Hw4 Report

Bei Gao 6863049908

1. Steps you followed to complete this assignment.

Environment installation and set up

- Install VirtualBox on Windows 10, and install Ubuntu 16.04.2 on VirtualBox
- Install Apache2 and PHP using 'apt-get' tool
- install Java 8 (also export the path) and Eclipse & add jsoup library to my Java project'
- unzip Solr-6.5.0 to 'home' directory & start Apache2 and Solr.
- Install solr-php-client in my Apache web root directory 'apacheroot'.
- Install Python 2.7 (built-in), easy_install tool & install networkx-1.11 using easy_install

Indexing html files and initial search

- Create a core called 'myhw' in Solr installation directory in terminal.
- Index 19086 html files from the LATimes site using 'post' to 'myhw' core in terminal.
- Create initial web server for querying those files in PHP—'wel.php' in 'gedit' editor.

Ps: I pre-processed the possible ',' in CSV file and output a JSON file with same content.

- Search the 8 queries using Solr default ranking method in web browser, url:http://localhost/wel.php

Generate external pagerank file

- Extract outgoing links using **jsoup** in Java for 19086 html files from LATimes site and outputs a text file 'edgeList.txt' which contains all encoding link pairs in which one points to another within 19086 files. The format of text file is 'linkcode1 linkcode2' where link1 points to link2. The IDE for Java is Eclipse.
- Given 'edgeList.txt', build a correlated graph using **networkx** in Python, 'gedit' editor.
- Given the graph, generate PageRank scores for each node (link) using **networkx** in Python, 'gedit' editor, and outputs a text file 'external_pageRankFile.txt' ('doc ID'='score') .

Configuration of PageRank algorithm:

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

Configuration in Solr for pagerank file

- Place 'external_pagenRankFile.txt' in the data folder of 'myhw' directory
- Modify 'managed-schema' file and 'solrconfig.xml' file in config directory in 'myhw' core
- Reload 'myhw' core

Modify PHP web server file and Search using PageRank

- Modify my web server file 'wel.php'

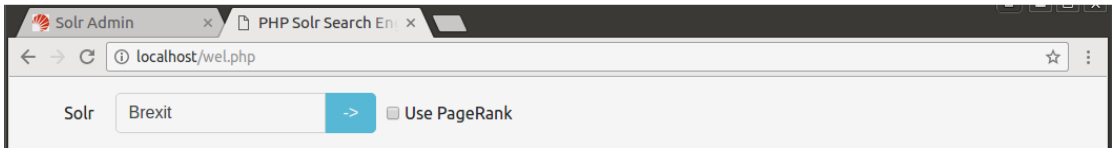
①refine the UI using HTML, CSS, Bootstrap

②add a checkbox for users to choose whether they want to use PageRank algorithm

- Search the 8 queries again using PageRank. Results will contain title, url, id and description for each record. url:http://localhost/wel.php

Analyze the results and write a report

2. Top 10 results generated by 2 ranking methods (I use ‘Brexit’ as example here)
   UI in query box



   Solr default

Solr Brexit -> Use PageRank

Results 1 - 10 of 91:

Brexit Articles, Photos, and Videos - Los Angeles Times
http://www.latimes.com/topic/politics-government/brexit-EVGAP00091-topic.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/39a7634a-199b-45a7-8ec0-4cd68eb36a1c.html
Description: News, Photos and Information about Brexit

Brexit leads IMF to cut forecast for global economy - LA Times
http://www.latimes.com/business/la-1468882806-snap-photo-photo.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/ebd91d9c-1e87-4c3f-8e77-2b665c3824e3.html
Description: Pro-European Union supporters and pro-"Brexit" supporters demonstrate in London's Green Park on July 9.

British scientists are freaking out about 'Brexit' too - LA Times
http://www.latimes.com/science/sciencenow/la-sci-sn-brexit-science-20160624-snap-story.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/78ebd410-efa6-4a45-a4c4-b5eaba56dfde.html
Description: Scientists in the U.K. were overwhelmingly opposed to 'Brexit.' Now that it's happening, here's what they want politicians to keep in mind.

Brexit - LA Times
http://www.latimes.com/travel/deals/la-1467744419-snap-photo-photo.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/e1d8349d-303b-4650-b62f-9da84d7a9498.html
Description: The vote to leave the European Union has weakened the British pound against the U.S. dollar, giving travelers more for their money.

'Brexit' roils currency markets - LA Times
http://www.latimes.com/business/la-1467824818-snap-photo-photo.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/19793351-63e0-4dea-8ced-1a56ea4c70f1.html
Description: The dollar looks to gain strength against the euro and pound in the wake of Britain's vote to leave the European Union.

Brexit sign - LA Times
http://www.latimes.com/world/la-brdelossantos-1485251826-snap-photo-photo.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/bf11ea70-7d9a-458d-881b-48be6f3deb51.html
Description: Britain's Supreme Court ruled Tuesday that the government must receive Parliament's approval to trigger the process of taking Britain out of the European Union.

Secretary of State John Kerry attempts to soothe 'Brexit' fears in Europe - LA Times
http://www.latimes.com/world/la-fg-brexit-updates-secretary-of-state-john-kerry-attempts-1467127537-htmlstory.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/c4e94f55-e2e5-4704-af20-cda2bcfe8a1f.html
Description: America's top diplomat sought to soothe fears Monday on both sides of the Atlantic as aftershocks of Great Britain's vote to withdraw f...

Britain votes to leave the European Union - LA Times
http://www.latimes.com/world/europe/87674965-132.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/217d7d7e-ebf8-4384-89c2-7139d2aa2c77.html
Description: Britons voted Thursday to leave the 28-nation European Union, a historic vote that sent shock waves across the continent and prompted Prime Minister David

British Prime Minister May lays out plans for a 'stronger, fairer, more united' country after 'Brexit' - LA Times
http://www.latimes.com/world/la-fg-theresa-may-speech-20170117-story.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/b41f1263-0f37-4140-afaa-9eb34e0f00b2.html
Description: British Prime Minister Theresa May says there will be no 'half in, half out' deal in the nation's exit from the European Union.

Essential Arts & Culture: A controversial Santa Monica play, a terrifying architectural slide, how 'Brexit' affects the arts - LA Times
http://www.latimes.com/entertainment/arts/miranda/la-et-cam-newsletter-essential-arts-and-culture-20160624-snap-story.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/7557772c-8c0d-4c8d-a220-d9188a47989f.html
Description: The most important Arts & Culture news from the Los Angeles Times

PageRank

Solr Brexit -> Use PageRank

Results 1 - 10 of 91:

European Union Articles, Photos, and Videos - Los Angeles Times
http://www.latimes.com/topic/business/economy/european-union-ORGOV000067-topic.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/b3683e60-3c9a-420b-8ff6-e09c14d662b9.html
Description: News, Photos and Information about European Union

EUROPE - Los Angeles Times
http://www.latimes.com/world/europe/
ID: /home/bei/solr-6.5.0/../shared/LATimes/39f138a5-53be-4ce0-9536-b48d6af324e5.html
Description: Breaking news and reporting from our European bureaus. Find what's happening overseas here.

John Kerry Articles, Photos, and Videos - Los Angeles Times
http://www.latimes.com/topic/politics-government/government/john-kerry-PEPLT003513-topic.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/e881192a-2d82-48ca-98f9-547123a35a47.html
Description: News, Photos and Information about John Kerry

Theater Reviews - Los Angeles Times
http://www.latimes.com/entertainment/arts/theater/reviews/
ID: /home/bei/solr-6.5.0/../shared/LATimes/e37c7a07-2752-4d5e-94bf-a33bad0a3b50.html
Description: Comprehensive coverage of Los Angeles theater reviews and dance productions. Read play reviews from the L.A. Times.

Updates on California politics: Lawmakers send gun measures to Gov. Brown, initiative on parole overhaul makes the Nov. 8 ballot - LA Times
http://www.latimes.com/politics/la-pol-sac-essential-politics-20160601-htmlstory.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/34d639d9-83db-4800-accb-38873f52f893.html
Description: Updates on California politics: Lawmakers send gun measures to Gov. Brown, initiative on parole overhaul makes the Nov. 8 ballot June 30, 2016, 5:45 p.m. Welcome to Essential Politics, our daily feed on California government and politics news. Here's what we're watching: Legislators sent Gov. Jerry...

Brandpoint - Travel - Los Angeles Times
http://www.latimes.com/bp/travel/
ID: /home/bei/solr-6.5.0/../shared/LATimes/460edc6c-8115-4272-8348-639426b6e347.html
Description: None

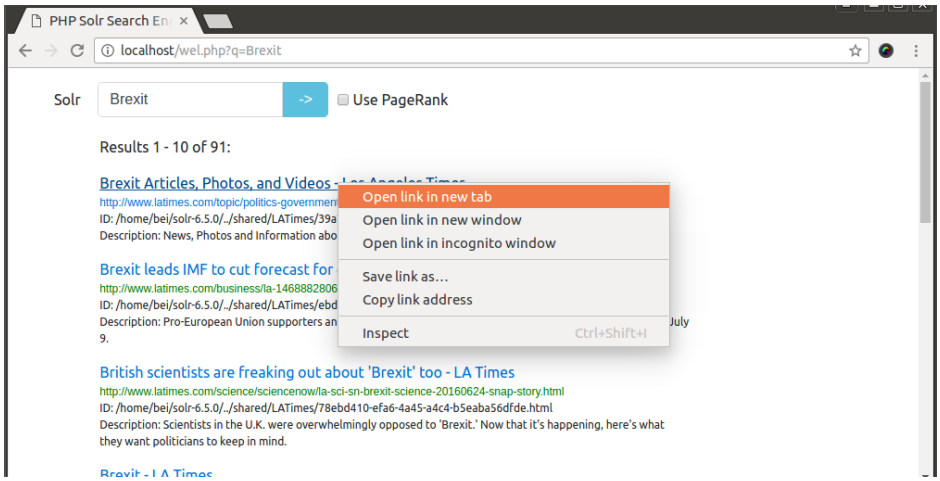
Kyle Kim - LA Times
http://www.latimes.com/la-bio-kyle-kim-staff.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/174f373c-e646-434f-b7c8-2e32ad438eb2.html
Description: Kyle Kim is a graphics and data journalist for the Los Angeles Times. Previously, Kyle braved the long New England winters as deputy news and social media editor for GlobalPost. He's a self-admitted Swedophile, resident tortured artist, and Sriracha-obsessed.

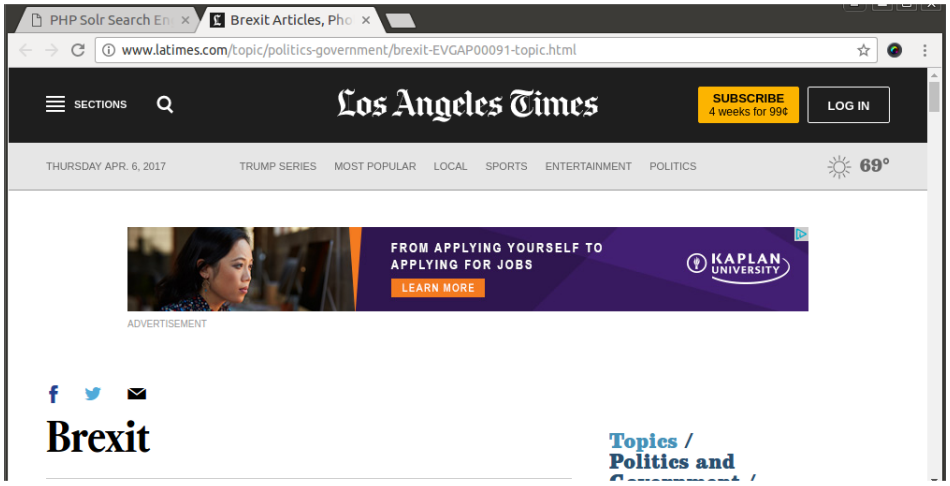
Peter Capaldi Articles, Photos, and Videos - Los Angeles Times
http://www.latimes.com/topic/entertainment/peter-capaldi-PECLB0000007802-topic.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/f1227a11-05b0-4685-8877-eff90cf9b900.html
Description: News, Photos and Information about Peter Capaldi

Christopher Knight Articles, Photos, and Videos - Los Angeles Times
http://www.latimes.com/topic/arts-culture/journalism/christopher-knight-PECLB00015016-topic.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/3dd3eb29-0a96-4bb4-8a4e-01660ed09496.html
Description: News, Photos and Information about Christopher Knight

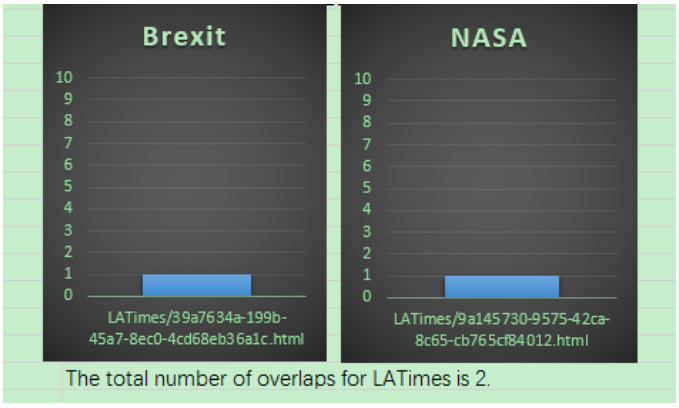
Brexit Articles, Photos, and Videos - Los Angeles Times
http://www.latimes.com/topic/politics-government/brexit-EVGAP00091-topic.html
ID: /home/bei/solr-6.5.0/../shared/LATimes/39a7634a-199b-45a7-8ec0-4cd68eb36a1c.html
Description: News, Photos and Information about Brexit

Click one link of results





3. Overlap graph showing the amount of overlap between the two ranking methods for each query



4. Explanation regarding why some pages have higher PageRank values

Because perhaps there are more internal links (maybe with high scores) pointing to them and/or less outgoing links compared to other pages. According to the definition of PageRank value, the PR value of each page depends on the PR of the pages pointing to it and the outgoing links the page has.