

Report of homework 4

Name: Chen Xu

Student ID: 9012039827

Email:cxu264@usc.edu

INDEX

Ch1:steps for implementation	3
Step1:environment setup:	3
Step2:Index the whole website of homework:	3
Step3:write the code for search index page	3
Step4: compute the edge List and page rank.....	6
Ch2. why some pages ranks higher:.....	8
Ch3.flow of searching Star Wars	10
Ch4.table of query.....	11
Query1:Elon Musk.....	11
Query2: Star Wars	13
Query3: North Korea.....	14
Query4: LA Dodgers	15
Query5: Puerto Rico	16
Query6: Hurricane Harvey	17
Query7: iPhone X	18
Query 8: Paris Climate Deal.....	20
Ch5.Overlap graph:	21

Ch1:steps for implementation

Step1:environment setup:

for the setup of the environment I have installed the Ubuntu 16.04 LTS in my virtualbox and java 1.8. So I just download solr-7.1.0 and unzip it in my Desktop of Ubuntu. But here the most annoying point of Ubuntu is it always have the access permission for user to access the root files. So that is the reason next of why I choose to install apache server in my windows. Because when you use `sudo apt-get install apache2`, it will automatically install the root website of your apache into `share/apache2` folder, this folder has the root permission, so it is really hard and inconvenient to move files and update webpages to that. So I choose wamp server 3.0.6 for my apache server, this is the combination of apache+mysql+php and it is easy to do things. Here above is the basic things I did for environment

Step2:Index the whole website of homework:

My mission in this homework is to analyze the USA_TODAY. So I download the files from google drive and then I create the core and name it:csci572.After that ,I use solr to index all of the data,here below is the screen shot of my core:

The screenshot displays the Solr Admin interface for a core named 'csci572'. It is divided into four main sections: Statistics, Instance, Replication (Master), and Healthcheck.

- Statistics:** Shows core metrics including Num Docs (18517), Max Doc (18517), Heap Memory (-1), Deleted Docs (0), Version (504), Segment Count (22), and status indicators for Optimized and Current (both green checkmarks).
- Instance:** Lists configuration paths for CWD, Instance, Data, Index, and Impl.
- Replication (Master):** A table showing the state of the master node.
- Healthcheck:** A message indicating that the ping request handler is not configured with a healthcheck file.

	Version	Gen	Size
Master (Searching)	1510431250183	80	500.63 MB
Master (Replicable)	-	-	-

At the bottom of the interface, there are links to Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

It took a while to finish it, I went to a lunch during that time, after I came back, it finished and there it took 500.63 MB

Step3:write the code for search index page

For wamp server, the website is stored in the folder of “www” like this below:

名称	修改日期	类型	大小
alias	2017/7/17 20:13	文件夹	
apps	2017/7/17 20:14	文件夹	
bin	2017/7/17 20:14	文件夹	
cgi-bin	2017/7/17 20:13	文件夹	
lang	2017/7/17 20:14	文件夹	
logs	2017/7/17 20:21	文件夹	
scripts	2017/7/17 20:14	文件夹	
tmp	2017/7/18 21:03	文件夹	
www	2017/11/11 15:02	文件夹	
barimage	2010/12/30 17:39	图片文件(.bmp)	5 KB
images_off	2016/8/1 1:21	图片文件(.bmp)	27 KB
images_on	2016/8/1 1:22	图片文件(.bmp)	27 KB
install-english	2016/10/16 0:09	文本文档	4 KB
license-english	2015/11/5 19:00	文本文档	8 KB
read_after_install-english	2016/3/8 3:45	文本文档	2 KB
unins000	2017/7/17 20:15	媒体文件(.dat)	950 KB
unins000	2017/7/17 20:13	应用程序	1,369 KB
uninstall_services	2017/7/17 20:13	Windows 批处理...	1 KB
wampmanager.conf	2017/7/17 20:15	CONF 文件	2 KB
wampmanager	2008/9/3 0:46	应用程序	1,205 KB
wampmanager	2017/11/12 12:46	配置设置	497 KB
wampmanager	2016/8/21 19:42	TPL 文件	24 KB

so below the www I create a folder of homework to store the files of homework4,here is the screen shot of homework folder:

Apache	2016/11/18 18:33	文件夹	
index	2017/11/12 0:53	PHP Script	4 KB

The Apache folder is the one I download from git like the homework instructions and Index is just the search page, in the next paragraph I am going to talk about how to write the code for index.php and before that I need to mention when I use the index.php what input in my browser(the address is):

<http://localhost:8080/homework/>

because my 80 port is occupied and reserved for my tomcat server(used for developing JSP) so I change the port for my apache server.

The code is pretty straight forward:the first part of the code:

```
<?php
header('Content-Type:text/html;charset=utf-8');
error_reporting(E_ALL ^ E_WARNING);
$limit=10;
$query=isset($_REQUEST['q'])?$_REQUEST['q']:false;
$results=false;

if($query){

    require_once('Apache/Solr/Service.php');
    $solr=new Apache_Solr_Service('192.168.0.31',8983,'/solr/csci572/');
    if(get_magic_quotes_gpc()==1){
        $query=stripslashes($query);
    }
    try{
        if(isset($_GET['way'])==false){
            $choice="Lucence";
        }
        else{
            $choice=$_GET['way'];
        }
        if(strcmp($choice,"Lucence")==0){
            $results=$solr->search($query,0,$limit);
        }
        if(strcmp($choice,"PageRank")==0){
            $param=array('sort'=>'pageRankFile desc');
            $results=$solr->search($query,0,$limit,$param);
        }
    }
    catch(Exception $e){
        die("<html><head><title>SEARCH ERROR</title></head><body><pre>{$e->__toString()}</pre></body></html>");
    }
}
?>
```

What I just want to mention here is the `$_GET['way']`: this is just used for selecting the algorithm for search, because besides the search button and input I add radio button to that. here I also made a mistake during I implement it because I miss type the variable “results” into “result”....So in the next section of fetching the result, I got noting initially because the json data is stored in “result” not “results”, it took me nearly 2 hours to find it out, what a stupid problem.....(I even used wireshark to track the data and found I have received the data from solr.....)

Then the next part is the form of search:

```
<?php
<html>
<head>
<title>csci572 homework4</title>
</head>
<body bgcolor="#FAEBD7">
<div align="center">
<form accept-charset="utf-8" method="get">
<label for="q">Search:</label>
<input id="q" name="q" type="text" value="<?php echo htmlspecialchars($query,ENT_QUOTES,'utf-8');?>" />
<input type="submit" /><br/>
<input type="radio" name="way" value="PageRank" />PageRank<br/>
<input type="radio" name="way" value="Lucence" />Lucene<br/>
</form>
</div>
<?php // display results
```

Here you can see the radio button I use and the name of input is ‘q’.

Finally the last part is :

```
<?php // iterate document fields / values
foreach ($doc as $field => $value) {

    $flag1=strcmp($field,'id');
    $flag2=strcmp($field,'description');
    $flag3=strcmp($field,'title');
    $flag4=strcmp($field,'og_url');
    if($flag1!=0 && $flag2!=0 && $flag3!=0 && $flag4!=0){
        continue;
    }
    else if($flag1==0){
        $ID=$value;
    }
    else if($flag2==0){
        $DES=$value;
    }
    else if($flag3==0){
        $title=$value;
    }
    else if($flag4==0){
        $url=$value;
    }
}

echo '<tr>';
echo '<th>Title:</th>';
echo '<td width="100%"><a href=".' . $url . '>' . htmlspecialchars($title,ENT_QUOTES,'utf-8') . '</a></td>';
echo '</tr>';

echo '<tr>';
echo '<th>URL:</th>';
echo '<td width="100%"><a href=".' . $url . '>' . htmlspecialchars($url,ENT_QUOTES,'utf-8') . '</a></td>';
echo '</tr>';

echo '<tr>';
echo '<th>ID:</th>';
echo '<td width="100%">' . htmlspecialchars($ID,ENT_QUOTES,'utf-8') . '</td>';
echo '</tr>';

echo '<tr>';
echo '<th>Description:</th>';
echo '<td width="100%">' . htmlspecialchars($DES,ENT_QUOTES,'utf-8') . '</td>';
echo '</tr>';
?>
```

Because the order is id,description,title and og_url,so I use several variables to store them and output them in the table of html.

Also for the param: `$param=array('sort'=>'pageRankFile desc');`

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

The two line of codes above are the ones I used for selecting page rank algorithm.

Step4: compute the edge List and page rank

(Jsoup+networkx)

Then I write the java code to construct the edgelist of dataset. The website pages is just a directed graph, so below the code is used for get edge List:

```
import java.io.BufferedReader;

public class ExtractLinks {
    Map<String,String> file_url=new HashMap<String,String>();
    Map<String,String> url_file=new HashMap<String,String>();
    Set<String> edges=new HashSet<String>();
    public void presolve() throws IOException{

        FileReader reader=new FileReader("USA Today Map.csv");
        BufferedReader br=new BufferedReader(reader);
        String str=null;
        while((str=br.readLine())!=null){
            String[] group=str.split(",");
            file_url.put(group[0],group[1]);
            url_file.put(group[1],group[0]);
        }
        System.out.println(file_url.size());
        br.close();
        reader.close();
    }
    public void extract() throws IOException{
        String dirpath="C:\\Users\\xuchen\\Desktop\\CRAWL\\USA Today";
        File dir=new File(dirpath);
        for(File file:dir.listFiles()){
            // System.out.println(file.getName());
            Document doc=Jsoup.parse(file,"UTF-8",file_url.get(file.getName()));
            Elements links=doc.select("a[href]");
            for(Element link:links){
                String url=link.attr("abs:href").trim();
                if(url_file.containsKey(url)){
                    edges.add(file.getName()+" "+url_file.get(url));
                    System.out.println(file.getName()+" "+url_file.get(url));
                }
            }
        }
        BufferedWriter bw=null;
        FileWriter fw=null;
        fw=new FileWriter("edgeList.txt");
        bw=new BufferedWriter(fw);

        for(String cur:edges){
            bw.write(cur);
            bw.write("\n");
        }
        bw.close();
        fw.close();
    }
}
```

File_url and url_file hashmap used for mapping between name of page and real url address. So presolve() is used for read and fill in two hashmap using USA_TODAY And extract() is used for construct the edge list here below is the methond to illustrate the things I do for getting edge list:

Foreach file in dir

Doc = parse(file);

Links=doc.fetch(<a>)

Foreach link in links:

If(link in url=>file map) then add the edge (self,link) into the result set

Else continue;

After this I get the edge list file of about 226.4MB:

d453c44841a4f2d56248f95a0934b251ecc829a04d2c83a503723ab23a6bb537. html 2072b79c3aa28dfdf06cb9d16fbf218c6a3ee7e1459cdab8fa1b9e21cf89330a. html
 7e08fedcc5d1c95112749f2eece33f8def9e59a21bfb9c23eb6df46931a3a391. html e7e18f8de632cf4027cd88a70e540e1437fc608ee02ef5834681e6354df1e34a. html
 9a0de7b5d2bf927489756554c23b0b9660e5ad1d598bb5a828c496bac10ee78. html 8b87c5af6f5c25ed01c3d667d56eb5f5c287d3f1d6d2c14120123803f89f7bc8de. html
 d89acbc05d8f4c4096929d0c684e200a9566cf3f660d26aaab8fad021894671. html ba7c682c01a6e39b130002fdd33517193d0a428bca415512f0ca3df817d8f4. html
 79c38d452f120906268e839c2e76adad1bb743bd0190949936ae4938ea. html e6e0045e3f8db214ede493d8c61f9f208f2283d70d6c6a3af9e0673764c523b. html
 2cf6c53d51c08597db203554c44599f67524636ab06e3d0907fc05286a4fe970. html 570285c2a82e2d4b9592c29873e91d4d2d11c30db99b26a5d3d3a06b53c84dd2. html
 e489b6371208ca5f29426fa88a2aad0c194c52b8f39d1babe420ecf19878208c. html 75ba2e9b9c567d8d0a6ced0e339e1ce4ed338520296bb9536665dae1fadcc00. html
 b8fd38d5a0c31bf288b3e9642900666d46ab364d51a579030f44703e1304f6. html 55720a7b133d1029a9e6dbb5c876c64cc88d1735e4c1b54af58d0e538bcd0d0. html
 b89f42f5c0f115a5bdf717fe7e55146c1819b26248cd4539018c1587edf1da. html b31002346645134fb828c1c85f7bfb9e9a7fa3226e79215f8be9ae52f0608658. html
 c52c501508e275f09cade139bd216d45823ace28fb279364d7637e46d0b29af. html ae004cd1d6d1bf535c07181df9651760f6e486bbbed4adbe48b5f0f65a82c2ec91. html
 a9a20672cb45ec064f0b3a471a0e2a3c0179fac9add36d0af94c3c91e7d941d8. html ba7c682c01a6e39b130002fdd33517193d0a428bca415512f0ca3df817d8f4. html
 ef6743b15b232dfc34e924a23152f42d8b2fe7c71fe532476758c2de0e85e4f3. html c7e023f180966c1dcf82df0dd77330849b5906156607b03f409a2ccb7870b4e. html
 2fcdaf70279324fdb8e66c5509e5eb479c275f23f4d7bf6bc9778eff86defe93b. html ce2a73398c64cfdaa9b566172b2dd6be30055053bf834a0eb7870e79fe79ef54. html
 42e4a7832e3e1619db2c05055eb479c275f23f4d7bf6bc9778eff86defe93b. html 570285c2a82e2d4b9592c29873e91d4d2d11c30db99b26a5d3d3a06b53c84dd2. html
 533a8040e8eae63b540c9ba7f5891b8b32a932b6ad26a83ab6b67e1a37f470d4. html 7583e55257f38b4db725307be266d6ea030c3ba0e003e6b31a5539db9c9342f. html
 9d7ede595a96f0b4dcf3174e2a2a527aa4b3786f84ed913afb9a0d0aa5cc6d5c. html 23016bd5f3d6df3f9df2c3217ac554531ccc9f7011a81d9d03f2473a66ee79e. html
 61401225a18db7b105e77e29b196ce7431b989ed250d265f1c1494db681dd66. html ac0040f1fb27f1f62ea8cd4a4a039f7f86d00c278d6ed8facda2504753a2e825bc. html
 1a21bb39423939c9cf3f36d96f5cfb2e6f2cc7bc557a0b37573c6baebccb49851. html d236c993a97338813c8dbdf90efad4c990b610b36dd3d93c2ad247e258cc7206. html
 536c6aaccd6c6b367b0c8faa0fbb5b28166f317bfa71286d267bfcd592b2ab. html 08f9ba87c9a24e6df2b2b62c3f71c6c3076c841d0c29ab23254601f3bff36a2. html
 348855a6f5f6d543521f4f79800e219d6cd6823711e73a262713b366980e998d. html ce5d082601a4917d11c20aba701b6c9c3ed9314ecd4b7e989a5bd33e03b296d6. html
 c6e060787ca70004b6ee6bec0d498ade0a73ac488bc3a085cbb41b4588f880. html 6166eeaa9acebb5088a1146af7fff687f96163bf845d1c75b552329dcf66102. html
 53cb9e4f7e9e206c36ec6b047430169724bf63ae93787264b796df461a3a9ad. html ba7c682c01a6e39b130002fdd33517193d0a428bca415512f0ca3df817d8f4. html
 7e32dc8591b2189a30a1b2a9b011865f8b256ced587590a854551c0603a6918. html 651c9f55914416a45ce036e7574b2b191efa7d2b5532042ae70094f3426d0ff4. html
 f17fd11499a4bb1a742f708422ca5700e7fab84aba61ca18399c9feaddb3a2. html ea79169eda0e3e48f75833996ce05fb6b6ae7efef61a35d710c5927e54686fc1. html
 e5428ee4d85084464fd095cd8c2e7d7d7dad6e91d0f6367ab96bde4be1. html c884751aa7e2cf7d9812deb457b1215d9a5a3782351ca8aa8cb3e2cd2a814d4. html
 e677793e6a397de0b98642ba0c9f90ddbc73f1d08d0e01f1d6a7bf319959ec25. html a42e79929f10f1f14197f7a88484fbec33378082eab266c2c9f20ed78765f9. html
 0453e88a702e20c498e21e34b64d49b1820517653845ff8450a19599a32718c. html e52425b61ad5ba34d39bbe318cbe1b3cc1388db64c5c485b647000bc2b02fafd. html
 ad72e15ee1501ab1d5aad46117bf0c20866f10dc8a246cb0804d3cde856e62d. html 5bf458ba43bb85af3be54e5abccc5bb21dece66257d799b125862f086b9312e6. html
 7aeb40109cb6f057646317ebd62c193e86c249422f7e367bf62346b4c93d. html 81c26ad6b067f0337c1ac74d54e78ccf083ba5e6e7af7ff2424d40cf178724. html
 4b096123bce380d4f616b21d3ccf3368f3865b6755a4fc889127c165ef23e627. html 77b23ce870b3a169d8ca3ff3762c12c595f9005ac31226eac5b91d7349d6d87. html
 9e3aa086938c77af420214c5ccc86096d821ca1fcd06d44a7a2db5369c24041. html 7c0fab94b252be957ec083703df97c89e3957fbfe5089f47f657e2f1dd4b0c9. html
 49046da4bdf4de08c35b9e80f9cfc4434b9637422d4a63cc9af52bbf3fcccbe. html 2072b79c3aa28dfdf06cb9d16fbf218c6a3ee7e1459cdab8fa1b9e21cf89330a. html
 2a05c79360783cf158a8b181e50c01aa987ee7fe15b9816fddedf1ea52998151. html a10ff92a9df16cf9d06d6870305b411ff3212f40398fc9895e08aca33099382f9. html
 812495ce5f7eaaad11366b74bdf3f3f65f6006f965949533cf56872c70fa7b0fd. html 98e10e07ca1a675eeffa431339b0d4cd22f50515a0619244792c5f2563f852dbe. html
 5610bc389d7c93cd8a7e872ef94aaa96d7b880b30f6d19fb6e2306f821e2b. html 472eed1071fa6be870ba86bb084d18bb540ce31316a6b49e6bb0e592144d9d1. html
 999cd898950671e126659185b4abce5a9fcd4bd83cdd5b647d211cae969b848. html 75ba2e9b9c567d8d0a6ced0e339e1ce4ed338520296bb9536665dae1fadcc00. html
 36a9f8015662a478a92405c3e46be256572b31d4bf52f726e0bf9d6ed297f8f. html 36cc419dae00b21e242e2573a865226719e794c0320f18e8b5f5ded59a06728e7. html
 6c2c975f0b2433dd543e1b3a6474f1e26c3a5b1c1ef34336010a71f10d94dc8. html 472eed1071fa6be870ba86bb084d18bb540ce31316a6b49e6bb0e592144d9d1. html
 e33fcd00362df0326726dc74f68ea37efbb85c26fe77e6f65df3a7ab6d3. html a86707529ab7018648e42cd63bd0964228728506dccc6e12e05f481be16dd0f54. html
 0857a7b728351d4dd4f4209687c9452b884296f6560e3533f1b9c965a96781ee. html 2072b79c3aa28dfdf06cb9d16fbf218c6a3ee7e1459cdab8fa1b9e21cf89330a. html

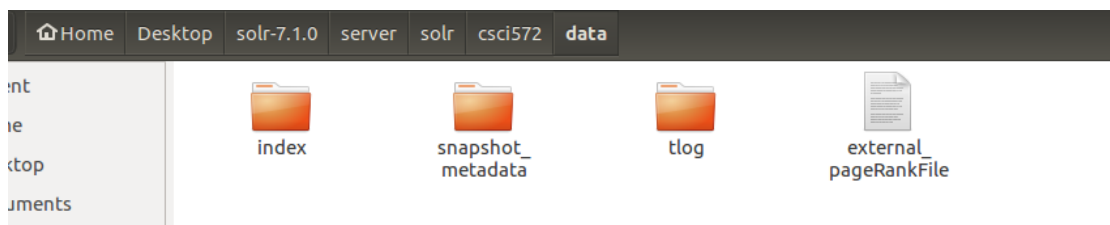
After that , I write a piece of python code to calculate the pagerank score for every page: (**python 2.7.12**),here below is the python code:

```

import networkx as nx;
print("start processing-----");
G=nx.read_edgelist("edgeList.txt",create_using=nx.DiGraph());
pr=nx.pagerank(G,alpha=0.85,personalization=None,max_iter=30,
    tol=1e-06,nstart=None,weight='weight',dangling=None);
file=open("external_pageRankFile.txt","w");
for key in pr:
    file.write("/home/xuchen/Desktop/solr-7.1.0/crawl/USA/"+key+"="+str(pr[key]));
    file.write("\n");
file.close();
print("processing end-----");
  
```

The parameter I use for pageRank are: **alpha=0.85 personalization=None max_iter=30 tol=1e-6 nstart=None weight='weight' dangling=None**

Then I get the external_pageRankFile.txt. And one thing I need to mention here: for computing edge list and pagerank: I use my host windows machine because it is faster than Ubuntu virtual machine. And then I transfer my external file in the Ubuntu and put into my core folder:



And change the information in managed-schema.xml:

```

<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalTextField"/>
<field name="pageRankFile" type="external" stored="false" indexed="false"/>
  
```

And:

```
<!-- QuerySenderListener takes an array of NamedList and executes a
     local query request for each NamedList in sequence.
-->
<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
```

Here I also made a mistake: because I stored the file name as pageRankFile.txt. and in field name I write: pageRankFile. So there are some mistakes when I use the page Rank for searching.

Here above are all the steps I did for implementation of my homework4. So next I am going to talk about why some get higher page rank and the screen shots of my homework:

Ch2. why some pages ranks higher:

For example: when I choose CNN as search content and use page rank for that:

Results 1 - 10 of 190:

1.	Title: Corrections & Clarifications URL: https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/ ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/ba7c682c01a6e39b130002fbd33517193d0a4248bca415512f0ca3df817d8f4.html Description: The following corrections and clarifications have been published by USA TODAY.
2.	Title: Videos, Photos - USA TODAY URL: https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/ ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/cc8235a5941634bfe55f56183150245da5935873b713b6c01d6b8490a2160943.html Description: View videos and photo galleries from USA TODAY
3.	Title: Dove apologizes racially insensitive Facebook advertising image URL: https://www.usatoday.com/story/money/business/2017/10/08/dove-sorry-racially-insensitive-facebook-ad/744545001/ ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/730242df4bd4eef3ba4421fb181c87e3b7cecac96400d028d2148179f971773d.html Description: Dove has apologized for an advertising image that many considered racially insensitive.
4.	Title: Disneyland: Will it stay open during Anaheim Hills fire? URL: https://www.usatoday.com/story/money/business/2017/10/09/disneyland-open-anaheim-hills-fire-spreads-smoky-clouds-park/748398001/ ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/07531b496b96970043d06db071f822dcd41ddbd1774d3de506cdfbc6f6d89ef.html Description: Orange clouds frame park's Halloween decorations.
5.	Title: North Korea: President Trump's veiled threats could lead to miscalculation. URL: https://www.usatoday.com/story/news/world/2017/10/08/trump-north-korea-tweets-may-confuse-kim-long-un-his-administration/744451001/ ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/5d01ab6c82352f20c96d39e0170deffe36385690224528077173e7174412b210.html Description: Trump's tweets may also confuse his own administration, analysts say
6.	

Then we click into the first item we get:

Sports: A Nov. 3 story on the World Series included the incorrect year for when Cleveland last won the World Series and left out a team among those who have never won the Series. The Indians won in 1948, and the Brewers are among those who have never won. <https://www.usatoday.com/story/sports/mlb/columnist/bob-nightengale/2017/11/03/world-series-over-free-agency/828166001/>

News: A story Oct. 27 about proposed tax changes for Americans misstated the effective date for mortgage-interest changes. The proposed changes would take effect after Nov. 2.

Social media: An Instagram post on @USATODAY on Nov. 2 used the wrong photo with a quote from the Houston Astros' Carlos Correa. The post was deleted.

October 2017

Your Say: A previous version of this letter to the editor mislabeled Kevin Spacey's behavior. It is now labeled as "predatory." <https://www.usatoday.com/story/opinion/2017/10/31/readers-sound-off-outrage-over-kevin-spacey-misplaced/817397001/>

Sports: A previous version of this story included an incorrect position for the Pirates' Bill Mazeroski in the 1960 World Series. <https://www.usatoday.com/story/sports/mlb/2017/11/01/world-series-game-7-classics/820041001/>

Sports: An item in Sportsline Oct. 26 incorrectly referred to Jerry Jones of the Dallas Cowboys. He is the owner/president/general manager.

Sports: A previous version of this story included an incorrect nickname for Don Meredith. <https://www.usatoday.com/story/sports/2017/10/26/middle-finger-often-middle-sports-controversy/804550001/>

Sports: A previous version of this story incorrectly referred to Martinsville as the oldest track on the NASCAR circuit. <https://www.usatoday.com/story/sports/nascar/2017/10/26/martinsville-playoff-race-first-data-500-preview/803115001/>

News: An earlier version of the story mischaracterized the Edison Electric Institute. The story also misstated the payment status of workers from Jacksonville Electric Authority for work performed. <https://www.usatoday.com/story/news/world/2017/10/30/puerto-rico-performed/803115001/>

This is actually the index page of usa today and then we try the 5:

The screenshot shows the USA Today website. On the left is a social media sidebar with icons for Facebook, Twitter, LinkedIn (145), and a comment icon (25). The main content area features a large video player showing President Trump speaking, with an 'Autoplay: On | Off' toggle. Below the video is a news article titled 'If President Trump hoped to keep North Korea guessing with his latest threats, he has also left many Americans — and perhaps even his own administration — equally puzzled.' The article text mentions that Trump said his predecessors in the White House have talked to North Korea for the past 25 years without success, and that Trump tweeted 'Sorry, but only one thing will work!' on Saturday. A quote from Jeffrey Lewis, an analyst at the Middlebury Institute of International Studies, is also included: 'What the heck is that?' said Jeffrey Lewis, an analyst at the Middlebury Institute of International Studies. 'You just can't parse any of it.' Below the article is a small image of a skyscraper. On the right side of the page is a SmartFares advertisement. It shows two flight options: 'Los Angeles to Beijing' starting from \$222 and 'Los Angeles to Kuala Lumpur' starting from \$207. Both options have a 'Click' button.

We get one of the exactly news.

Compare about these two result, we can easily know that first page contains a lot links to the other pages. Such the page in a website is the index page which is the most frequent page people may click it. So it actually rank higher than the exactly

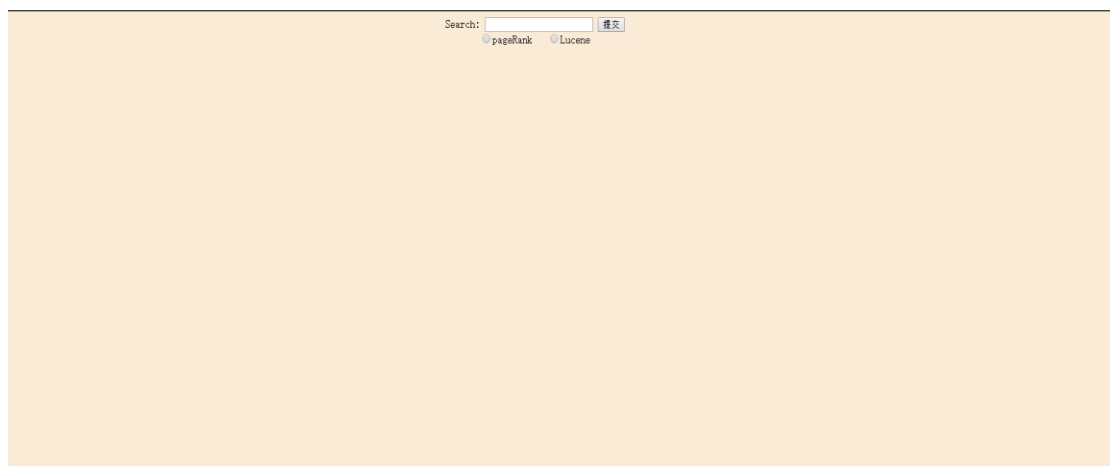
page which introduce the exactly news. In other words: people tend to back to this page after they search for some exactly news and this index page is an important hub for exactly news. Also, many news page also point to the index page.

Another condition is if there is some breaking news, some of the news page may also have the link for that news, which makes some pages of breaking news higher.

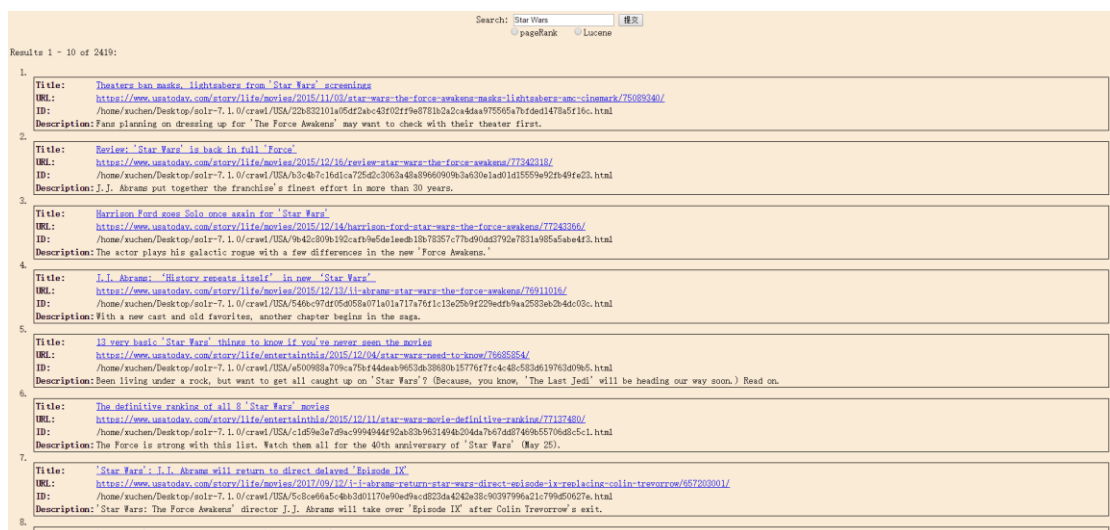
So what I understand about all of these is: the page rank compute the rank of each page just base on the web graph which seems has nothing to do with human preference. However, as a human being or web designer, we have our own preference of some pages, when you just write your own page, you may add the pages you like into your own pages and also in order to improve your own page rank, you may add some links that most people like best to improve your website visit rate. So page rank actually is the reflection of human preference and psychology.

Ch3.flow of searching Star Wars

1. before the searching, screen shot of it:



2. choose lucene as search method:



3. choose pagerank as search method:

Search:

☒ pageRank ☐ Lucene

Results 1 - 10 of 2419:

- Title:** [Virtually There from USA TODAY](#)

URL: [None](#)

ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/30c70eba96bbaecb0179d90997986374ed31d1298547952e41f74ca2aa8a.html

Description: From the USA TODAY NETWORK and YouTube, it's Virtually There, your front-row seat to amazing. Every week, we give you three cool VR experiences: one just for the thrill of it, one epic adventure and one dream destination. Take a breath. Take it in. And don't forget to look around.
- Title:** [Corrections & Clarifications](#)

URL: <https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/>

ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/ba7c682c01a6e39b130002fde-d33517193d0a4248bca415512f0ca3df917d8f4.html

Description: The following corrections and clarifications have been published by USA TODAY.
- Title:** [Videos/Photos - USA TODAY](#)

URL: <https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/>

ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/112af5e51d228598ad4948ca034462b2f3d4e482856f28b9efa204610725dc4.html

Description: View videos and photo galleries from USA TODAY
- Title:** [Financial, Economic and Money News - USATODAY.com](#)

URL: <https://www.usatoday.com/news/>

ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/027e6ff30c19504d2c32431a0a92a45d1d9e5e82d42728eca0c13fc09a63e.html

Description: The latest breaking financial news on the USA and world economy, personal finance, money markets and real estate.
- Title:** [Technology and Electronics Reviews - USATODAY.com](#)

URL: <https://www.usatoday.com/tech/>

ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/4079a8a0740bb41affe1a52940c7d04e21f77ba25f13157cca7ad3ea137da1.html

Description: Power up with breaking news on personal technology, electronics, gaming and computers.
- Title:** [Entertainment News: Celebrity gossip, photos, videos & stories - USATODAY.com](#)

URL: <https://www.usatoday.com/life/>

ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/957940f7500be1da55f55d3c277bc75adb0cc34bf7489e6f47b776dde2a5bd.html

Description: The latest news in entertainment, pop culture, celebrity gossip, movies, music, books and tv reviews.
- Title:** [USA TODAY: Latest World and US News - USATODAY.com](#)

URL: <https://www.usatoday.com>

ID: /home/xuchen/Desktop/solr-7.1.0/crawl/USA/4cb60d74e66770d7eb949b4ef059e266c1ba179500865099a07b7d02a636f53.html


4. click on one item:

USA TODAY SUBSCRIBE NOW to get home delivery

NEWS SPORTS LIFE MONEY TECH TRAVEL OPINION 47° CROSSWORDS WASHINGTON THANKSGIVING MORE

Harrison Ford goes Solo once again for 'Star Wars'

Brian Truitt, USA TODAY Published 2:05 p.m. ET Dec. 14, 2015 | Updated 3:31 p.m. ET Dec. 15, 2015



Final trailer for the movie "Star Wars: The Force Awakens." | Lucadfilm

CONNECT TWEET LINKEDIN COMMENT EMAIL MORE

NEW YORK — Fans collectively freaked out when they saw Han Solo money back onto the Millennium Falcon in the first trailer for *Star Wars: The Force Awakens* and pronounce "Chewie, we're home" to his furry best pal.

So did the cast of the new movie (in theaters Friday), says co-star Daisy Ridley (as Rey), who is with John Boyega (Finn) the new face of the franchise. "Everything felt special when we were doing it, especially on the Falcon with him."

Photo: Lucadfilm

Share your feedback to help improve our site experience!

Ch4.table of query

Query1:Elon Musk

Lucene	PageRank
1. https://www.usatoday.com/story/tech/talkingtech/2017/09/05/elon-musk-artificial-intelligence-battle-most-likely-cause-wwiii/632362001/	https://www.usatoday.com/tech/

2. https://www.usatoday.com/story/tech/2017/10/06/elon-musk-delays-self-driving-truck-focus-model-3-puerto-rico-power/741050001/	https://www.usatoday.com/travel/
3. https://www.usatoday.com/story/money/cars/2017/07/03/tesla-model-3/447465001/	https://www.usatoday.com/money/cars/
4. https://www.usatoday.com/story/money/2017/10/06/elon-musk-tesla-can-help-fix-puerto-ricos-ruined-electrical-grid/738782001/	https://www.usatoday.com/money/business/
5. https://www.usatoday.com/story/tech/nation-now/2017/09/29/elon-musk-speech-what-we-learned/715703001/	https://www.usatoday.com/topic/Hurricane-Maria/local
6. https://www.usatoday.com/story/news/nation-now/2017/10/01/tesla-eyes-hurricane-ravaged-caribbean-could-shape-power-grids/721986001/	https://www.usatoday.com/sports/ncaaf/sagarin/
7. https://www.usatoday.com/story/tech/talkingtech/2017/09/14/you-must-watch-elon-musk-video-rocket-fails/665166001/	https://www.usatoday.com/sports/ncaab/teams/
8. https://www.usatoday.com/videos/money/2017/10/06/elon-musk-believes-tesla-can-rebuild-puerto-ricos-power-grid/106363188/	https://www.usatoday.com/sports/ncaab/standings/
9. https://www.usatoday.com/story/life/entertainthis/2017/04/24/amber-heard-elon-musk-dating-public-debut-instagram-posts/100834568/	https://www.usatoday.com/sports/ncaab/sagarin/
10. https://www.usatoday.com/story/driveon/2015/10/09/tesla-elon-musk-apple-hate/73689838/	https://www.usatoday.com/sports/ncaab/scores/

Query2: Star Wars

Lucene	PageRank
1. https://www.usatoday.com/story/life/movies/2015/11/03/star-wars-the-force-awakens-masks-lightsabers-amc-cinemark/75089340/	https://www.usatoday.com/section/global/virtuallythere/
2. https://www.usatoday.com/story/life/movies/2015/12/16/review-star-wars-the-force-awakens/77342318/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
3. https://www.usatoday.com/story/life/movies/2015/12/14/harrison-ford-star-wars-the-force-awakens/77243366/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
4. https://www.usatoday.com/story/life/movies/2015/12/13/jj-abrams-star-wars-the-force-awakens/76911016/	https://www.usatoday.com/money/
5. https://www.usatoday.com/story/life/entertainthis/2015/12/04/star-wars-need-to-know/76685854/	https://www.usatoday.com/tech/
6. https://www.usatoday.com/story/life/entertainthis/2015/12/11/star-wars-movie-definitive-ranking/77137480/	https://www.usatoday.com/life/
7. https://www.usatoday.com/story/life/movies/2017/09/12/j-j-abrams-return-star-wars-direct-episode-ix-replacing-colin-trevorrow/657203001/	https://www.usatoday.com
8. https://www.usatoday.com/story/life/movies/2017/08/31/star-wars-five-best-ways-celebrate-the-last-jedi-force-friday-ii/618268001/	https://www.usatoday.com/policing/
9. https://www.usatoday.com/story/money/business/small-business/	https://www.usatoday.com/travel/

ess-central/2017/10/08/star-wars-last-jedi-trailer-gets-debut-monday-night-football/744286001/	
10. https://www.usatoday.com/story/life/movies/2017/10/09/new-star-wars-last-jedi-trailer-hints-troubling-tie-between-rey-and-kylo-ren/748730001/	https://www.usatoday.com/topic/Las-Vegas-shooting/local

Query3: North Korea

Lucene	PageRank
1. https://www.usatoday.com/story/news/world/2017/06/17/three-americans-still-detained-north-korean-jails/102948008/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
2. https://www.usatoday.com/story/news/2017/06/19/north-korean-treatment-western-prisoners-bizarre-not-always-physically-brutal/103022748/	https://www.usatoday.com/news/
3. https://www.usatoday.com/story/news/world/2016/01/06/north-korea-hydrogen-bomb/78345074/	https://www.usatoday.com/news/
4. https://www.usatoday.com/story/news/world/2017/04/28/seoul-north-korea-test-fires-missile/101044820/	https://www.usatoday.com/policing/
5. https://www.usatoday.com/story/news/2017/07/04/north-korea-ballistic-missile-south-korea/449341001/	https://www.usatoday.com/opinion/
6. https://www.usatoday.com/story/news/world/2017/09/02/north-korea-nuclear-test-seismic-activity/629486001/	https://www.usatoday.com/policing/data-casualties
7. https://www.usatoday.com/story/news/world/2017/07/17/united-states-missile-defense-north-korea-threats/470177001/	https://www.usatoday.com/sports/nc-aaf/

8. https://www.usatoday.com/story/news/world/2017/08/29/north-korea-fires-ballistic-missile-over-japan/611119001/	https://www.usatoday.com/olympics-rio-2016/
9. https://www.usatoday.com/story/news/world/2017/08/08/report-north-korea-has-uke-fits-inside-missile/549188001/	https://www.usatoday.com/travel/experience-america/
10. https://www.usatoday.com/story/news/politics/2017/06/13/north-korea-releases-us-citizen-otto-warmbier-tillerson-says/102806222/	https://www.usatoday.com/shop/

Query4: LA Dodgers

Lucene	PageRank
1. https://www.usatoday.com/sports/mlb/salaries/2016/player/all/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
2. https://www.usatoday.com/sports/mlb/salaries/2015/player/all/	https://www.usatoday.com/life/
3. https://www.usatoday.com/sports/mlb/	https://www.usatoday.com/life/
4. https://www.usatoday.com/sports/mlb/	https://www.usatoday.com/olympics-rio-2016/
5. https://www.usatoday.com/sports/mlb/salaries/2012/player/all/	https://www.usatoday.com/travel/experience-america/
6. https://www.usatoday.com/sports/mlb/salaries/	https://www.usatoday.com/sports/boxing/

7. https://www.usatoday.com/sports/mlb/salaries/2017/playerr/all/	https://www.usatoday.com/travel/usa-today-eats/
8. https://www.usatoday.com/story/sports/mlb/2017/10/09/leading-off-watching-weather-in-boston-greinke-vs-dodgers/106458414/	https://www.usatoday.com/sports/mlb/
9. https://www.usatoday.com/sports/mlb/salaries/2014/playerr/all/	https://www.usatoday.com/sports/ncaa/b/
10. https://www.usatoday.com/sports/mlb/salaries/2009/playerr/all/	https://www.usatoday.com/sports/nba/

Query5: Puerto Rico

Lucene	PageRank
1. https://www.usatoday.com/story/news/politics/2017/09/29/acting-homeland-security-director-calls-reach-hurricane-maria-a-good-news-story-mayor-san-juan-outra/716464001/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
2. https://www.usatoday.com/story/news/nation-now/2017/10/04/torres-aid-mission-puerto-rico/734371001/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
3. https://www.usatoday.com/story/news/nation-now/2017/10/08/aid-makes-needy-hurricane-maria-survivors-puerto-rico/744978001/	https://www.usatoday.com/tech/
4. https://www.usatoday.com/story/news/politics/2017/09/29/trump-hurricane-damaged-puerto-rico-weve-never-seen-situation-like-this/716531001/	https://www.usatoday.com/news/

5. https://www.usatoday.com/story/news/nation-now/2017/10/05/husband-wife-have-surprise-reunion-puerto-rico-airport/737828001/	https://www.usatoday.com/weather/
6. https://www.usatoday.com/story/news/world/2017/09/20/puerto-rico-braces-direct-hit-hurricane-maria/684121001/	https://www.usatoday.com/weather/
7. https://www.usatoday.com/story/news/nation/2017/09/26/why-puerto-rico-faces-monumental-recovery-effort/703515001/	https://www.usatoday.com
8. https://www.usatoday.com/story/news/2017/09/24/puerto-rico-dam-holding-some-local-officials-say-threat-overblown/697748001/	https://www.usatoday.com/opinion/
9. https://www.usatoday.com/story/news/world/2017/09/21/puerto-ricans-hurricane-maria/688060001/	https://www.usatoday.com/tech/talkingtech/
10. https://www.usatoday.com/story/weather/hurricanes/2017/10/05/puerto-rico-clean-water/734813001/	https://www.usatoday.com/travel/roadwarriorvoices/

Query6: Hurricane Harvey

Lucene	PageRank
1. https://www.usatoday.com/story/opinion/2017/08/28/hurricane-harvey-shadowed-two-lessons-2005-editorials-debates/610503001/	https://www.usatoday.com/section/global/virtuallythere/
2. https://www.usatoday.com/story/news/nation/2017/09/04/houston-open-business/631151001/	https://www.usatoday.com/s hop/

3. https://www.usatoday.com/story/news/nation/2017/08/28/hearbreak-texas-harvey-drive-30-000-shelters-fema-says/607224001/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
4. https://www.usatoday.com/story/news/2017/08/27/cajun-navy-heads-texas-aid-rescues/606883001/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
5. https://www.usatoday.com/story/opinion/faith-in-america/2017/08/30/storm-the-church-bigger-than-joel-osteens-building/618795001/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
6. https://www.usatoday.com/story/news/nation-now/2017/08/28/harvey-nursing-home-residents-rescued-after-viral-photo-safe-warm-dry/610575001/	https://www.usatoday.com/money/
7. https://www.usatoday.com/story/news/nation-now/2017/08/27/harvey-nursing-home-residents-rescued-floodwaters-after-photo-goes-viral/606987001/	https://www.usatoday.com/tech/
8. https://www.usatoday.com/story/news/nation-now/2017/08/31/hurricane-harvey-devastated-texas-see-wreckage-drone/619985001/	https://www.usatoday.com/money/markets/
9. https://www.usatoday.com/story/news/nation/2017/08/30/joe-l-osteen-we-never-turned-away-hurricane-harvey-flooding-victims/615569001/	https://www.usatoday.com/news/
10. http://www.usatoday.com/picture-gallery/news/nation/2017/08/23/tropical-depression-harvey-batters-texas/104899472/	https://www.usatoday.com/sports/

Query7: iPhone X

Lucene	PageRank
--------	----------

1. https://www.usatoday.com/story/tech/reviewedcom/2017/09/12/should-you-upgrade-to-the-new-iphone-8-or-iphone-x/105546142/	https://www.usatoday.com/section/global/virtuallythere/
2. https://www.usatoday.com/story/tech/2017/09/13/iphone-x-too-good-iphone-8-8-plus-may-soon-find-out/663345001/	https://www.usatoday.com/s hop/
3. https://www.usatoday.com/story/tech/reviewedcom/2017/09/25/7-great-cases-that-will-work-with-the-new-iphone-8-and-8-plus/105985358/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
4. https://www.usatoday.com/story/tech/columnist/baig/2017/09/12/iphone-8-how-removing-home-button-change-your-iphone-experience/641540001/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
5. https://www.usatoday.com/story/tech/talkingtech/2017/09/11/apples-iphone-event-heres-how-watch/653065001/	https://www.usatoday.com/story/news/2013/01/09/corrections-clarifications/1821023/
6. https://www.usatoday.com/story/tech/columnist/2017/09/09/iphone-8-how-sell-your-old-iphone-most-money/649373001/	https://www.usatoday.com/money/
7. http://www.usatoday.com/picture-gallery/tech/2017/06/23/iphone-8-evolution-of-the-iphone-over-the-years/103144056/	https://www.usatoday.com/tech/
8. http://www.usatoday.com/picture-gallery/tech/2017/06/23/iphone-8-evolution-of-the-iphone-over-the-years/103144056/	https://www.usatoday.com/money/markets/
9. https://www.usatoday.com/story/tech/talkingtech/2017/09/13/skip-iphone-x-and-8-if-you-need-iphone-better-deal/660743001/	https://www.usatoday.com/news/
10.	

https://www.usatoday.com/story/tech/talkingtech/2017/09/19/apple-iphone-consumers-waiting-iphone-x/681952001/	https://www.usatoday.com/sports/
---	---

Query 8: Paris Climate Deal

Lucene	PageRank
https://www.usatoday.com/story/news/nation/2017/05/31/onl-y-two-nations-out-197-oppose-climate-pact---and-us-may-ne xt/102360164/	https://www.usatoday.com/story/n ews/2013/01/09/corrections-clarifi cations/1821023/
https://www.usatoday.com/story/news/2017/06/01/trump-wit hdraw-paris-climate-deal-5-things-you-need-know-thursday/ 102286628/	https://www.usatoday.com/story/n ews/2013/01/09/corrections-clarifi cations/1821023/
https://www.usatoday.com/story/news/politics/2017/09/16/re ports-trump-administration-may-not-pull-out-paris-climate-a greement/673988001/	https://www.usatoday.com/tech/
https://www.usatoday.com/story/news/world/2017/06/01/tru mp-paris-agreement-climate-change-world-reaction/1023909 36/	https://www.usatoday.com/news/
https://www.usatoday.com/story/news/politics/2016/02/09/su preme-court-halts-obamas-emissions-rule/80085182/	https://www.usatoday.com/sports/
https://www.usatoday.com/story/news/politics/2016/02/09/su preme-court-halts-obamas-emissions-rule/80085182/	https://www.usatoday.com/life/
https://www.usatoday.com/story/news/2017/04/20/science-m arch-war-truth-political-polarization/100636124/	https://www.usatoday.com
https://www.usatoday.com/story/opinion/2017/10/02/north-k orea-trump-needs-china-but-hes-doing-it-wrong-max-baucus	https://www.usatoday.com/policin g/

-ryan-hass-column/721378001/	
https://www.usatoday.com/story/news/2014/12/01/five-things-to-know-monday/19703739/	https://www.usatoday.com/travel/
https://www.usatoday.com/story/news/2014/12/01/five-things-to-know-monday/19703739/	https://www.usatoday.com/opinion/

Ch5.Overlap graph:

There is no overlap of all the query between pagerank and lucene. However, for query: LA Dodgers

There are two links similar:

Lucene: a series links whose root dir is: <https://www.usatoday.com/sports/mlb/>

PageRank: <https://www.usatoday.com/sports/mlb/> which is index page for all of items in Lucene.

The screenshot shows the USA Today website. At the top, there's a navigation bar with links for NEWS, SPORTS, LIFE, MONEY, TECH, TRAVEL, OPINION, CROSSWORDS, WASHINGTON, THANKSGIVING, and MORE. A 'SUBSCRIBE NOW' button is also visible. Below the navigation bar is a large banner for 'LEARN SOMETHING NEW TODAY. Up to 75% off online courses.' with a 'START LEARNING' button. The main content area is divided into sections: 'SPORTS' (with sub-links for NFL, MLB, NBA, NHL, NCAAF, OLYMPICS, NCAAB, GOLF, NASCAR, UFC, FANTASY, and MORE), 'MLB' (with sub-links for Main, Scores, Schedule, Standings, Statistics, Odds, Salaries, Teams, and More), and 'TOP STORIES'. The 'TOP STORIES' section lists several headlines related to MLB, including 'Twins outfielder named defensive player of year', 'Ready to manage again, Wedge interviews with NY Yankees', 'FTW: Why trading Giancarlo Stanton makes no sense', 'Yankees interview former Indians, M's manager', 'Upton named 2017 Tiger of the Year, despite trade', 'Halladay family 'heartbroken,' to hold memorial', 'MLB free agents: Top 73 on the market', 'Stanton's 2018 status with Marlins in doubt', and 'Agent: D-Backs expected to pursue J.D. Martinez'. A 'FOLLOW USA TODAY SPORTS' section with social media icons is also present. The bottom of the page features a 'RIGHT NOW' section.