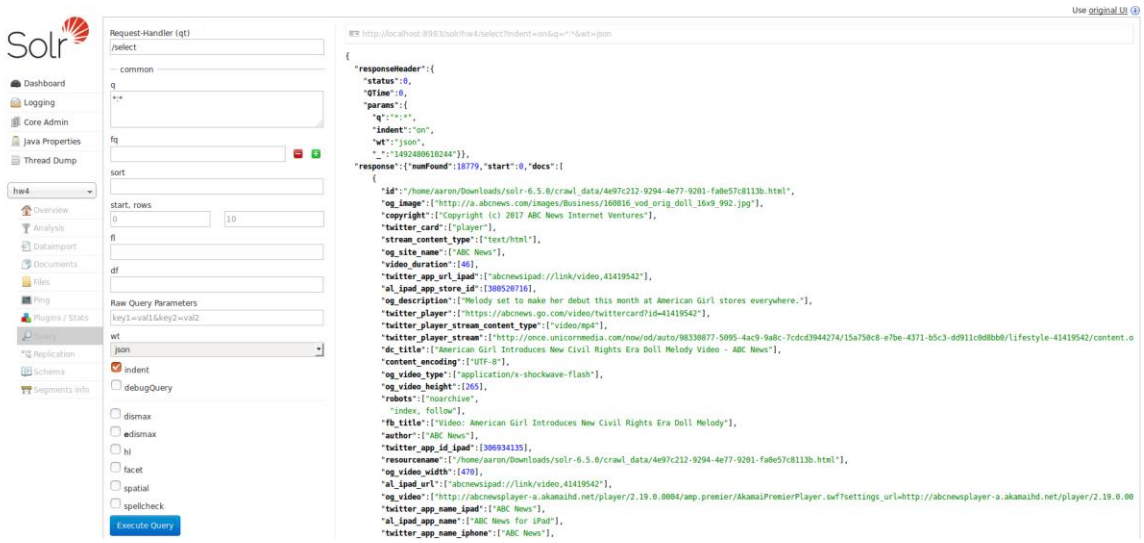
CSCI572 Homework 4 Report
Name: Shurui Liu, 9999-9586-70

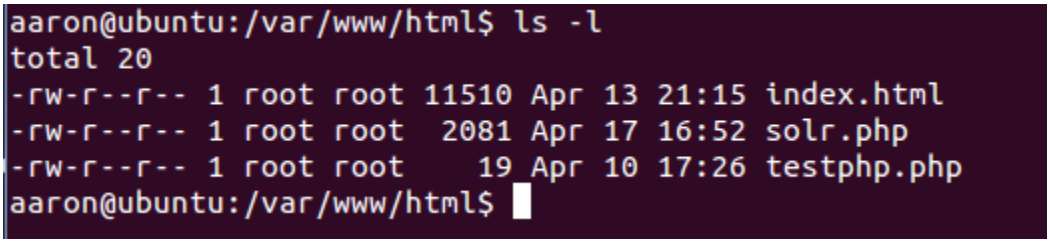
I. Steps

1. The first step is to use Solr to index all the HTML files that assigned to me. I am responsible for **ABCNews**. This screenshot shows you there are **18779** files are indexed by Solr.

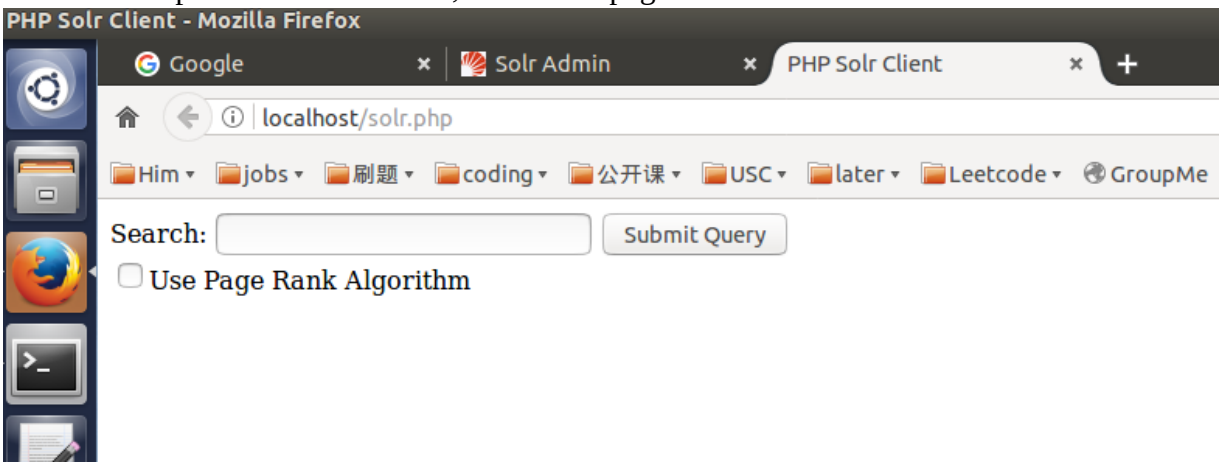


Before indexing these files, I made proper changes in **managed-schema** and **solrconfig.xml** based on the tutorial we got from 572 website.

2. The second step is setup the PHP page on my Apache server. I place my code, **solr.php**, at this location like the screenshot below. **/var/www/html**



When both Apache and Solr are on, I have this page below.

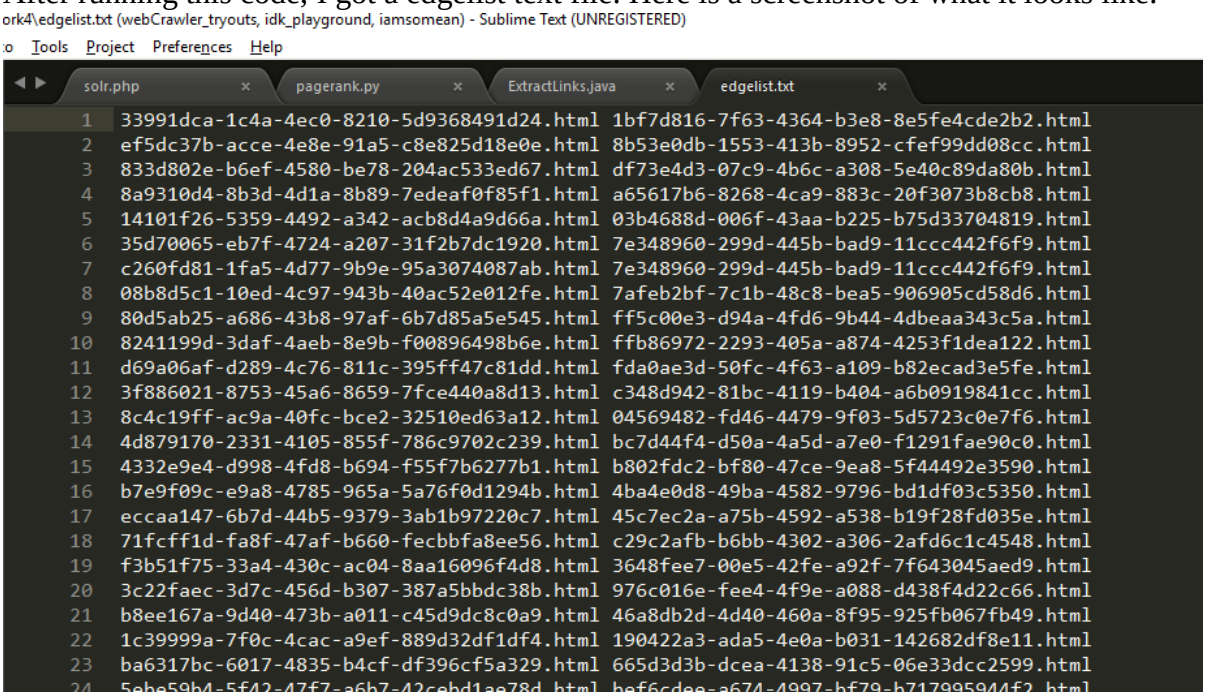


3. Then I write a Java snippet to calculate the edgelist and save it as a text file called edgelist.txt. The Java code uses a library called Jsoup to calculate the edgelist. This screenshot below shows the critical part of the code.

   For variable ***url***, we have to use **abs:href** for **link.attr**.

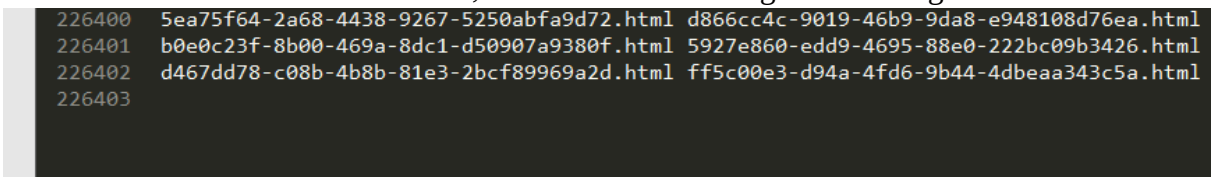
```
// file counter
int i = 1;
for (File file : dir.listFiles()) {
    Document doc = Jsoup.parse(file, "UTF-8", fileUrlMap.get(file.getName()));
    System.out.println(i++);
    Elements links = doc.select("a[href]");
    // Elements pngs = doc.select("[src]");
    // Element tag = doc.select("html").first();
    // String t = tag.text();

    for (Element link : links) {
        String url = link.attr("abs:href").trim();
        if (urlFileMap.containsKey(url)) {
            String edge = file.getName() + " " + urlFileMap.get(url);
            edges.add(edge);
        }
    }
}

Iterator<String> iter = edges.iterator();
while (iter.hasNext()) {
    System.out.println(iter.next());
    writer.write(iter.next() + "\n");
}
writer.flush();
writer.close();
```

After running this code, I got a edgelist text file. Here is a screenshot of what it looks like.



The file means that there is outgoing links from the HTML code on the **left** side point to the HTML code on the **right** side.

The number of HTML files is **18779**, and the number of edges in this edgelist.txt file is **226402**.



4. Based on the tutorial we got for the homework assignment, here is how I computed the pagerank. ***pr = nx.pagerank(G, alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None);***

   alpha is 0.85, personalization is None, max iteration time is 30, tolerance (decimal number) is 1e-06, nstart is None, weight is 'weight', and dangling is None. Code is below. Notice that I attached the crawl_data folder path in the beginning of the file name.

```
__author__ = 'Shurui Liu'
import networkx as nx

def generatePageRank():
    hashmap = readFile();
    G = nx.DiGraph();
    for key in hashmap:
        G.add_node(key);
        for url in hashmap[key]:
            G.add_edge(key, url);

    pr = nx.pagerank(G, alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None);
    output = open("external_PageRank.txt", "w+");
    for key in pr:
        output.write("/home/aaron/Downloads/solr-6.5.0/crawl_data/" + key + "=" + ("%f" % pr[key]) + "\n");
        # output.write("/" + key + "=" + ("%.6f" % pr[key]) + "\n");
    output.close();

def readFile():
    hashmap = {};
    filename = "edgelist.txt";
    with open(filename, "r") as lines:
        for line in lines:
            # format is .html, space, .html
            urls = line.split();

            if urls[0] not in hashmap:
                hashmap[urls[0]] = [];

            for url in urls[1:]:
                hashmap[urls[0]].append(url);
    return hashmap;

if __name__ == "__main__":
    generatePageRank();
```

```
 for key in pr:
     output.write("/home/aaron/Downloads/solr-6.5.0/crawl_data/" + key + "=" + ("%f" % pr[key]) + "\n");
     # output.write("/" + key + "=" + ("%.6f" % pr[key]) + "\n");
 output.close();
```

After running this code, I got a text file called **external_PageRank.txt**, which is what we need to import into Solr for the PageRank algorithm. Here's what it looks like below. The format is

/filepath/filename.html=0.xxx

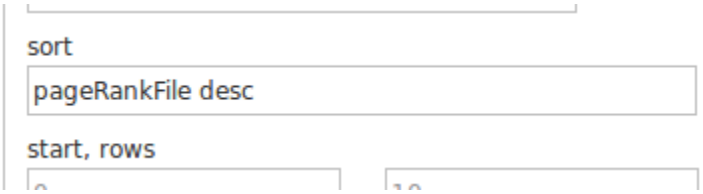
external_PageRank.txt

```
/home/aaron/Downloads/solr-6.5.0/crawl_data/495fa265-cf57-4527-a6b5-b95e44e17058.html=0.000009
/home/aaron/Downloads/solr-6.5.0/crawl_data/f3823a79-cabc-4c7e-a976-f4511f3778a4.html=0.000015
/home/aaron/Downloads/solr-6.5.0/crawl_data/342ec473-4e75-43fa-b9a3-fd52c1705aca.html=0.000015
/home/aaron/Downloads/solr-6.5.0/crawl_data/743f4d20-f4f2-4c70-9396-786ed57cbfbc.html=0.000041
/home/aaron/Downloads/solr-6.5.0/crawl_data/cceedeca-29fd-450b-a561-0f28d8af0c20.html=0.000009
/home/aaron/Downloads/solr-6.5.0/crawl_data/ce5671de-edb3-41c2-993c-b0eba8fcea5e.html=0.000013
/home/aaron/Downloads/solr-6.5.0/crawl_data/434eaaf7-ed70-4312-9ac7-29cee7a756e0.html=0.000015
/home/aaron/Downloads/solr-6.5.0/crawl_data/323b51c7-a1df-4a1d-a3b0-cddc31a79928.html=0.000223
/home/aaron/Downloads/solr-6.5.0/crawl_data/819f58d7-ad56-45e6-ac32-23635529b561.html=0.000013
/home/aaron/Downloads/solr-6.5.0/crawl_data/049b75cf-cd44-4d13-a0f8-0873f3452737.html=0.000014
/home/aaron/Downloads/solr-6.5.0/crawl_data/e98d1b6c-ac9e-4846-89fb-47d146a26877.html=0.000009
/home/aaron/Downloads/solr-6.5.0/crawl_data/45fe244c-d1d3-4567-9d82-0c39ab5bca3b.html=0.000018
/home/aaron/Downloads/solr-6.5.0/crawl_data/4b3a7c64-ebe2-4a86-9afb-293a3b29b3ed.html=0.000009
/home/aaron/Downloads/solr-6.5.0/crawl_data/e258994d-2761-474b-9986-45db67279397.html=0.000010
```

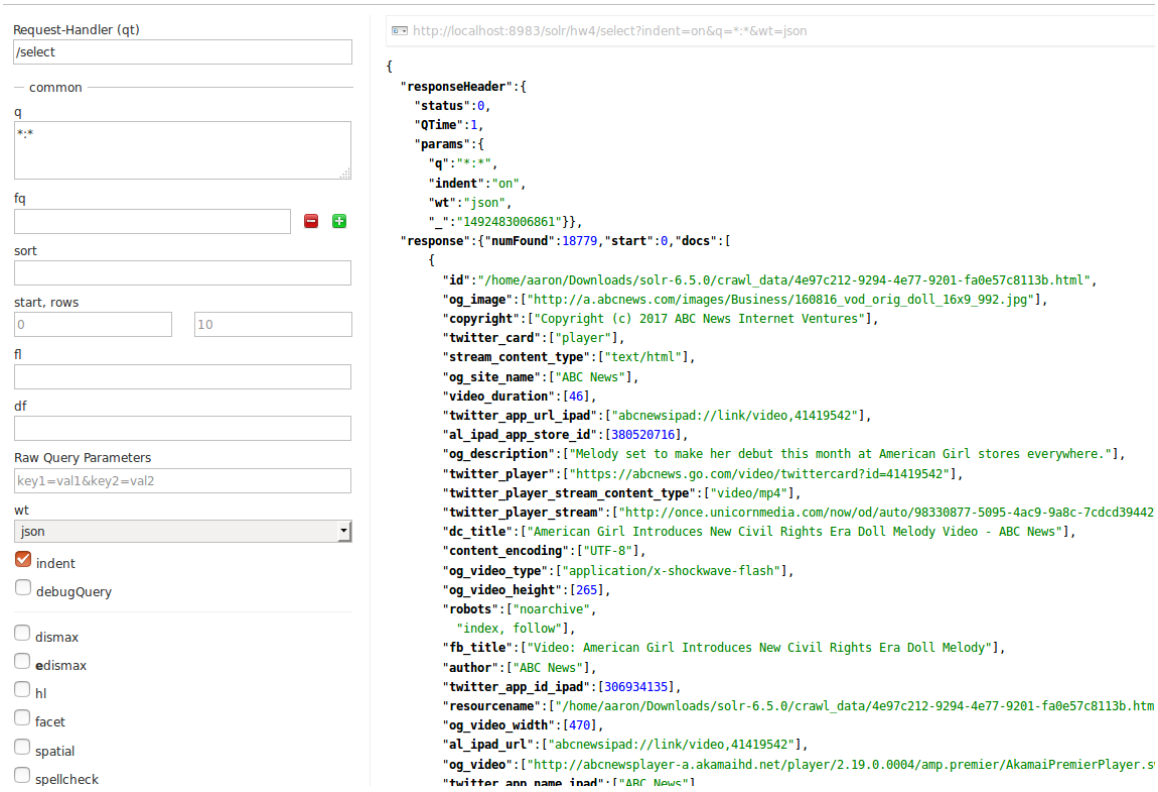
5. After having this **external_PageRank.txt** file, I place it at here, **solr-6.5.0/server/solr/hw4/data** Notice that **hw4** is my core name.

```
aaron@ubuntu:~/Downloads/solr-6.5.0/server/solr/hw4/data$ ls -l
total 1756
-rw-rw-r-- 1 aaron aaron 1761680 Apr 17 11:18 external_PageRank.txt
drwxrwxr-x 2 aaron aaron   20480 Apr 17 18:37 index
drwxrwxr-x 2 aaron aaron    4096 Apr 17 13:51 snapshot_metadata
drwxrwxr-x 2 aaron aaron    4096 Apr 17 18:28 tlog
aaron@ubuntu:~/Downloads/solr-6.5.0/server/solr/hw4/data$
```

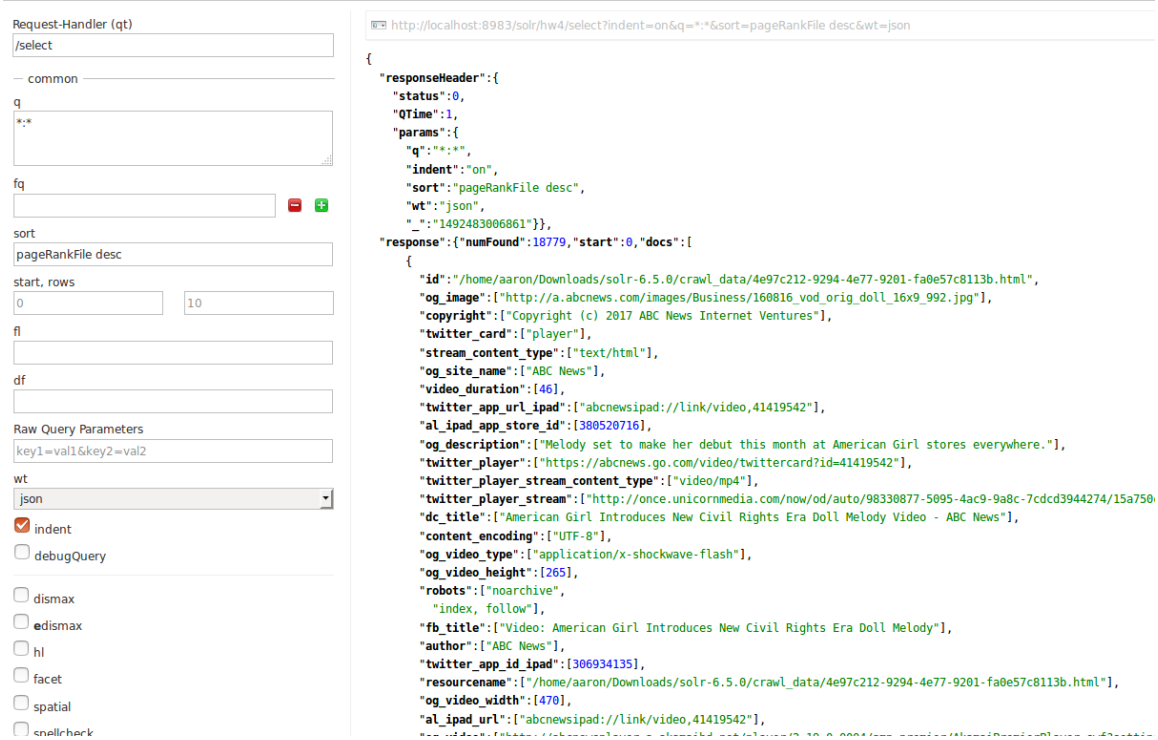
6. Then going back to the Solr UI, make a query with this change to apply the PageRank algorithm.



But unfortunately there is **NOTHING** changed when I applied this sorting algorithm. The query results are the **SAME**. I don't think the problem comes from the edgelist or the pagerank calculation, it might come from the xml configuration. Screenshots are below.




```
{
  "responseHeader":{
    "status":0,
    "QTime":1,
    "params":{
      "q":"*:*",
      "indent":"on",
      "wt":"json",
      "_":"1492483006861"}},
  "response":{"numFound":18779,"start":0,"docs":[
      {
        "id":"/home/aaron/Downloads/solr-6.5.0/crawl_data/4e97c212-9294-4e77-9201-fa0e57c8113b.html",
        "og_image":["http://a.abcnews.com/images/Business/160816_vod_orig_doll_16x9_992.jpg"],
        "copyright":["Copyright (c) 2017 ABC News Internet Ventures"],
        "twitter_card":["player"],
        "stream_content_type":["text/html"],
        "og_site_name":["ABC News"],
        "video_duration":[46],
        "twitter_app_url_ipad":["abcnewsipad://link/video,41419542"],
        "al_ipad_app_store_id":[380520716],
        "og_description":["Melody set to make her debut this month at American Girl stores everywhere."],
        "twitter_player":["https://abcnews.go.com/video/twittercard?id=41419542"],
        "twitter_player_stream_content_type":["video/mp4"],
        "twitter_player_stream":["http://once.unicornmedia.com/now/od/auto/98330877-5095-4ac9-9a8c-7cdcd39442
        "dc_title":["American Girl Introduces New Civil Rights Era Doll Melody Video - ABC News"],
        "content_encoding":["UTF-8"],
        "og_video_type":["application/x-shockwave-flash"],
        "og_video_height":[265],
        "robots":["noarchive",
          "index, follow"],
        "fb_title":["Video: American Girl Introduces New Civil Rights Era Doll Melody"],
        "author":["ABC News"],
        "twitter_app_id_ipad":[306934135],
        "resourcename":["/home/aaron/Downloads/solr-6.5.0/crawl_data/4e97c212-9294-4e77-9201-fa0e57c8113b.htm
        "og_video_width":[470],
        "al_ipad_url":["abcnewsipad://link/video,41419542"],
        "og_video":["http://abcnewsplayer-a.akamaihd.net/player/2.19.0.0004/amp.premier/AkamaiPremierPlayer.s
        "twitter_app_name_ipad":["ABC News"]
```


```
{
  "responseHeader":{
    "status":0,
    "QTime":1,
    "params":{
      "q":"*:*",
      "indent":"on",
      "sort":"pageRankFile desc",
      "wt":"json",
      "_":"1492483006861"}},
  "response":{"numFound":18779,"start":0,"docs":[
      {
        "id":"/home/aaron/Downloads/solr-6.5.0/crawl_data/4e97c212-9294-4e77-9201-fa0e57c8113b.html",
        "og_image":["http://a.abcnews.com/images/Business/160816_vod_orig_doll_16x9_992.jpg"],
        "copyright":["Copyright (c) 2017 ABC News Internet Ventures"],
        "twitter_card":["player"],
        "stream_content_type":["text/html"],
        "og_site_name":["ABC News"],
        "video_duration":[46],
        "twitter_app_url_ipad":["abcnewsipad://link/video,41419542"],
        "al_ipad_app_store_id":[380520716],
        "og_description":["Melody set to make her debut this month at American Girl stores everywhere."],
        "twitter_player":["https://abcnews.go.com/video/twittercard?id=41419542"],
        "twitter_player_stream_content_type":["video/mp4"],
        "twitter_player_stream":["http://once.unicornmedia.com/now/od/auto/98330877-5095-4ac9-9a8c-7cdcd3944274/15a750
        "dc_title":["American Girl Introduces New Civil Rights Era Doll Melody Video - ABC News"],
        "content_encoding":["UTF-8"],
        "og_video_type":["application/x-shockwave-flash"],
        "og_video_height":[265],
        "robots":["noarchive",
          "index, follow"],
        "fb_title":["Video: American Girl Introduces New Civil Rights Era Doll Melody"],
        "author":["ABC News"],
        "twitter_app_id_ipad":[306934135],
        "resourcename":["/home/aaron/Downloads/solr-6.5.0/crawl_data/4e97c212-9294-4e77-9201-fa0e57c8113b.html"],
        "og_video_width":[470],
        "al_ipad_url":["abcnewsipad://link/video,41419542"],
```

II. Query results

- Brexit
  Results 1 - 10 of 132:
  Document Brexit European Union Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 46271
  Document Brexit European Union Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 51693
  Document Tony Blair's new mission is to change minds on Brexit

  Author: ABC News | Size: 25167
  Document Brexit European Union Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 43041
  Document Brexit European Union Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 42704
  Document Brexit European Union Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 42701
  Document Brexit European Union Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 42761
  Document Brexit European Union Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 42761
  Document AP Interview: Sweden PM: Brexit deal in 2 years 'very tough'

  Author: ABC News | Size: 28917
  Document Tony Blair's new mission is to change minds on Brexit - ABC News

  Author: ABC News | Size: 76964

- NASDAQ
  Results 1 - 10 of 67:
  Document NASDAQ Composite Index Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 43866
  Document NASDAQ 100 Index Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50615
  Document NASDAQ Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 44503
  Document Indexes News, Photos and Videos - ABC News

  Author: None | Size: 29670
  Document How major US stock market indexes fared on Friday

  Author: ABC News | Size: 23618
  Document Markets Right Now: Late push renews records for US indexes

  Author: ABC News | Size: 23199
  Document Stocks Party Like It&#xFFFD;s 2000 During Record Highs - ABC News

  Author: ABC News | Size: 76232
  Document Russell 2000 Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 43605
  Document Triple-Digit Decline for Dow Jones Industrials - ABC News

  Author: ABC News | Size: 98595
  Document Facebook Shares Back at the IPO Price - ABC News

  Author: ABC News | Size: 77900

- NBA
  Results 1 - 10 of 526:
  Document National Basketball Association NBA

  Author: None | Size: 51512
  Document NBA warns Texas over proposed 'bathroom bill'

  Author: ABC News | Size: 20113
  Document Stephen Curry, Warriors still top NBA merchandise sales - ABC News

  Author: ABC News | Size: 69263
  Document #NBArank Rising Stars: Your list of the NBA's top players under 25 - ABC News

  Author: ABC News | Size: 68628
  Document #NBArank Rising Stars: Your list of the NBA's top players under 25 - ABC News

  Author: ABC News | Size: 68630
  Document #NBArank Rising Stars: Your list of the NBA's top players under 25 - ABC News

  Author: ABC News | Size: 68630
  Document National Basketball Association NBA

  Author: None | Size: 46904
  Document National Basketball Association NBA

  Author: None | Size: 46847
  Document National Basketball Association NBA

  Author: None | Size: 46104
  Document National Basketball Association NBA

  Author: None | Size: 46160

- Snapchat
  Results 1 - 10 of 95:
  Document Snapchat Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 48462
  Document Snapchat Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50407
  Document Snapchat Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50704
  Document Snapchat Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50406
  Document Snapchat Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50409
  Document Snapchat Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50405
  Document Snapchat Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50849
  Document Snapchat Filters Among Top Halloween Costume Trends of 2016 - ABC News

  Author: ABC News | Size: 79753
  Document Police: 200 Descend on Philadelphia Mall in 2nd Night of Violence; Fracas Organized on Snapchat - ABC News

  Author: ABC News | Size: 78049
  Document Snapchat Files for IPO Video - ABC News

  Author: ABC News | Size: 65517

- Illegal Immigration
  Results 1 - 10 of 1694:
  Document Illegal Immigration Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 45964
  Document Illegal Aliens Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 46763
  Document Arizona Immigration Law: Judge Susan Bolton Blocks Parts of SB 1070 - ABC News

  Author: ABC News | Size: 77985
  Document Nationality Law Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 48561
  Document Dying Dad Jesus Navarro Denied Kidney Transplant Over Immigration Status, but Supporters Try to Help - ABC News

  Author: ABC News | Size: 79729
  Document Dying Dad, Jesus Navarro, to Get Kidney Transplant Despite Undocumented Immigration Status - ABC News

  Author: ABC News | Size: 76962
  Document Immigration News, Photos and Videos - ABC News

  Author: None | Size: 30757
  Document Immigration News, Photos and Videos - ABC News

  Author: None | Size: 30761
  Document Judge Susan Bolton Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 46864
  Document Businesses nationwide participate in Day Without Immigrants protest

  Author: ABC News | Size: 27907

- Donald Trump
  Results 1 - 10 of 10141:
  Document Donald Trump Jr. Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 48293
  Document Donald Trump's Inauguration in Photos Photos - ABC News

  Author: ABC News | Size: 122105
  Document Tiffany Trump: News, Videos and Photos-ABC News

  Author: None | Size: 51537
  Document The Moment Donald Trump and His Family Knew He Won the Election - ABC News

  Author: ABC News | Size: 78493
  Document Trump's First 100 Days Photos - ABC News

  Author: ABC News | Size: 125591
  Document Donald Trump Latest News and Videos

  Author: None | Size: 50757
  Document ANALYSIS: Donald Trump Sees Calm After Storm in First DC Trip as President-Elect - ABC News

  Author: ABC News | Size: 72805
  Document GOP Strategist Likely Faces Hurdles in Defamation Suit Against Donald Trump - ABC News

  Author: ABC News | Size: 74133
  Document President Donald Trump Vows to Launch 'Major Investigation' Into Alleged Voter Fraud - ABC News

  Author: ABC News | Size: 77476
  Document OPINION: What Type of President Will Donald Trump Be - Lincoln or Buchanan? - ABC News

  Author: ABC News | Size: 72281

- Russia
  Results 1 - 10 of 1553:
  Document Top Trump envoys signal no quick changes to US-Russia ties

  Author: ABC News | Size: 28396
  Document US secretary of state says Russia must honor Ukraine deal

  Author: ABC News | Size: 25419
  Document Russia Sanctions Explained: A Look at the Measures Donald Trump Could Roll Back - ABC News

  Author: ABC News | Size: 81310
  Document Russia Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 47345
  Document US official says Russia deployed missile in treaty violation - ABC News

  Author: ABC News | Size: 73777
  Document What we know - and don't know - about the Trump team's contacts with Russia before the election - ABC News

  Author: ABC News | Size: 78070
  Document Nikki Haley Blasts Russia in First Remarks at UN Security Council - ABC News

  Author: ABC News | Size: 78046
  Document Trump: Reports that campaign had contact with Russia are 'fake news'

  Author: ABC News | Size: 27173
  Document EU Diplomats Back Continued Russia Sanctions - ABC News

  Author: ABC News | Size: 75468
  Document Danes See Russia as a 'Leader' With 'Advanced Capabilities' - ABC News

  Author: ABC News | Size: 65616

- NASA
  Results 1 - 10 of 4937:
  Document NASA

  Author: None | Size: 47764
  Document Incredible Images Captured From Space Photos - ABC News

  Author: ABC News | Size: 227987
  Document Picture | Incredible Images Captured From Space - ABC News

  Author: ABC News | Size: 228899
  Document NASA Targets New Date for Mars InSight Drilling Mission - ABC News

  Author: ABC News | Size: 65094
  Document Two Earth-Size Planets Found by NASA Kepler Mission - ABC News

  Author: ABC News | Size: 68683
  Document Pluto's Moon Charon Had a 'Hulk' Period, New NASA Images Reveal - ABC News

  Author: ABC News | Size: 66624
  Document NASA Wants to Blast Your Art Work Into Space - ABC News

  Author: ABC News | Size: 64284
  Document Kepler Space Telescope Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 51246
  Document Mars Exploration Videos at ABC News Video Archive at abcnews.com

  Author: None | Size: 50943
  Document Breathtaking Views of Earth From Above Photos - ABC News

  Author: ABC News | Size: 100529

III. Graphs

Since I didn't get the PageRank algorithm work on Solr, there is no graphs to show.

IV. Explanation about why some pages have higher page rank

The logic of PageRank algorithm is ranking different webpages based on the number of outgoing links (outgoing edges) they have. For any pages except those isolated pages (with no outgoing links), if you have more numbers of outgoing links, you will have a relatively higher rank; but if you have more incoming links, it will lower your ranking.