

CSCI 572 Information Retrieval and Web Search Engines

Homework 4: Indexing the Web using Solar

Name : Anubhav Jindal

USC ID : 5963113610

Steps followed to complete the Assignment:

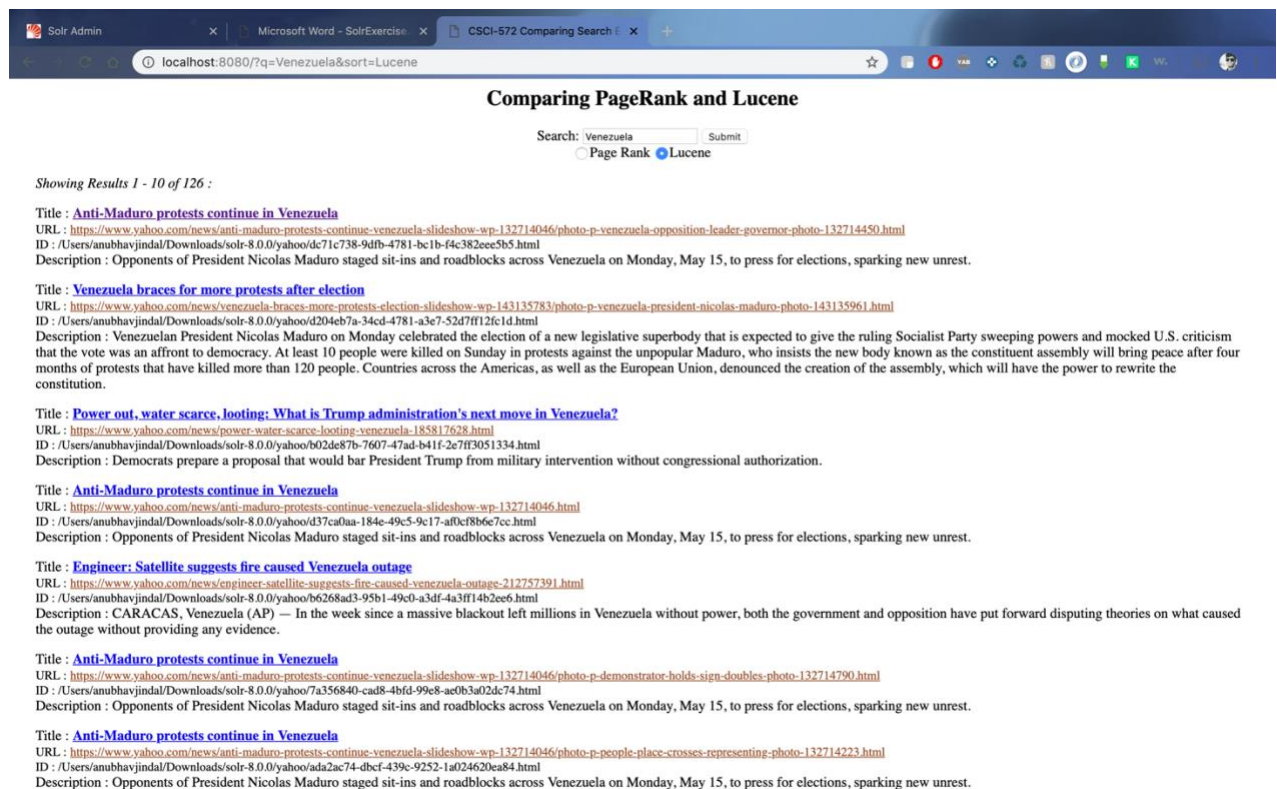
1. Installations and Environment Setup
 - a. I am using a Macbook machine which has a Unix based Operating System, so I didn't install Ubuntu.
 - b. Solr installation:
 - i. Downloaded the Solr Binary release as a zip file from <http://www.apache.org/dyn/closer.lua/lucene/solr/8.0.0/solr-8.0.0.zip>
 - ii. Unzipped the zipped file which gave me a solr-8.0.0 directory
 - c. Downloaded Yahoo html files obtained from a crawl that provided to us as a Google drive link and placed them in the solr-8-0-0 directory.
 - d. Solr Setup:
 - i. Opened Terminal and change the current directory to **solr-8.0.0**
 - ii. Executed **bin/solr start** to start solr
 - iii. Executed **bin/solr create -c csci572** to create a core.
 - iv. Executed **bin/post -c myexample -filetypes html yahoo/** to index html files in the yahoo folder
 - v. Accessed <http://localhost:8983/solr/> and verified the working of solr.
 - vi. Ensured that the **<str name="df">_text_</str>** line is uncommented in solrconfig.xml
2. Edgelist computation:
 - a. Wrote a program in Java using **JSoup Library** and HashMaps to compute the Edgelist from the provided Yahoo html pages.
 - b. The Edgelist was written to a **edgeList.txt** file for further use.
3. PageRank computation:
 - a. **NetworkX** was already installed on my machine as part of Anaconda framework.
 - b. Wrote a python script which used NetworkX to compute the PageRanks.
 - c. The **edgeList.txt** and **DiGraph** were used to compute PageRanks.
 - d. The parameters used were as follows:
 - i. **alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None**
 - e. The results of the script were stored in **external_pageRankFile.txt** for further use.
4. Setting up Solr for PageRank algorithm compatibility
 - a. Copied external_pageRankFile.txt to solr-8.0.0/server/solr/csci572/data/
 - b. Modified managed-schema for using pageRankFile
 - i. **<FieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" valType="pfloat" />**
 - ii. **<field name="pageRankFile" type="external" stored="false" indexed="false" />**
 - c. Modified solarconfig.xml file to reload external file
 - i. **<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />**
 - ii. **<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />**
5. Reloaded Solr Interface using the reload button in the Core Admin section.
6. For using PageRank algorithm instead of default algorithm use the **parameter: pageRankFile desc**
7. Created **Custom Interface** to the solr search engine using **solr-php-client, PHP, HTML and CSS**.
8. Ran the queries as described in the homework and computed the overlap.

SCREENSHOTS

- a. Screenshot of initial page where I enter the query in input box



- b. Page which shows the results for Lucene(Default).



c. Page which shows the results for PageRank

Solr Admin | Microsoft Word - SolrExercise | CSCI-572 Comparing Search E | +

localhost:8080/?q=Venezuela&sort=pagerank

Comparing PageRank and Lucene

Search: Venezuela Submit

☒ Page Rank ☐ Lucene

Showing Results 1 - 10 of 126 :

Title : [Science | Yahoo News - Latest News & Headlines](#)
URL : <https://www.yahoo.com/news/science/>
ID : /Users/anubhavjindal/Downloads/solr-8.0.0/yahoo/73f85d91-7903-4110-a92f-90fa0380179.html
Description : The latest news and headlines from Yahoo! News. Get breaking news stories and in-depth coverage with videos and photos.

Title : [How a CIA agent trained agents who found Soviet missiles in Cuba](#)
URL : <https://www.yahoo.com/news/operation-cobra-untold-story-cia-officer-trained-network-agents-found-soviet-missiles-cuba-100005794.html>
ID : /Users/anubhavjindal/Downloads/solr-8.0.0/yahoo/fa310394-6c90-422f-ba1c-36fd81f4f5a6.html
Description : The canoe had been the idea of Tom Hewitt, the agents' case officer in the CIA's huge Miami station, where he waited for word of the mission. It would be Hewitt's job to guide their actions from afar, now that they were back in their homeland.

Title : [PHOTOS: Malnourished Venezuelans hope urgently needed aid arrives soon](#)
URL : <https://www.yahoo.com/photos/malnourished-venezuelans-hope-urgently-needed-aid-arrives-soon-184505075.html>
ID : /Users/anubhavjindal/Downloads/solr-8.0.0/yahoo/55584706-e01a-4d86-a574-f0e7097be56.html
Description : Yaneidi Guzman has lost a third of her weight over the past three years as Venezuela's economic collapse made food unaffordable and she now hopes the opposition will succeed in bringing urgently needed foreign aid to the South American country.

Title : [Beto O'Rourke jumps into 2020 presidential race](#)
URL : <https://www.yahoo.com/news/beto-o-rourke-jumping-into-2020-presidential-race-235309920.html>
ID : /Users/anubhavjindal/Downloads/solr-8.0.0/yahoo/0468db21-1d76-410d-830c-2fb2caff49c.html
Description : The former Texas congressman announced his candidacy on Thursday.

Title : [Trump, a self-proclaimed champion of Americans detained abroad, quiet on former Marine held in Russia](#)
URL : <https://www.yahoo.com/news/trump-a-self-proclaimed-champion-of-americans-detained-abroad-quiet-on-former-marine-held-in-russia-130009990.html>
ID : /Users/anubhavjindal/Downloads/solr-8.0.0/yahoo/396e8fca-8c36-44a9-a03c-e74d31fb0bb0.html
Description : David Whelan, a lawyer who specializes in information and technology, was diving into books about Russian negotiation tactics and using Google Translate to glean clues about how his twin brother got wrapped up in what he views as a fabricated spying scandal. Meanwhile, Paul Whelan turned a year older

Title : [The top 10 undercover stories of 2016](#)
URL : <https://www.yahoo.com/news/the-top-10-undercover-stories-of-2016-100099854.html>
ID : /Users/anubhavjindal/Downloads/solr-8.0.0/yahoo/739d4318-1147-4165-bcc4-0b4aa3da4c8f.html
Description : Big stories dominated the news in 2016 like no other year: Donald Trump won. The U.K. backed the Brexit. Right-wing populism surged around the globe. And more.

Title : [Venezuela opposition calls nationwide strike against Maduro](#)
URL : <https://www.yahoo.com/news/maduro-urges-peaceful-opposition-vote-venezuela-012524217.html>
ID : /Users/anubhavjindal/Downloads/solr-8.0.0/yahoo/821c7bad-81b8-433d-8f0b-67c8b4021a2c.html
Description : Venezuela's opposition called a nationwide strike for Thursday to press President Nicolas Maduro to back off a rewriting of the constitution, ratcheting up tensions after an unofficial vote rejecting Maduro's plan and amid months of deadly protests. The strike call, issued on Monday, was part

d. The actual webpage that opens when clicking on one of the result's link

Solr Admin | Microsoft Word - SolrExercise | CSCI-572 Comparing Search | + | Anti-Maduro protests continue

https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-venezue...

Home Mail News Finance Sports Entertainment Search Mobile More

YAHOO! Search Search News Search web Sign in 54 Mail

Anti-Maduro protests continue in Venezuela

YAHOO! news Yahoo News Photo Staff · May 16, 2017



2 / 21

Anti-Maduro protests continue in Venezuela

Venezuela's opposition leader and Governor of Miranda state Henrique Capriles (L) poses for a picture with supporters during a blockade in an avenue while rallying against President Nicolas Maduro in Caracas, Venezuela, May 15, 2017. (Marco Bello/Reuters)

Waiting for www.yahoo.com...

EXPLANATION WHY SOME PAGES HAVE HIGHER PAGERANK THAN OTHERS

PageRank is an algorithm that is used to rank webpages in order to analyze which webpages are more important than the others. It is computed based on the number of links to a particular webpage (in-links) and the number of links out from the webpage (out-links). The more in-links a webpage has, the higher is the PageRank, the lesser out-links a page has, more is the PageRank. Also, higher the PageRank of the website which contributes to the in-link, higher will be the PageRank. Due to these reasons, some pages have higher PageRank than other pages.

QUERY RESULTS

Venezuela	
LUCENE	PAGERANK
https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-venezuela-opposition-leader-governor-photo-132714450.html	https://www.yahoo.com/news/science/
https://www.yahoo.com/news/venezuela-braces-more-protests-election-slideshow-wp-143135783/photo-p-venezuela-president-nicolas-maduro-photo-143135961.html	https://www.yahoo.com/news/operation-cobra-untold-story-cia-officer-trained-network-agents-found-soviet-missiles-cuba-100005794.html
https://www.yahoo.com/news/power-water-scarce-looting-venezuela-185817628.html	https://www.yahoo.com/news/photos-malnourished-venezuelans-hope-urgently-needed-aid-arrives-soon-184505075.html
https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046.html	https://www.yahoo.com/news/beto-o-rourke-jumping-into-2020-presidential-race-235309920.html
https://www.yahoo.com/news/engineer-satellite-suggests-fire-caused-venezuela-outage-212757391.html	https://www.yahoo.com/news/trump-a-selfproclaimed-champion-of-americans-detained-abroad-quiet-on-former-marine-held-in-russia-130000990.html
https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-demonstrator-holds-sign-doubles-photo-132714790.html	https://www.yahoo.com/news/the-top-10-undercovered-stories-of-2016-100009854.html
https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-people-place-crosses-representing-photo-132714223.html	https://www.yahoo.com/news/maduro-urges-peaceful-opposition-vote-venezuela-012524217.html
https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-opposition-supporters-carry-materials-photo-132714045.html	https://www.yahoo.com/news/u-top-court-backs-trump-travel-ban-targeting-142429044.html
https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-opposition-supporters-carry-materials-photo-132714045.html	https://www.yahoo.com/news/cias-communications-suffered-catastrophic-compromise-started-iran-090018710.html

132714046/photo-p-demonstrator-boy-wearing-helmet-photo-132714648.html	
https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-demonstrator-sits-discarded-stove-photo-132714988.html	https://www.yahoo.com/news/close-house-race-republican-karen-handel-gets-help-paul-ryan-220451019.html

Senate	
LUCENE	PAGERANK
https://www.yahoo.com/news/senate-skinny-repeal-obamacare-falls-apart-senate-floor-mccain-defects-072042708.html	https://www.yahoo.com/news/
https://www.yahoo.com/news/senate-judiciary-committee-vote-kavanaugh-slideshow-wp-170920887/photo-p-senate-judiciary-committee-chairman-photo-170920546.html	https://www.yahoo.com/news/originals/
https://www.yahoo.com/news/senate-judiciary-committee-vote-kavanaugh-slideshow-wp-170920887/photo-p-senate-judiciary-committee-member-photo-170920673.html	https://www.yahoo.com/news
https://www.yahoo.com/news/senate-judiciary-committee-vote-kavanaugh-slideshow-wp-170920887/photo-p-senate-judiciary-committee-members-photo-170920113.html	https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-herescia-stop-100048764.html
https://www.yahoo.com/news/jeff-session-testifies-senate-hearing-slideshow-wp-191620787/photo-p-senate-intelligence-committee-chairman-photo-191620949.html	https://www.yahoo.com/news/politics/
https://www.yahoo.com/news/protesters-across-country-oppose-gops-slideshow-wp-133309283/photo-p-senate-democrats-gather-senate-photo-150009242.html	https://www.yahoo.com/news/tagged/skulduggery
https://www.yahoo.com/news/kavanaugh-ford-testify-senate-judiciary-slideshow-wp-145913238/photo-p-senate-judiciary-committee-member-photo-213813950.html	https://www.yahoo.com/news/topics
https://www.yahoo.com/news/paul-ryan-premature-say-whether-house-pass-senate-healthcare-bill-174352395.html	https://www.yahoo.com/news/habitat-for-sale-an-oil-and-gas-group-calls-the-tune-at-the-interior-department-215244528.html
https://www.yahoo.com/news/cbo-likely-say-millions-lose-coverage-senate-health-care-bill-215135603.html	https://www.yahoo.com/news/maverick-republican-william-weld-looking-run-malignant-narcissist-trump-142709863.html
https://www.yahoo.com/news/trump-senate-republicans-look-like-fools-dont-dump-60-vote-rule-135937103.html	https://www.yahoo.com/news/politics

Democrats	
LUCENE	PAGERANK
https://www.yahoo.com/news/midterms/house	https://www.yahoo.com/news/
https://www.yahoo.com/news/midterms/house/seat	https://www.yahoo.com/news/originals/
https://www.yahoo.com/news/democrats-rolling-2018-midterms-message-next-week-191517316.html	https://www.yahoo.com/news
https://www.yahoo.com/news/will-democrats-filibuster-gorsuch-republicans-dont-think-so-111241581.html	https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html
https://www.yahoo.com/news/white-house-tries-tag-democrats-blame-possible-shutdown-203618659.html	https://www.yahoo.com/news/ann-coulter-lunatic-trump-challenged-2020-right-165049018.html
https://www.yahoo.com/news/democrats-hope-better-deal-message-will-resonate-190849661.html	https://www.yahoo.com/news/politics/
https://www.yahoo.com/news/democrats-2018-many-candidates-many-candidates-162905320.html	https://www.yahoo.com/news/habitat-for-sale-an-oil-and-gas-group-calls-the-tune-at-the-interior-department-215244528.html
https://www.yahoo.com/news/democrats-vow-fight-lifetime-keep-brett-kavanaugh-off-supreme-court-051334537.html	https://www.yahoo.com/news/maverick-republican-william-weld-looking-run-malignant-narcissist-trump-142709863.html
https://www.yahoo.com/news/mcconnell-plan-fails-will-democrats-republicans-work-together-fix-health-care-220450895.html	https://www.yahoo.com/news/politics
https://www.yahoo.com/news/democrats-wrest-power-away-republicans-state-level-183309710.html	https://www.yahoo.com/news/democrat-predicts-house-will-get-trumps-tax-returns-complications-142301949.html

Republicans	
LUCENE	PAGERANK
https://www.yahoo.com/news/midterms/house	https://www.yahoo.com/news/
https://www.yahoo.com/news/midterms/house/seat	https://www.yahoo.com/news/tagged/photos/
https://www.yahoo.com/news/mcconnell-plan-fails-will-democrats-republicans-work-together-fix-health-care-220450895.html	https://www.yahoo.com/news/originals/
https://www.yahoo.com/news/grim-warnings-white-house-republicans-ahead-election-171354295--election.html	https://www.yahoo.com/news
https://www.yahoo.com/news/democrats-wrest-power-away-republicans-state-level-183309710.html	https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html

https://www.yahoo.com/news/republicans-rush-obamacare-repeal-and-replace-through-house-committees-233538837.html	https://www.yahoo.com/news/politics/
https://www.yahoo.com/news/gop-obamacare-replacement-narrowly-passes-out-of-committee-after-3-republicans-defect-145519718.html	https://www.yahoo.com/news/tagged/when%20presidents%20lead/
https://www.yahoo.com/news/will-democrats-filibuster-gorsuch-republicans-dont-think-so-111241581.html	https://www.yahoo.com/news/tagged/skullduggery
https://www.yahoo.com/news/house-republicans-reveal-obamacare-replacement-bill-013510100.html	https://www.yahoo.com/news/world/
https://www.yahoo.com/news/down-ticket-5-republicans-finally-000000015.html	https://www.yahoo.com/news/science/

Patriot Movement	
LUCENE	PAGERANK
https://www.yahoo.com/news/jeffersons-tree-liberty-blood-schoolchildren-171918999.html	https://www.yahoo.com/news
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-wing-supporter-patriot-prayer-photo-212529706.html	https://www.yahoo.com/news/tagged/through-her-eyes
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-wing-supporter-patriot-prayer-photo-212529025.html	https://www.yahoo.com/news/crackdown-opioids-victims-people-need-live-100058361.html
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-counter-protester-argues-police-photo-212529993.html	https://www.yahoo.com/news/admissions-scam-adds-insult-injury-minority-applicants-202405114.html
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-left-wing-counter-protesters-photo-212529339.html	https://www.yahoo.com/news/death-penalty-outlawed-203116351.html
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-far-protester-receives-treatment-photo-212529076.html	https://www.yahoo.com/news/ilhan-omar-and-andre-carson-respond-to-horror-of-new-zealand-massacre-of-muslims-171412117.html
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-counter-demonstrators-held-back-photo-212529207.html	https://www.yahoo.com/news/latest-students-delhi-protest-cleaner-143006709.html
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-	https://www.yahoo.com/news/photos-in-gaza-women-walk-thin-line-between-hope-and-despair-132405730.html

212529094/photo-p-police-push-protesters-back-photo-212529889.html	
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094.html	https://www.yahoo.com/news/2016-pictures-news-slideshow-wp-100040397.html
https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-counter-protester-raises-arms-photo-212529025.html	https://www.yahoo.com/news/trump-mocks-beto-o-rourkes-crazy-hand-gestures-183443014.html

Oscar 2019	
LUCENE	PAGERANK
https://www.yahoo.com/news/spike-lee-pushes-back-trump-spat-continues-oscar-speech-162915413.html	https://www.yahoo.com/news/
https://www.yahoo.com/news/skulduggery-power-vice-000835154.html	https://www.yahoo.com/news/tagged/photos/
https://www.yahoo.com/news/topics/star-wars	https://www.yahoo.com/news/originals/
https://www.yahoo.com/news/trump-bashes-spike-lee-racist-hit-oscars-123716258.html	https://www.yahoo.com/news
https://www.yahoo.com/news/topics/royal-family	https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-herescia-stop-100048764.html
https://www.yahoo.com/news/kamala-harris-trump-racist-185542673.html	https://www.yahoo.com/news/ann-coulter-lunatic-trump-challenged-2020-right-165049018.html
https://www.yahoo.com/news/the-names-behind-the-college-cheating-charges-141742295.html	https://www.yahoo.com/news/politics/
https://www.yahoo.com/news/topics/technology	https://www.yahoo.com/news/tagged/when%20presidents%20lead/
https://www.yahoo.com/news/gore-us-getting-close-political-shift-climate-change-223933667--politics.html	https://www.yahoo.com/news/us-grounds-boeing-737-max-amid-growing-safety-200009938.html
https://www.yahoo.com/news/heres-know-orlando-victims-000000001.html	https://www.yahoo.com/news/tagged/skulduggery

Channel	
LUCENE	PAGERANK
https://www.yahoo.com/news/topics/fox-news-channel	https://www.yahoo.com/news/topics
https://www.yahoo.com/news/topics/cnn	https://www.yahoo.com/news/crackdown-opioids-victims-people-need-live-100058361.html

https://www.yahoo.com/news/service-charlottesville-victim-heather-heyer-call-righteous-action-172114070.html	https://www.yahoo.com/news/operation-cobra-untold-story-cia-officer-trained-network-agents-found-soviet-missiles-cuba-100005794.html
https://www.yahoo.com/news/topics/megyn-kelly	https://www.yahoo.com/news/skullduggery-washington-think-tank-says-no-saudi-funds-124450269.html
https://www.yahoo.com/news/topics/marijuana-legalization	https://www.yahoo.com/news/topics/president-trump
https://www.yahoo.com/news/topics/space-exploration	https://www.yahoo.com/news/rory-mcilroy-blames-pga-tour-slow-play-epidemic-004543461.html
https://www.yahoo.com/news/topics/health	https://www.yahoo.com/news/the-names-behind-the-college-cheating-charges-141742295.html
https://www.yahoo.com/news/topics/sex-crimes	https://www.yahoo.com/news/death-toll-rises-nine-collapsed-lagos-school-building-074812489.html
https://www.yahoo.com/news/topics/sanctuary-cities	https://www.yahoo.com/news/topics/jared-kushner
https://www.yahoo.com/news/topics/adam-rippon	https://www.yahoo.com/news/fire-trump-says-misspoke-russian-interference-u-s-election-192113980.html

Wall	
LUCENE	PAGERANK
https://www.yahoo.com/news/wall-wall-trump-defends-border-plan-kelly-suggests-changed-145144558.html	https://www.yahoo.com/news/
https://www.yahoo.com/news/imagining-trumps-big-beautiful-wall-might-183350584.html	https://www.yahoo.com/news/tagged/photos/
https://www.yahoo.com/news/gop-leaders-back-at-least-12-billion-for-border-wall-153042941.html	https://www.yahoo.com/news/originals/
https://www.yahoo.com/news/trumps-big-beautiful-wall-collides-congress-102804089.html	https://www.yahoo.com/news
https://www.yahoo.com/news/border-wall-funding-threatens-government-shutdown-budget-battle-heats-180534239.html	https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html
https://www.yahoo.com/news/on-the-border-waiting-for-the-big-beautiful-wall-225959295.html	https://www.yahoo.com/news/ann-coulter-lunatic-trump-challenged-2020-right-165049018.html
https://www.yahoo.com/news/sanders-calls-obamas-400k-wall-street-speaking-fee-unfortunate-150857262.html	https://www.yahoo.com/news/politics/

https://www.yahoo.com/news/steve-king-has-a-model-of-the-border-wall-he-wants-to-build-to-protect-our-superior-civilization-183248446.html	https://www.yahoo.com/news/tagged/when%20presidents%20lead/
https://www.yahoo.com/news/rebuilding-great-wall-china-slideshow-wp-163045771/photo-p-people-rest-working-reconstruction-photo-153045115.html	https://www.yahoo.com/news/tagged/skullduggery
https://www.yahoo.com/news/rebuilding-great-wall-china-slideshow-wp-163045771/photo-p-people-wait-bricks-other-photo-153045472.html	https://www.yahoo.com/news/world/

OVERLAPS

QUERY	NO. OF OVERLAPS
Venezuela	0
Senate	0
Democrats	0
Republicans	0
Patriot Movement	0
Oscar 2019	0
Channel	0
Wall	0

There were no overlaps in any query. So, no bar graph has been drawn as per the instructions given.