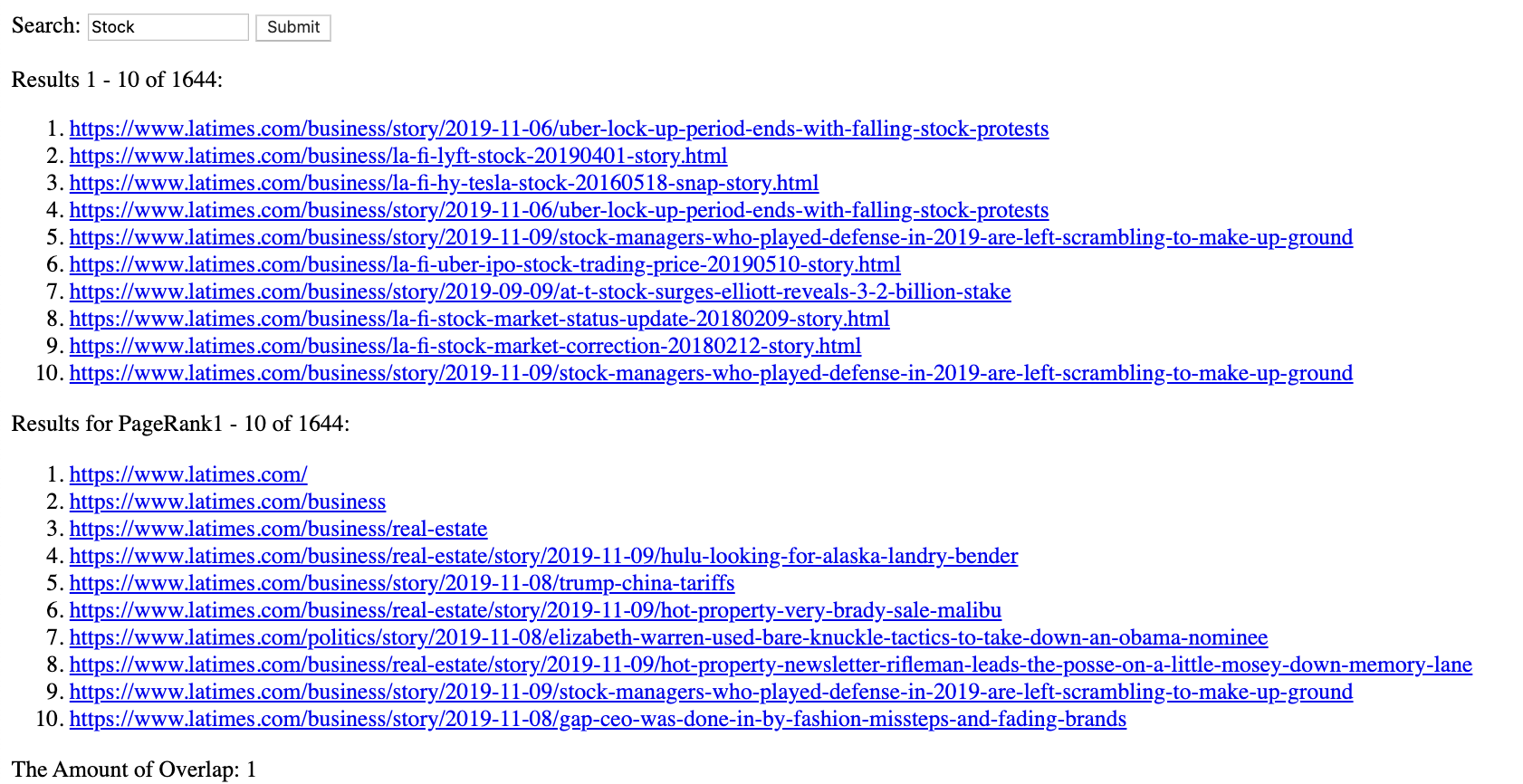
|  |  |
| --- | --- |
| 1. <https://www.latimes.com/politics/story/2019-11-07/gop-is-already-thinking-about-how-to-turn-a-senate-impeachment-trial-to-trumps-advantage> 2. <https://www.latimes.com/politics/story/2019-10-30/california-donors-spend-millions-on-2020-senate-races-across-the-country> 3. <https://www.latimes.com/politics/story/2019-10-04/vulnerable-senate-republicans-impeachment> 4. <https://www.latimes.com/politics/la-na-pol-kavanaugh-hearing-20180927-story.html> 5. <https://www.latimes.com/nation/la-pol-scotus-confirmation-votes-over-the-years-20181005-htmlstory.html> 6. <https://www.latimes.com/politics/story/2019-08-14/2020-senate-control-presidential-race> 7. <https://www.latimes.com/politics/story/2019-08-14/john-hickenlooper-quits-presidential-race-for-senate-run> 8. <https://www.latimes.com/politics/story/2019-10-30/california-donors-spend-millions-on-2020-senate-races-across-the-country> 9. <https://www.latimes.com/politics/la-na-pol-william-barr-senate-confirm-attorney-general-20190214-story.html> 10. <https://www.latimes.com/opinion/story/2019-10-15/wmcaleen-homeland-security-trump-nielsen-senate-acting-secretary> | 1. <https://www.latimes.com/> 2. <https://www.latimes.com/opinion> 3. <https://www.latimes.com/california/story/2019-11-09/prop-187-anniversary-commemoration> 4. <https://www.latimes.com/opinion/la-letter-to-the-editor-htmlstory.html> 5. <https://www.latimes.com/opinion/story/2019-11-09/facebook-twitter-political-ads-lies> 6. <https://www.latimes.com/opinion/story/2019-11-08/berlin-wall-30th-anniversary-cold-war-donald-trump> 7. <https://www.latimes.com/politics/story/2019-11-08/elizabeth-warren-used-bare-knuckle-tactics-to-take-down-an-obama-nominee> 8. <https://www.latimes.com/opinion/story/2019-11-09/elizabeth-warren-trump> 9. <https://www.latimes.com/opinion/story/2019-11-09/devin-nunes-impeachment-witness-list> 10. <https://www.latimes.com/opinion/story/2019-11-09/sat-uc-college-admissions> |



|  |  |
| --- | --- |
| 1. <https://www.latimes.com/espanol/internacional/articulo/2019-11-09/monos-se-volvieron-inmunes-tras-inocularles-virus-del-ebola-con-mutacion> 2. <https://www.latimes.com/science/sciencenow/la-sci-science-of-zika-five-ways-20160223-htmlstory.html> 3. <https://www.latimes.com/espanol/internacional/articulo/2019-11-09/monos-se-volvieron-inmunes-tras-inocularles-virus-del-ebola-con-mutacion> 4. <https://www.latimes.com/science/story/2019-10-21/virus-afm-illness-paralyzing-kids> 5. <https://www.latimes.com/science/story/2019-10-21/virus-afm-illness-paralyzing-kids> 6. <https://www.latimes.com/socal/daily-pilot/news/tn-dpt-me-west-nile-20170707-story.html> 7. <https://www.latimes.com/science/sciencenow/la-sci-sn-acute-flaccid-myelitis-polio-20181017-story.html> 8. <https://www.latimes.com/local/lanow/la-me-ln-virulent-newcastle-disease-outbreak-in-southern-california-20190607-story.html> 9. <https://www.latimes.com/world-nation/story/2019-09-04/anti-vaxxers-helping-polio-comeback-pakistan> 10. <https://www.latimes.com/local/obituaries/la-me-deborah-asnis-20150921-story.html> | 1. <https://www.latimes.com/terms-of-service> 2. <https://www.latimes.com/espanol/> 3. <https://www.latimes.com/science> 4. <https://www.latimes.com/espanol/internacional> 5. <https://www.latimes.com/espanol/deportes/articulo/2019-11-09/como-ser-campeon-de-league-of-legends> 6. <https://www.latimes.com/business/story/2019-10-13/body-parts-harvesting-hinders-coroner-autopsies> 7. <https://www.latimes.com/espanol/mexico/articulo/2019-11-09/crean-en-mexico-plataforma-para-medir-impacto-ambiental-de-basura-electronica> 8. <https://www.latimes.com/espanol/internacional/articulo/2019-11-09/el-nuevo-mitsubishi-mirage-pudiera-arribar-a-mediados-del-proximo-ano> 9. <https://www.latimes.com/espanol/internacional/articulo/2019-11-09/monos-se-volvieron-inmunes-tras-inocularles-virus-del-ebola-con-mutacion> 10. <https://www.latimes.com/espanol/internacional/articulo/2019-11-09/una-tortuga-tecnologica-para-detectar-microplasticos-en-los-oceanos> |



|  |  |
| --- | --- |
| 1. <https://www.latimes.com/sports/olympics/story/2019-08-15/2020-tokyo-olympics-searching-for-answers-amid-heat-wave> 2. <https://www.latimes.com/sports/olympics> 3. <https://www.latimes.com/people/david-wharton> 4. <https://www.latimes.com/people/david-wharton> 5. <https://www.latimes.com/sports/olympics/story/2019-10-16/2020-tokyo-olympics-marathon-sapporo-avoid-heat> 6. <https://www.latimes.com/sports/olympics/story/2019-07-31/deadly-heat-wave-2020-summer-olympics-tokyo> 7. <https://www.latimes.com/sports/story/2019-11-05/ioc-anti-doping-2020-tokyo-olympics> 8. <https://www.latimes.com/sports/olympics/story/2019-10-10/naomi-osaka-chooses-japan-over-u-s-2020-tokyo-olympics> 9. <https://www.latimes.com/sports/story/2019-07-23/tokyos-rough-road-to-2020-summer-olympics> 10. <https://www.latimes.com/people/david-wharton> | 1. <https://www.latimes.com/> 2. <https://www.latimes.com/sitemap> 3. <https://www.latimes.com/sports> 4. <https://www.latimes.com/opinion> 5. <https://www.latimes.com/business> 6. <https://www.latimes.com/entertainment-arts> 7. <https://www.latimes.com/world-nation> 8. <https://www.latimes.com/environment> 9. <https://www.latimes.com/newsroom-directory> 10. <https://www.latimes.com/politics> |



|  |  |
| --- | --- |
| 1. <https://www.latimes.com/business/story/2019-11-06/uber-lock-up-period-ends-with-falling-stock-protests> 2. <https://www.latimes.com/business/la-fi-lyft-stock-20190401-story.html> 3. <https://www.latimes.com/business/la-fi-hy-tesla-stock-20160518-snap-story.html> 4. <https://www.latimes.com/business/story/2019-11-06/uber-lock-up-period-ends-with-falling-stock-protests> 5. <https://www.latimes.com/business/story/2019-11-09/stock-managers-who-played-defense-in-2019-are-left-scrambling-to-make-up-ground> 6. <https://www.latimes.com/business/la-fi-uber-ipo-stock-trading-price-20190510-story.html> 7. <https://www.latimes.com/business/story/2019-09-09/at-t-stock-surges-elliott-reveals-3-2-billion-stake> 8. <https://www.latimes.com/business/la-fi-stock-market-status-update-20180209-story.html> 9. <https://www.latimes.com/business/la-fi-stock-market-correction-20180212-story.html> 10. <https://www.latimes.com/business/story/2019-11-09/stock-managers-who-played-defense-in-2019-are-left-scrambling-to-make-up-ground> | 1. <https://www.latimes.com/> 2. <https://www.latimes.com/business> 3. <https://www.latimes.com/business/real-estate> 4. <https://www.latimes.com/business/real-estate/story/2019-11-09/hulu-looking-for-alaska-landry-bender> 5. <https://www.latimes.com/business/story/2019-11-08/trump-china-tariffs> 6. <https://www.latimes.com/business/real-estate/story/2019-11-09/hot-property-very-brady-sale-malibu> 7. <https://www.latimes.com/politics/story/2019-11-08/elizabeth-warren-used-bare-knuckle-tactics-to-take-down-an-obama-nominee> 8. <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> 9. <https://www.latimes.com/business/story/2019-11-09/stock-managers-who-played-defense-in-2019-are-left-scrambling-to-make-up-ground> 10. <https://www.latimes.com/business/story/2019-11-08/gap-ceo-was-done-in-by-fashion-missteps-and-fading-brands> |