CSCI 572 - Information Retrieval
Homework 5
Lohith Paripati - paripati@usc.edu - 8299564286

Steps followed to accomplish Homework 5

1. Downloaded **Tika Jar file** from here.

2. Wrote **htmlContentExtracter.java** file and generated **big.txt** with help of Tika jar file as it parses through all html files of news site.

3. Downloaded Peter Norvig's PHP Client (**SpellCorrector.php**) from here.

4. Created php program to invoke php client program and generate **serialized_dictionary.txt** file by sending big.txt as input.
   Note: *ini_set('memory_limit', '-1');* line is added to SpellCorrector program else dictionary won't generate as memory goes out. big.txt, spellcorrector, simple php program invoking spell_corrector are all placed in same folder.

5. Updated **solrconfig.xml** with below lines to implement auto suggest and refreshed my core in solr gui-

```
<searchComponent class="solr.SuggestComponent" name="suggest">
<lst name="suggester">
    <str name="name">suggest</str>
    <str name="lookupImpl">FuzzyLookupFactory</str>
    <str name="field">_text_</str>
    <str name="suggestAnalyzerFieldType">string</str>
 </lst>
 </searchComponent>
  <requestHandler class="solr.SearchHandler" name="/suggest">
    <lst name="defaults">
```

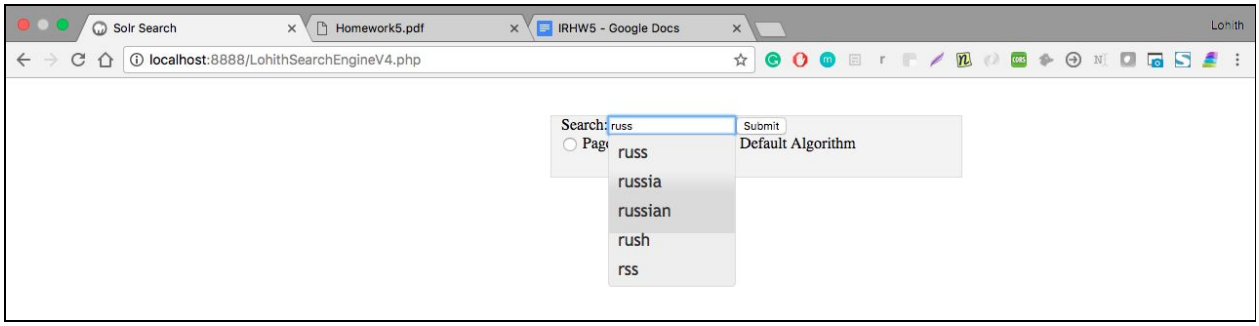
```
<str name ="suggest">true</str>
<str name="suggest.count">5</str>
<str name="suggest.dictionary">suggest</str>
</lst>
<arr name="components">
<str>suggest</str>
</arr>
</requestHandler>
```

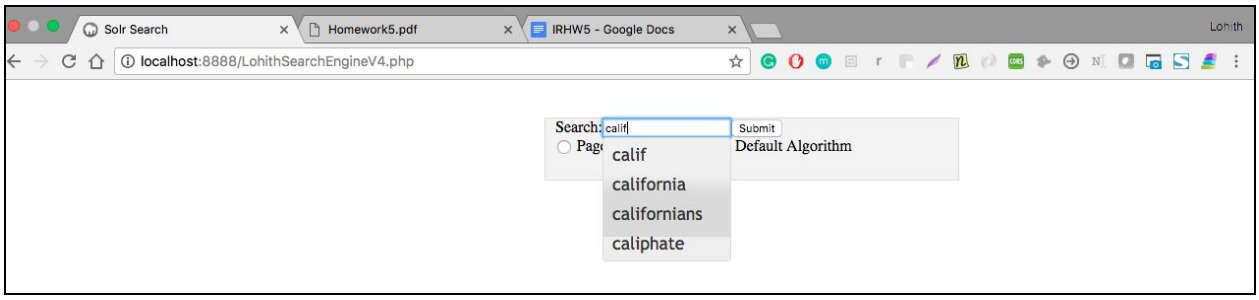
6. Now updated my php file from Assignment-4 to implement auto suggest and spell checker with the resources generated in previous steps such as big.txt, SpellCorrector.php, serialized_dictionary.txt and suggest component in solr.

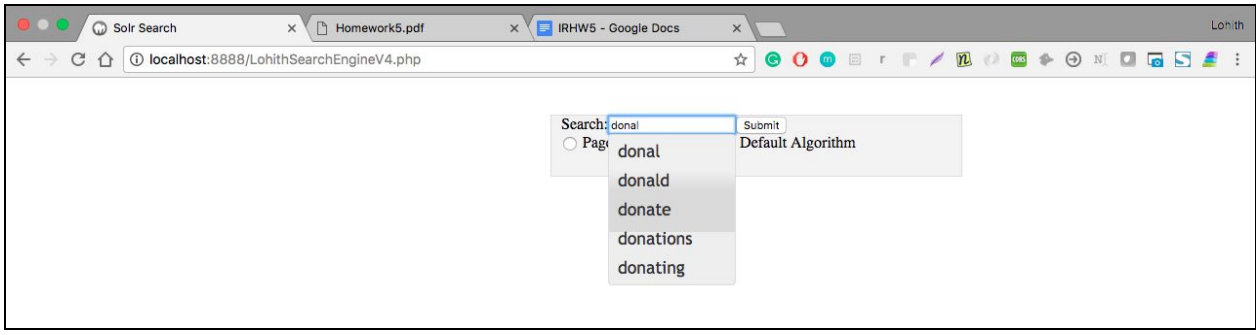
7. For implementing snippet, I used **simple_html_dom.php** which helps in retrieving plain text from all html files indexed and can be used for matching with query terms to display in snippet if match exists.

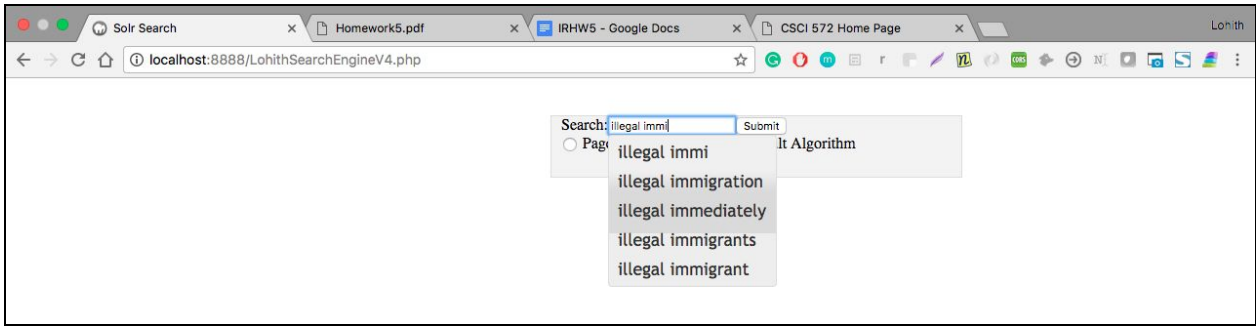
Analysis of the results

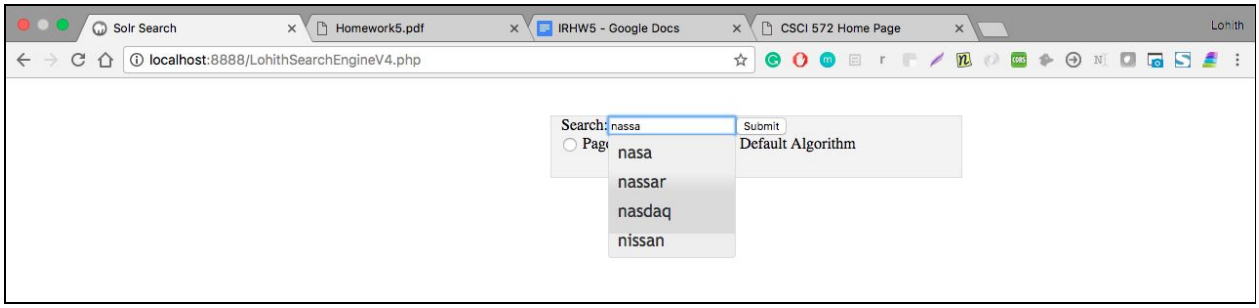
1. **5 sample examples' Screenshots for Auto Suggestions**



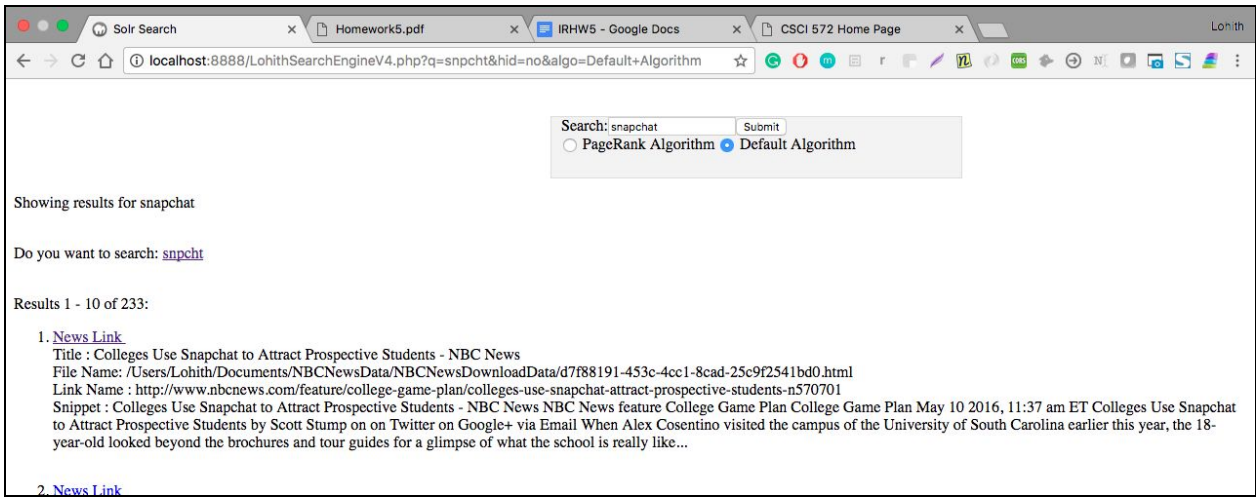

Solr Search
Homework5.pdf
IRHW5 - Google Docs
Lohith
localhost:8888/LohithSearchEngineV4.php
Search: calif
Submit
Default Algorithm
calif
california
californians
caliphate

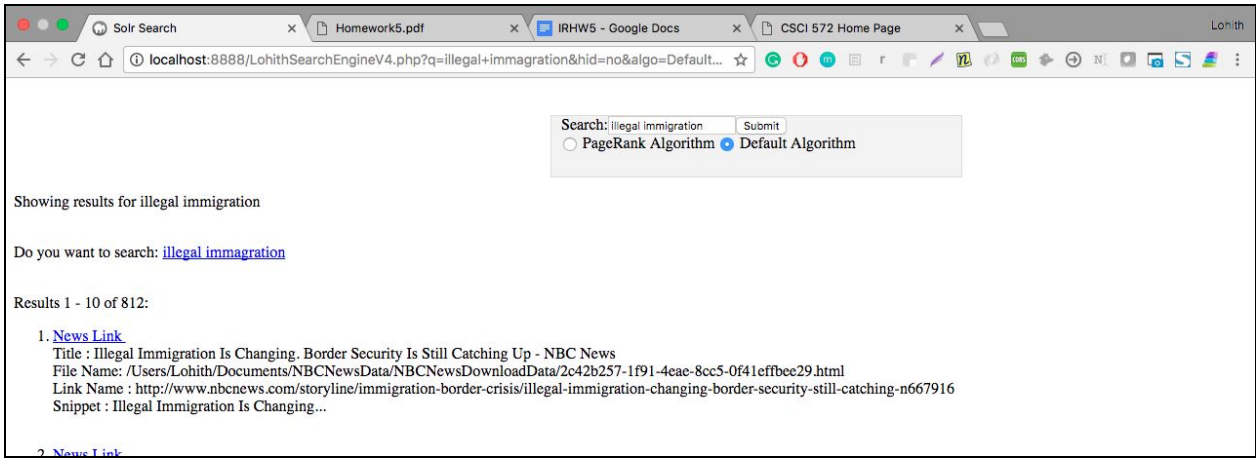

Solr Search
Homework5.pdf
IRHW5 - Google Docs
Lohith
localhost:8888/LohithSearchEngineV4.php
Search: donal
Submit
Default Algorithm
donal
donald
donate
donations
donating

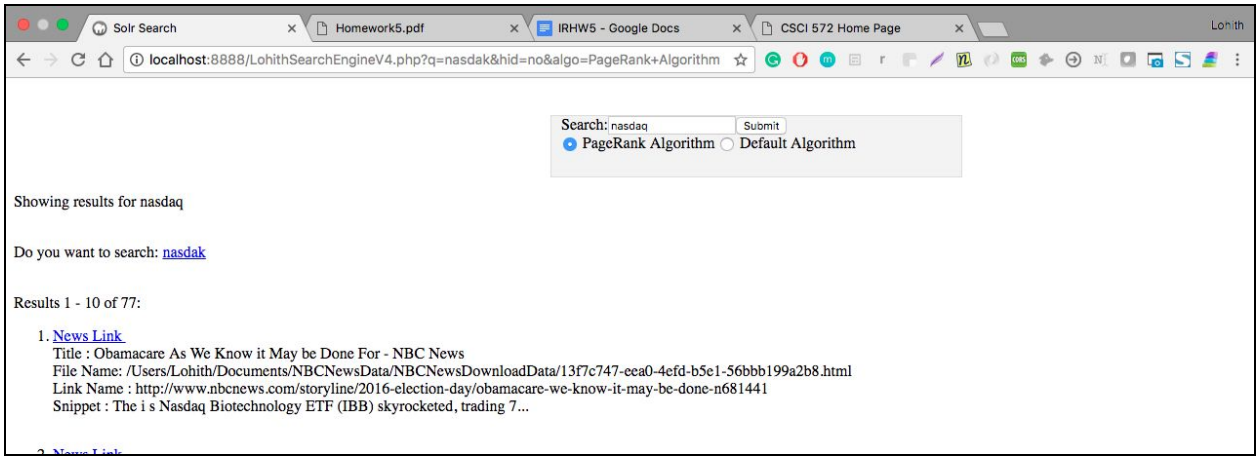

Solr Search
Homework5.pdf
IRHW5 - Google Docs
CSCI 572 Home Page
Lohith
localhost:8888/LohithSearchEngineV4.php
Search: illegal immi
Submit
illegal immi
illegal immigration
illegal immediately
illegal immigrants
illegal immigrant


Solr Search
Homework5.pdf
IRHW5 - Google Docs
CSCI 572 Home Page
Lohith
localhost:8888/LohithSearchEngineV4.php
Search: nassa
Submit
Default Algorithm
nasa
nassar
nasdaq
nissan

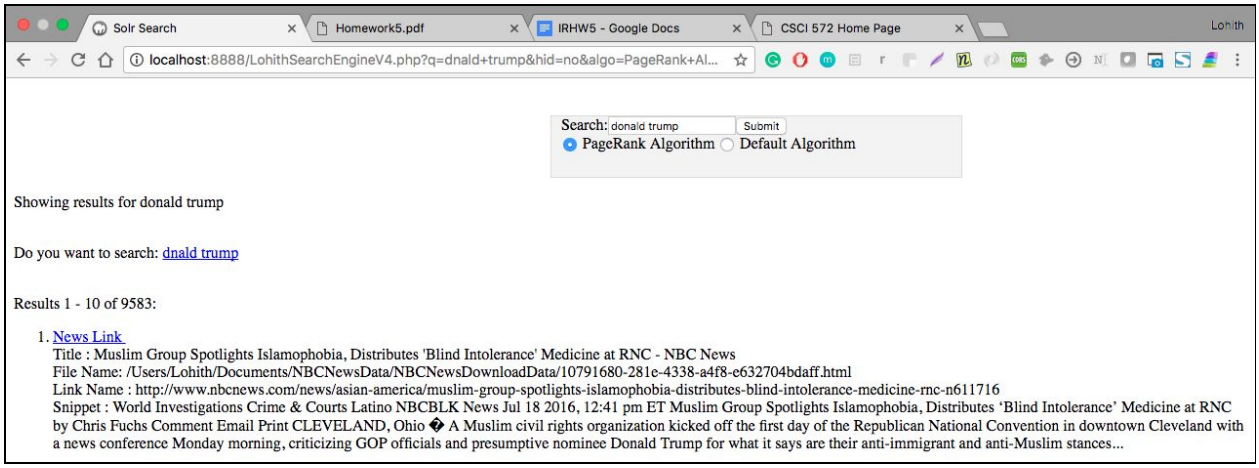
2. 5 sample examples' Screenshots for Spell Correction








Solr Search
Homework5.pdf
IRHW5 - Google Docs
CSCI 572 Home Page
Lohith
localhost:8888/LohithSearchEngineV4.php?q=naca&hid=no&algo=PageRank+Algorithm
Search: nasa
Submit
PageRank Algorithm
Default Algorithm
Showing results for nasa
Do you want to search: naca
Results 1 - 10 of 536:
1. News Link
Title : NASA's First Tragedy: 50 Years Since Apollo 1 Fire - NBC News
File Name: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/10e138a8-0a1a-4a6e-9972-8fac90ffe397.html
Link Name : http://www.nbcnews.com/slideshow/nasa-s-first-tragedy-50-years-apollo-1-fire-n713416
Snippet : NASA's First Tragedy: 50 Years Since Apollo 1 Fire - NBC News NBC News Mach Space Mach gallery Jan 27 2017, 5:45 pm ET NASA's First Tragedy: 50 Years Since Apollo 1 Fire on on Twitter on Google+ via Email On Jan...
2. News Link
Title : Bob Ebeling, Engineer Who Predicted Space Shuttle Challenger Explosion, Dies - NBC News


Solr Search
Homework5.pdf
IRHW5 - Google Docs
CSCI 572 Home Page
Lohith
localhost:8888/LohithSearchEngineV4.php?q=dnald+trump&hid=no&algo=PageRank+Al...
Search: donald trump
Submit
PageRank Algorithm
Default Algorithm
Showing results for donald trump
Do you want to search: dnald trump
Results 1 - 10 of 9583:
1. News Link
Title : Muslim Group Spotlights Islamophobia, Distributes 'Blind Intolerance' Medicine at RNC - NBC News
File Name: /Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/10791680-281e-4338-a4f8-e632704bdaff.html
Link Name : http://www.nbcnews.com/news/asian-america/muslim-group-spotlights-islamophobia-distributes-blind-intolerance-medicine-rnc-n611716
Snippet : World Investigations Crime & Courts Latino NBCBLK News Jul 18 2016, 12:41 pm ET Muslim Group Spotlights Islamophobia, Distributes 'Blind Intolerance' Medicine at RNC by Chris Fuchs Comment Email Print CLEVELAND, Ohio A Muslim civil rights organization kicked off the first day of the Republican National Convention in downtown Cleveland with a news conference Monday morning, criticizing GOP officials and presumptive nominee Donald Trump for what it says are their anti-immigrant and anti-Muslim stances...