Frequently Asked Questions (FAQ)

1.Where and how to download website HTML files?

You can download them from

- **https://drive.google.com/drive/folders/14qtJywCRk00wEbmt-Pv3efyPIBqwcoZr**

Please use your USC email to request access to your news website files. You should download the folder for the newssite that is assigned to you based on your USC ID.

USC ID ends with	News Sites to Crawl	NewsSite Name	Root URL
01~30	NY Times	nytimes	https://www.nytimes.com
31~60	Fox News	foxnews	https://www.foxnews.com
61~00	Los Angeles Times	latimes	https://www.latimes.com

2. Error while manually mounting the guest additions?

You might encounter an error –“**modprobevboxsf failed”**while running the VBoxLinuxAdditions.run. That’s a problem with the virtual box. Just restart the system and verify that the guest additions are working.

3. The instruction of hw4 suggests that we should install ubuntu on windows so as to install solr more conveniently. However. if I do not want to install ubuntu, may I install solr and other software needed directly to windows system?

You can. However, if you run into some issues when installing directly on Windows, we may not be able to guide you. You will be on your own.

4. Are we allowed to use any language and platform instead of PHP for the Web pages?

Yes, you can use node.js or any other language, but you need to check its compatibility with Solr at your end only. However no help will be provided for the code implementation for this homework.

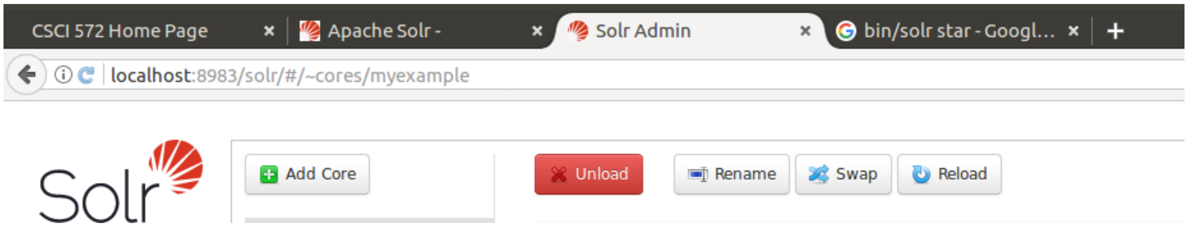
[1] Some news websites may produce fewer than 20,000 results, e.g. yahoo, which is not a problem

5. There is no step-wise instruction given for installing solr. Can you guide me in installation process? I downloaded Solr-7.1.0 zip.

You need to unzip it and then follow the instructions in "Using Solr to Index a Web Site" document.Once you are done with the Ubuntu installation steps, the Solr setup instructions are to be executed inside Ubuntu. So, setting up Java 8, downloading, unzipping Solr, Apache, PHP etc. all should be done inside Ubuntu.

6. How to reload our core and query with new configuration.

There will be a reload button in the Solr instance's core admin: Explore the Solr instance, it's various configurations and settings.



7.Do I need to install an Apache server on our computer in addition to running the Solr server?

Apache server will be installed by default in Linux machines. in the assignment we are using PHP script to communicate to the Solr. As we know PHP is a server-side scripting language and it requires Apache or any other web servers.

However, if you run into any problem then use the below commands on Ubuntu terminal.

To install Apache on Ubuntu

sudo apt-get install apache2

To install PHP on Ubuntu

sudo apt-get install libapache2-mod-php

8. Permission denied while running solr?

Try the binary version of solr instead of the src version.

If you are indeed running the binary version of solr – you need to check if you have permission to access and execute the files inside solr.

Specifically, to run bin/solr you might have to change access permissions -

Use: **chmod 777 bin/solr** or **sudo chmod 777 bin/solr**

9.Can I run solr in cloud mode?

You don't need cloud mode for this assignment. Only use "bin/solr start" to start running the Solr instance in stand-alone

To Restart Solr-

First do, "bin/solr stop -all". Then "bin/solr start"

10.I am not mentioning my top 10 results in the report. I am maintaining separate text files for each query. Is that allowed or shall I mention the results only in the report?

Include all the results in a single report.

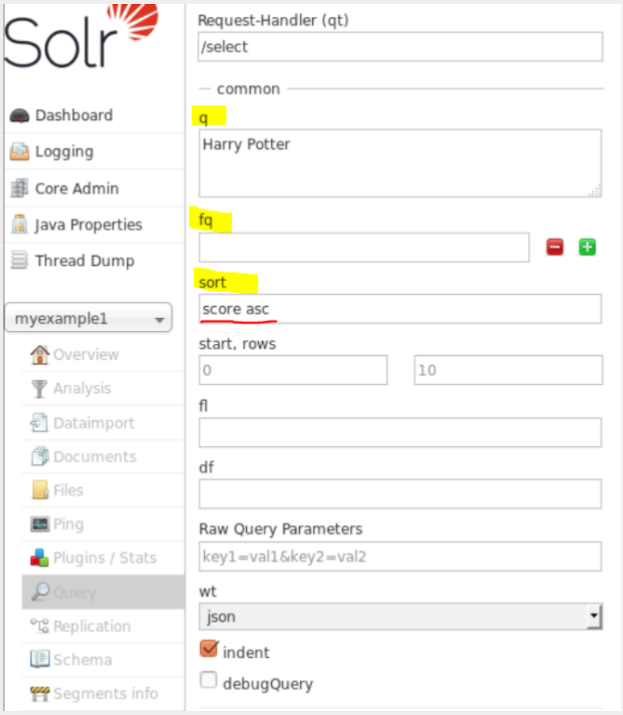
11.I am getting zero overlaps between Default and pagerank algorithms. Could you please let us know how many points would be deducted for zero overlaps?

As long as you follow the assignment steps and implement it correctly, you won't be penalized. Just make sure your steps are clear in your report and you should be fine.

12.How to use additional parameters from code?

Here is an example of using Additional Parameters:

Please notice the below parameters when querying Solr from the dashboard: q, fq, sort, etc.,



Request-Handler (qt): /select

q: Harry Potter

fq:

sort: score asc

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

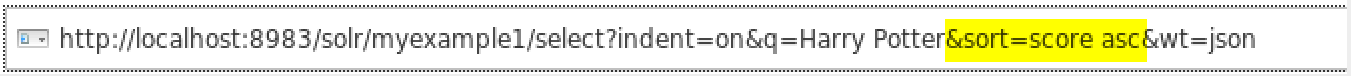
☐ debugQuery

When querying Solr from the dashboard, you enter the query in the q parameter input box-> “Harry Potter”.
You may use the other parameters as well to further refine the query.
Ex: sort parameter's default value when nothing is specified is “score desc”, i.e. the results are sorted and displayed from highest score to lowest score. You can change to “score asc”.(For demo purposes)

How do you specify this in the PHP code? By sending it as an additional parameter in the call for Solr:

```
$additionalParameters = array(
 'sort' => 'score asc'
);
$results = $solr->search($query, 0, $limit, $additionalParameters);
```

URL ->



http://localhost:8983/solr/myexample1/select?indent=on&q=Harry Potter&sort=score asc&wt=json

Hope this clears your doubts on how to use additional parameters from code, and why you were getting the error.
Further reading -> https://lucene.apache.org/solr/guide/7_1/common-query-parameters.html

13.Do we need to sort results by any attribute in the first algorithm?

No. It is already sorted by default value -> "score desc".

14.How to use NetworkX and JSOUP?

Both libraries need to be used in different parts of the assignment. JSOUP is used in JAVA program to extract links from given HTML pages and create edge list file. NetworkX is used in Python to calculate pagerank for all pages.

15.What is the use of Java code?

The Java code creates the edgelist.txt file. You must use this as input for Python code, which creates a graph using NetworkX library.

16. Do we have to write page rank algorithm?

The default search result is returned based on Lucene's ranking algorithm.

The next step is to get the results based on the page ranking score. You do not need to write the algorithm yourself. It is implemented in NetworkX library. You use it to obtain an external file with page ranking scorings, which you add to Solr. Please see the document “SolrExercise.docx [Step 2]”

17. If that page doesn't contain description information, we don't need to show description?

That's ok. You can show Null or N/A.

18.Getting much better results with ‘pageRankFileasc’ as opposed to ‘pageRankFile desc’. Which one shall I use?

Solr uses Boolean model to narrow down the range of the documents that needs to be scored. This will be done based on the query term. Page rank "desc" will next sort and return these from pages from highest score to lowest score.

The pages having highest page rank scores (maybe a homepage, or some other popular page with many links to it) may not be most relevant to query term. You can use anyone of them but it is recommended to rank the pages in the decreasing order of their ranks.

19.What should be the format of score in external_pageRankFile?

It can be exponential or decimal format, but exponential is better as it helps Solr(Lucene) in ranking the pages properly.

20. Handling duplicates?

Filter duplicates based on the ID of the returned page. i.e the ID or a page **must** be unique – it is okay if the URLs corresponding to two unique IDs are the same. If you find duplicate URLs in top 10 results, you can just handle them as if they are unique (No need to worry about it during calculations or display). Sometimes news websites publish the same page at 2 different places in their domain and forget to take old one down after the new one is published.

21.How should the front end look like?

You are free to implement any UI which resembles the results returned by any search engine. Don't forget to have any radio button or other equivalent implementation to toggle between Lucene and pageRank ranking algorithms.

22.What all information should be shown after search button is pressed?

Title, URL(og_url), ID, Description(og_description), If Description is not available, just display NA. If URL is missing, for HW4, fetch from the provided map file. Title and URL should be clickable.

23.Should we use quotes with 8 queries?

No, don't use any quotes while entering the queries.

24.Are we allowed to use whichever Solr Client Libraries within PHP and Node.js categories on this page https://cwiki.apache.org/confluence/display/solr/IntegratingSolr?

Yes, you can use any as long as search works fine.

25.what should be format in pageRankFile.txt file?

<full_file_path/doc_id>=<page_rank_score>

Note: the full file path should be absolute path of your local machine

26.Where can I find more information about how "href:abs" works?

https://jsoup.org/cookbook/extracting-data/working-with-urls

27.What should be the naming convention for page_rank file?

It should be external_<filename>.txt also your pagerank filename and field name in managed Schema should be same.

28.Where can I check if the Solr results doesn't show up properly?

Check error logs in /var/log/apache2/error.log

29.Order of the search results in the display page

Order doesn't matter but should contain the following results

1. Title (clickable)

2. URL (clickable)

3. ID

4. Description

All the URLs have to be retrieved from the Solr query. If the URL is not present, it needs to be accessed from the CSV file that contains the URL to HTML mappings.

30.Should we consider src and link[href] for edgeList creation?

Consider only a[href] as it is in the example given in assignment

31.external_pageRankFile should be the same for everyone for a given News website?

Yes

32. deliverables for home work 4

1. The external page rank file

2. The report containing all the written details - screenshots, table etc.

3. The working source code in a folder.

DO NOT include the code you used to generate EdgeList and PageRank values

33. Error while creating edgelist.

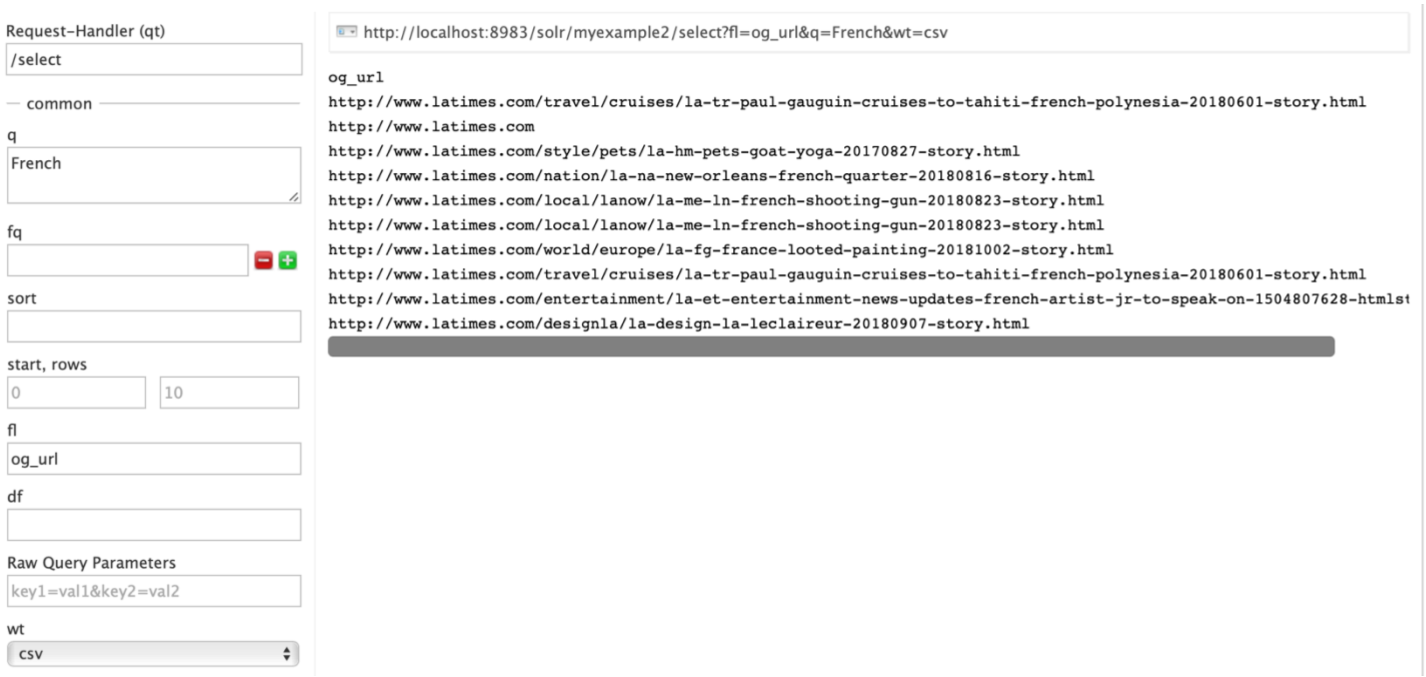
If you encounter an error - **BaseURI must not be null**, then check your folder where all the html files are present. There will be a file named .DS_store which is causing the error. Delete that file and re-run your code.

34. Which attribute should I use in JSoup? href or abs:href

Please use **abs:href**Refer https://jsoup.org/cookbook/extracting-data/working-with-urls for more details

35. What is the simplest way to collect the top ten results for each query?

When querying the Solr UI, you can configure the search to look like this:



Request-Handler (qt)
/select

common

q
French

fq

sort

start, rows
0 10

fl
og_url

df

Raw Query Parameters
key1=val1&key2=val2

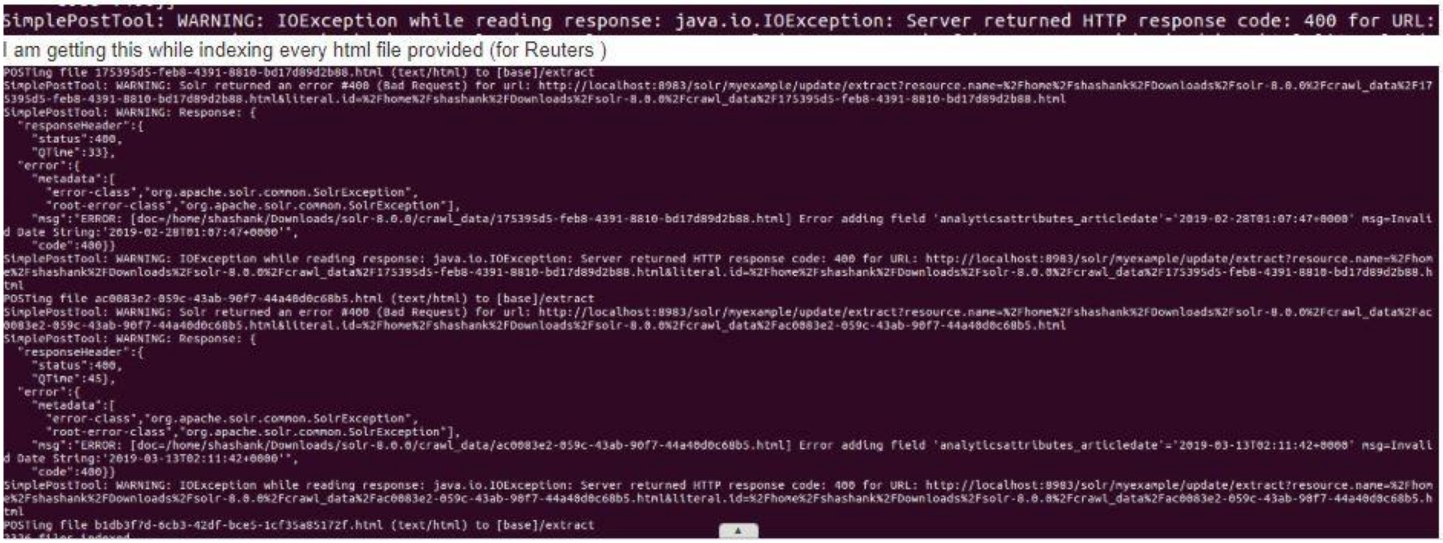
wt
csv

http://localhost:8983/solr/myexample2/select?fl=og_url&q=French&wt=csv

```
og_url
http://www.latimes.com/travel/cruises/la-tr-paul-gauguin-cruises-to-tahiti-french-polynesia-20180601-story.html
http://www.latimes.com
http://www.latimes.com/style/pets/la-hm-pets-goat-yoga-20170827-story.html
http://www.latimes.com/nation/la-na-new-orleans-french-quarter-20180816-story.html
http://www.latimes.com/local/lanow/la-me-ln-french-shooting-gun-20180823-story.html
http://www.latimes.com/local/lanow/la-me-ln-french-shooting-gun-20180823-story.html
http://www.latimes.com/world/europe/la-fg-france-looted-painting-20181002-story.html
http://www.latimes.com/travel/cruises/la-tr-paul-gauguin-cruises-to-tahiti-french-polynesia-20180601-story.html
http://www.latimes.com/entertainment/la-et-entertainment-news-updates-french-artist-jr-to-speak-on-1504807628-htmlst
http://www.latimes.com/designla/la-design-la-leclaireur-20180907-story.html
```

Be sure to set fl=og_url and wt=csv. You can also set a limit on the number of rows to display only the top 10. You can then copy-paste the URLs into a table in Word.

36. I am getting error "SimplePostTool: Warning: IOException while reading response.." while indexing the files for solr



```
SimplePostTool: WARNING: IOException while reading response: java.io.IOException: Server returned HTTP response code: 400 for URL:
```

I am getting this while indexing every html file provided (for Reuters)

```
POSTing file 175395d5-feb8-4391-8810-bd17d89d2b88.html (text/html) to [base]/extract
SimplePostTool: WARNING: Solr returned an error #400 (Bad Request) for url: http://localhost:8983/solr/myexample/update/extract?resource.name=%2Fhome%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2F17
5395d5-feb8-4391-8810-bd17d89d2b88.html&literal.id=%2Fhome%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2F175395d5-feb8-4391-8810-bd17d89d2b88.html
SimplePostTool: WARNING: Response: {
  "responseHeader":{
    "status":400,
    "QTime":33},
  "error":{
    "metadata":[
      "error-class","org.apache.solr.common.SolrException",
      "root-error-class","org.apache.solr.common.SolrException"],
    "msg":"ERROR: [doc=/home/shashank/Downloads/solr-8.0.0/crawl_data/175395d5-feb8-4391-8810-bd17d89d2b88.html] Error adding field 'analyticsattributes_articledate'='2019-02-28T01:07:47+0000' msg=Invali
d Date String:'2019-02-28T01:07:47+0000'",
    "code":400}}
SimplePostTool: WARNING: IOException while reading response: java.io.IOException: Server returned HTTP response code: 400 for URL: http://localhost:8983/solr/myexample/update/extract?resource.name=%2Fhom
e%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2F175395d5-feb8-4391-8810-bd17d89d2b88.html&literal.id=%2Fhome%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2F175395d5-feb8-4391-8810-bd17d89d2b88.h
tml
POSTing file ac0083e2-059c-43ab-90f7-44a40d0c68b5.html (text/html) to [base]/extract
SimplePostTool: WARNING: Solr returned an error #400 (Bad Request) for url: http://localhost:8983/solr/myexample/update/extract?resource.name=%2Fhome%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2Fac
0083e2-059c-43ab-90f7-44a40d0c68b5.html&literal.id=%2Fhome%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2Fac0083e2-059c-43ab-90f7-44a40d0c68b5.html
SimplePostTool: WARNING: Response: {
  "responseHeader":{
    "status":400,
    "QTime":45},
  "error":{
    "metadata":[
      "error-class","org.apache.solr.common.SolrException",
      "root-error-class","org.apache.solr.common.SolrException"],
    "msg":"ERROR: [doc=/home/shashank/Downloads/solr-8.0.0/crawl_data/ac0083e2-059c-43ab-90f7-44a40d0c68b5.html] Error adding field 'analyticsattributes_articledate'='2019-03-13T02:11:42+0000' msg=Invali
d Date String:'2019-03-13T02:11:42+0000'",
    "code":400}}
SimplePostTool: WARNING: IOException while reading response: java.io.IOException: Server returned HTTP response code: 400 for URL: http://localhost:8983/solr/myexample/update/extract?resource.name=%2Fhom
e%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2Fac0083e2-059c-43ab-90f7-44a40d0c68b5.html&literal.id=%2Fhome%2Fshashank%2FDownloads%2Fsolr-8.0.0%2Fcrawl_data%2Fac0083e2-059c-43ab-90f7-44a40d0c68b5.h
tml
POSTing file b1db3f7d-6cb3-42df-bce5-1cf35a85172f.html (text/html) to [base]/extract
```

If you are getting an error similar to – "SimplePostTool: WARNING: IOException while reading response: java.io.FileNotFoundException: http://localhost:8983/solr/myexample/update/extract?resource.name=%2Fhome%2Fhw4%2Fnytimes%2F39f07542-ee59-4254-ade2-0d393aa7e360.html&literal.id=%2Fhome%2Fhw4%2Fnytimes%2F39f07542-ee59-4254-ade2-0d393aa7e360.html" for almost all the HTML files, then you can either downgrade to Solr version 7.7.2, or follow the below steps.

- Add the below code to solrconfig.xml –

```
<lib dir="../../extract" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/extraction/lib" regex=".*\.jar" />
```

```
<lib dir="${solr.install.dir:../../../..}/contrib/dataimporthandler/lib/" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-dataimporthandler-.*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/extraction/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-cell-\d.*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/clustering/lib/" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-clustering-\d.*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/langid/lib/" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-langid-\d.*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/velocity/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-velocity-\d.*\.jar" />
<requestHandler name="/update" class="solr.UpdateRequestHandler">
</requestHandler>
<requestHandler name="/update/extract"
        startup="lazy"
        class="solr.extraction.ExtractingRequestHandler" >
<lst name="defaults">
<str name="lowernames">true</str>
<str name="uprefix">ignored_</str>
<!-- capture link hrefs but ignore div attributes -->
<str name="captureAttr">true</str>
<str name="fmap.a">links</str>
<str name="fmap.div">ignored_</str>
</lst>
</requestHandler>
```

After making the above changes to the file, save the file. Restart solr with “bin/solr restart” command and rerun the command ***bin/post –c myexample –filetypes html crawl_data/***
If you still get the errors, reinstall solr (with version 7.7.X).

37. I am getting SolrCore Initialization Failures in Solr UI/Dashboard

Try the following:

- Delete all other cores, and keep only one core on your system. To delete a core, start solr if not already running (bin/solr start), then run the command “bin/solr delete -c corename”
- Clear browser cache
- Restart browser

38. How to handle overlaps/duplicate URLs

Case 1

Lucene	**PageRank**
www.latimes.com/A	www.latimes.com/A
www.latimes.com/A	www.latimes.com/A
www.latimes.com/B	www.latimes.com/B

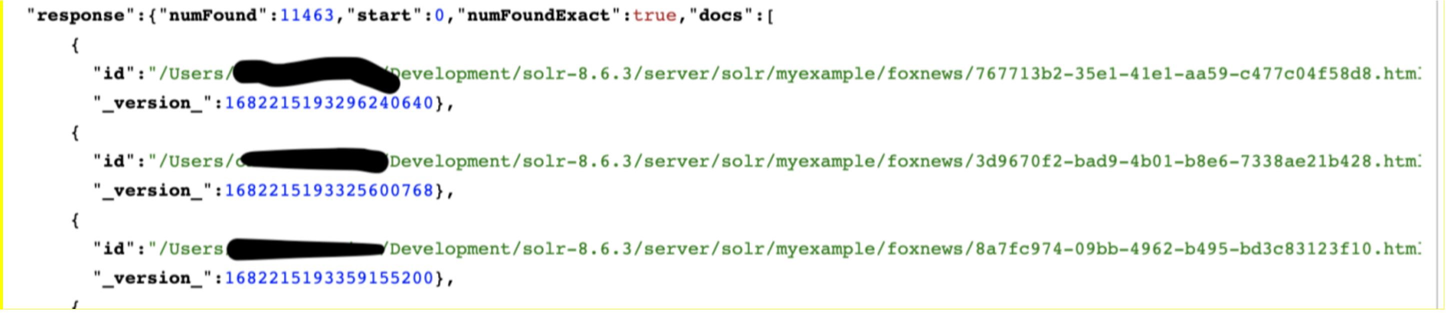
In this case, Lucene has 2 duplicates (www.latimes.com/A) but we will treat both as unique URLs. So you can consider the secondlink as different than first link. Similarly in PageRank, You can consider the secondlink as different from the first link.Now, in this case,there is an overlap of 3 in between Lucene results and PageRank results. All entries are the same for both the algorithms.

Case 2

Lucene	PageRank
www.latimes.com/A	www.latimes.com/A
www.latimes.com/B	www.latimes.com/A
www.latimes.com/C	www.latimes.com/B

In this case, theoverlap between Lucene results and PageRank results is 2. Here in PageRank, we are considering the second entrywww.latimes.com/Ato be different from thefirst link. So, in this case, www.latimes.com/Aand www.latimes.com/B each appeared once in both algorithms whereas duplicate www.latimes.com/A from PageRank did not appear in Lucene so it's not an overlap.

39. Post indexing the HTML files and upon running a query, only *id* and *_version_* fields are being seen in the returned JSON?



```
"response":{"numFound":11463,"start":0,"numFoundExact":true,"docs":[
    {
      "id":"/Users/[redacted]Development/solr-8.6.3/server/solr/myexample/foxnews/767713b2-35e1-41e1-aa59-c477c04f58d8.htm
      "_version_":1682215193296240640},
    {
      "id":"/Users/[redacted]Development/solr-8.6.3/server/solr/myexample/foxnews/3d9670f2-bad9-4b01-b8e6-7338ae21b428.htm
      "_version_":1682215193325600768},
    {
      "id":"/Users[redacted]/Development/solr-8.6.3/server/solr/myexample/foxnews/8a7fc974-09bb-4962-b495-bd3c83123f10.htm
      "_version_":1682215193359155200},
    {
```

Please use version Solr 7.7.3 for implementing the homework. There can be issues with indexing the files in the latest version of Solr.

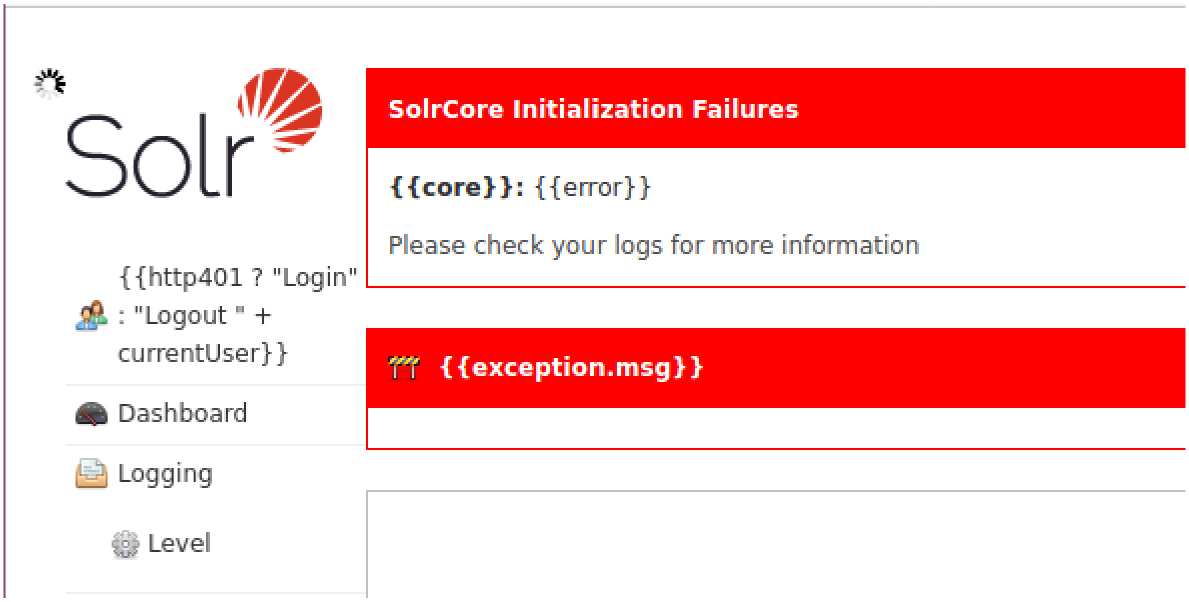
40. Do I need to host the website on any cloud platform for implementing hw#4?

No, Everything (Solr server, web server and other tasks) needs to run and can be performed locally. You do not need to host your website for this exercise.

41. How long does it typically take to index the news site files?

It should not take as long as the previous homeworks as we are not providing a politeness delay. Typically half hour should be sufficient to finish indexing. However, these times can vary based on the computation power of your machine.

42. Unable to start Solr UI



SolrCore Initialization Failures

{{core}}: {{error}}

Please check your logs for more information

{{exception.msg}}

Open in incognito. If you don't see this error when in incognito, switch back to a normal window, open Solr UI, and go to Developer Tools -> Application -> Clear Storage.

Here is a SO link to help resolve this issue:

https://stackoverflow.com/questions/56262704/solr-solrcore-initialization-failures-core-error

43. What should be the format of edgeList file?

It should be “ID ID”. For example: /some_path/abc.html /some_path/xyz.html

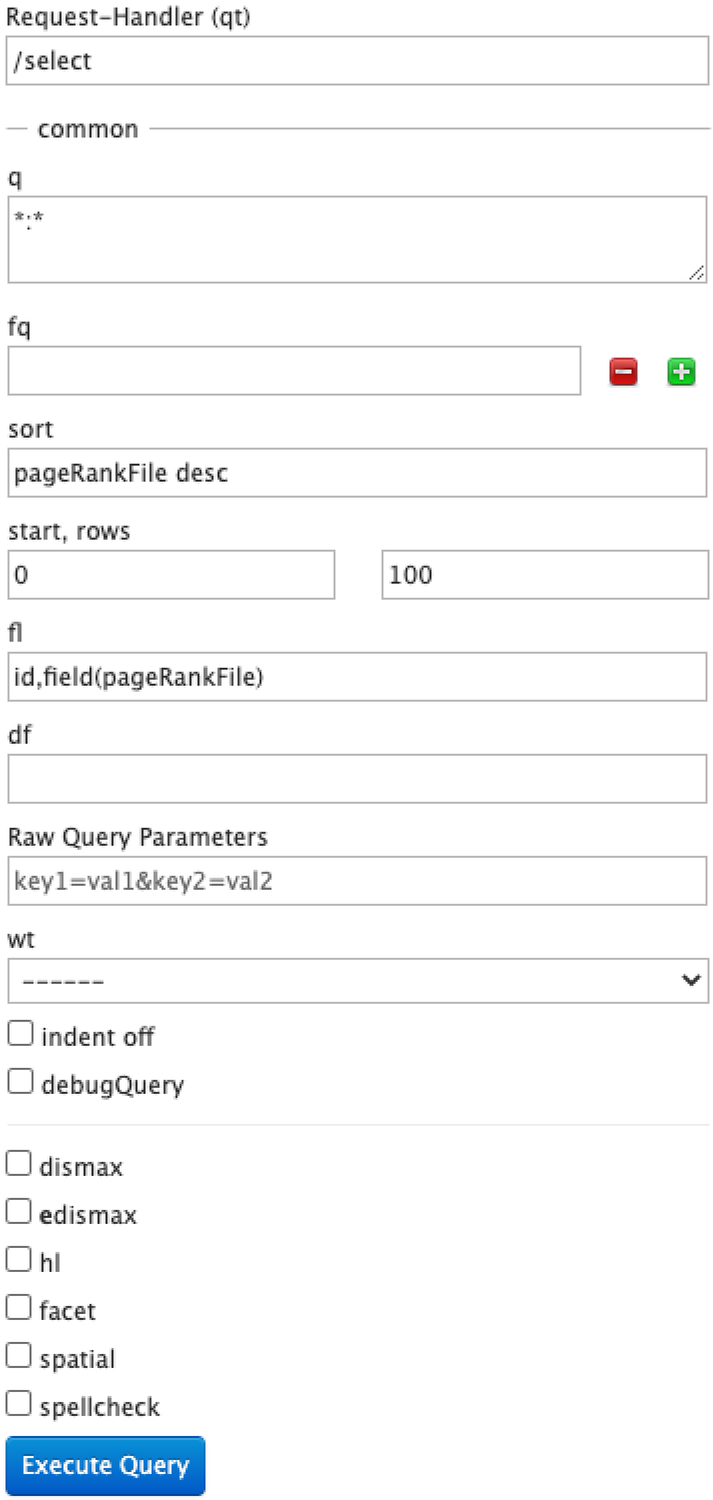
44. How to retrieve and display PageRank scores for results of a query

If you'd like to retrieve and display PageRank scores for each result item for a query,
set fl property to *,field(pageRankFile) (for all fields + PageRank score)
or id,field(pageRankFile) (for only ID and PageRank score).

Works in Solr web UI as well as API calls.

This may be helpful while debugging and to verify that the PageRank scores are correctly being retrieved from external_pageRankFile.

Web UI query example:



Request-Handler (qt)

/select

common

q

:

fq

sort

pageRankFile desc

start, rows

0

100

fl

id,field(pageRankFile)

df

Raw Query Parameters

key1=val1&key2=val2

wt

□ indent off

□ debugQuery

□ dismax

□ edismax

□ hl

□ facet

□ spatial

□ spellcheck

45. I have added the externalpageRankFile but it is not making any effect on the rankings

Please make sure that the pageRankFile is uploaded to the correct folder. The PageRank file must be put in **core-name/data/** folder.

46. Solution for students facing "SimplePostTool: WARNING: IOException" error during indexing

For students who are getting the error similar to "SimplePostTool: WARNING: IOException while reading response: java.io.FileNotFoundException:" for a large number of files, here's a solution:

1. Download the latest binary release of Solr (8.8.2) from https://solr.apache.org/downloads.html

2. After installing the latest version of Solr (8.8.2), follow steps 1-4 given in IndexingwithTIKAV3.pdf

3. Add the following code to solrconfig.xml:

```
<lib dir="../../extract" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/extraction/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/dataimporthandler/lib/" regex=".*\.jar"
/>
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-dataimporthandler-.*\.jar"
/>
<lib dir="${solr.install.dir:../../../..}/contrib/extraction/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-cell-\d.*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/clustering/lib/" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-clustering-\d.*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/langid/lib/" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-langid-\d.*\.jar" />
<lib dir="${solr.install.dir:../../../..}/contrib/velocity/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../..}/dist/" regex="solr-velocity-\d.*\.jar" />
<requestHandler name="/update" class="solr.UpdateRequestHandler">
 </requestHandler>
 <requestHandler name="/update/extract"
 startup="lazy"
 class="solr.extraction.ExtractingRequestHandler" >
 <lst name="defaults">
 <str name="lowernames">true</str>
 <str name="uprefix">ignored_</str>
 <!-- capture link hrefs but ignore div attributes -->
 <str name="captureAttr">true</str>
 <str name="fmap.a">links</str>
 <str name="fmap.div">ignored_</str>
 </lst>
 </requestHandler>
```

4. Since the latest version of Solr does not extract out all of the fields automatically, you need to change the managed-schema.xml to extract the required fields. The 'id' and '_version_' are the only fields extracted automatically. In order to extract the other required fields (title, URL, and description), add the following code to the managed-schema.xml file:

NOTE: Use 'title' instead of 'og_title'

```
 <field name="title" type="string" indexed="false"/>
 <field name="og_url" type="string" indexed="false"/>
 <field name="og_description" type="string"/>
```

5. Restart Solr using the command:

```
bin/solr restart
```

6. Follow step 5 from IndexingwithTIKAV3.pdf to index the HTML files. Now, the IOException should occur only for a few files or none at all.