

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 **htmlContentExtractor.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>

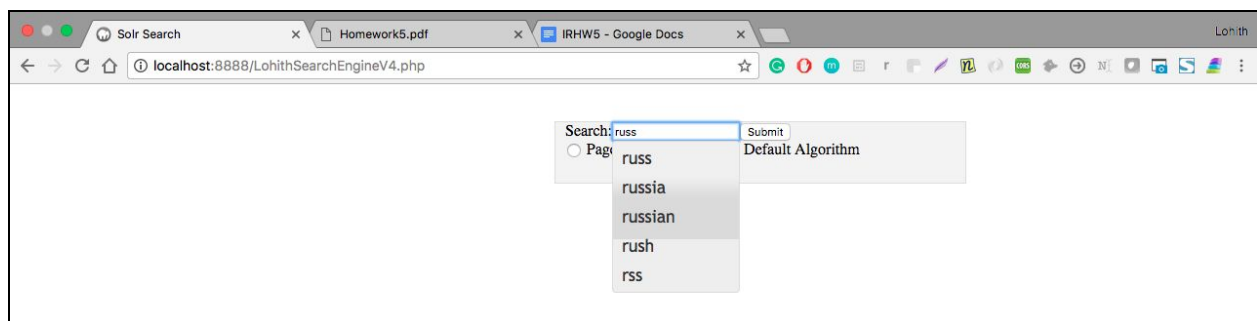
<arr name="components">
  <str>suggest</str>
</arr>
</requestHandler>

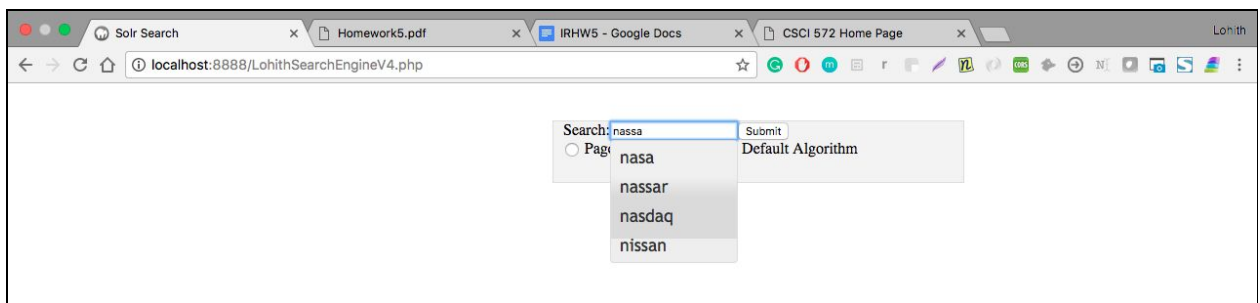
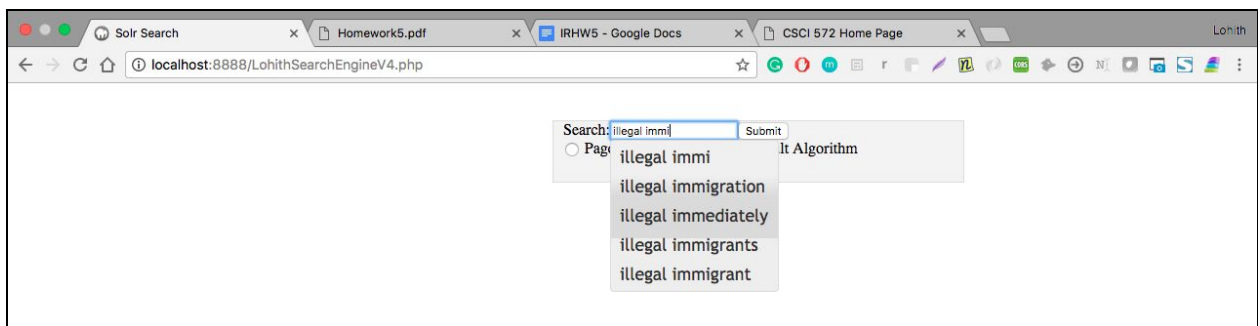
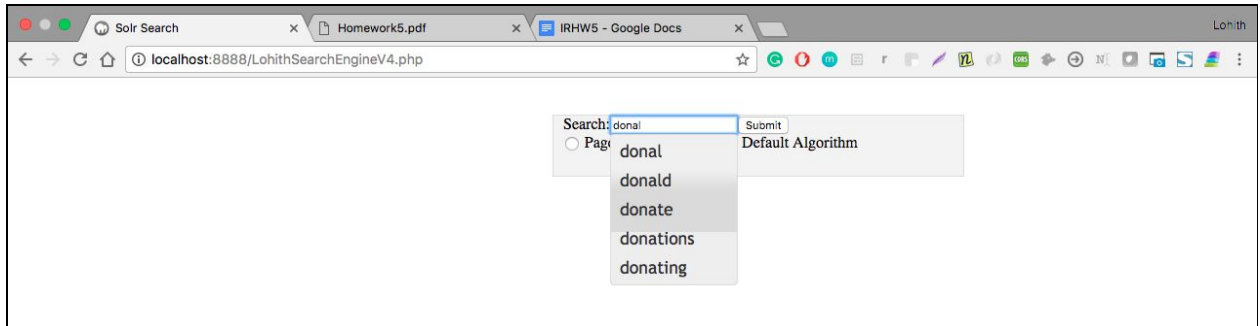
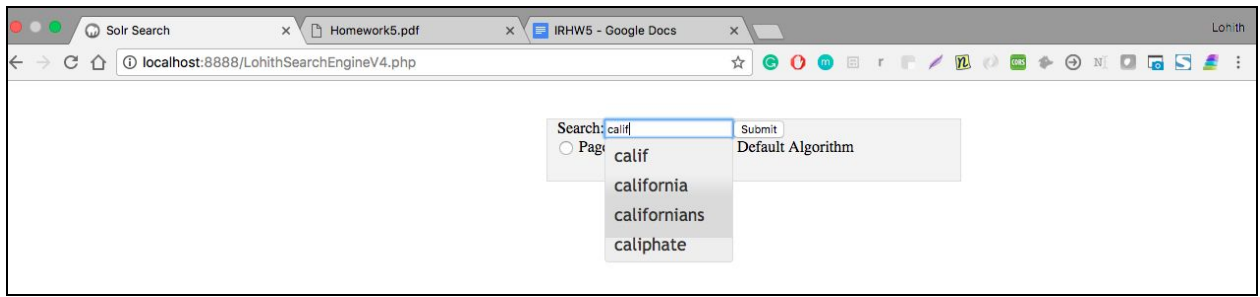
```

- 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.
- 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





2. 5 sample examples' Screenshots for Spell Correction

The screenshot shows a web browser with multiple tabs. The active tab is titled 'Solr Search'. The address bar shows the URL 'localhost:8888/LohithSearchEngineV4.php?q=snpcht&hid=no&algo=Default+Algorithm'. The search bar contains the text 'Search: snapchat' and a 'Submit' button. Below the search bar, there are two radio buttons: 'PageRank Algorithm' and 'Default Algorithm', with 'Default Algorithm' selected. The main content area displays 'Showing results for snapchat'. Below this, it says 'Do you want to search: [snpcht](#)'. The results section shows 'Results 1 - 10 of 233:'. The first result is a 'News Link' with the title 'Colleges Use Snapchat to Attract Prospective Students - NBC News'. The file name is '/Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/d7f88191-453c-4cc1-8cad-25c9f2541bd0.html'. The link name is 'http://www.nbcnews.com/feature/college-game-plan/colleges-use-snapchat-attract-prospective-students-n570701'. The snippet is 'Colleges Use Snapchat to Attract Prospective Students - NBC News NBC News feature College Game Plan College Game Plan May 10 2016, 11:37 am ET Colleges Use Snapchat to Attract Prospective Students by Scott Stump on on Twitter on Google+ via Email When Alex Cosentino visited the campus of the University of South Carolina earlier this year, the 18-year-old looked beyond the brochures and tour guides for a glimpse of what the school is really like...'. The second result is also a 'News Link'.

The screenshot shows a web browser with multiple tabs. The active tab is titled 'Solr Search'. The address bar shows the URL 'localhost:8888/LohithSearchEngineV4.php?q=illegal+immigration&hid=no&algo=Default...'. The search bar contains the text 'Search: illegal immigration' and a 'Submit' button. Below the search bar, there are two radio buttons: 'PageRank Algorithm' and 'Default Algorithm', with 'Default Algorithm' selected. The main content area displays 'Showing results for illegal immigration'. Below this, it says 'Do you want to search: [illegal immigration](#)'. The results section shows 'Results 1 - 10 of 812:'. The first result is a 'News Link' with the title 'Illegal Immigration Is Changing. Border Security Is Still Catching Up - NBC News'. The file name is '/Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/2c42b257-1f91-4eac-8cc5-0f41effbee29.html'. The link name is 'http://www.nbcnews.com/storyline/immigration-border-crisis/illegal-immigration-changing-border-security-still-catching-n667916'. The snippet is 'Illegal Immigration Is Changing...'. The second result is also a 'News Link'.

The screenshot shows a web browser with multiple tabs. The active tab is titled 'Solr Search'. The address bar shows the URL 'localhost:8888/LohithSearchEngineV4.php?q=nasdaq&hid=no&algo=PageRank+Algorithm'. The search bar contains the text 'Search: nasdaq' and a 'Submit' button. Below the search bar, there are two radio buttons: 'PageRank Algorithm' and 'Default Algorithm', with 'PageRank Algorithm' selected. The main content area displays 'Showing results for nasdaq'. Below this, it says 'Do you want to search: [nasdak](#)'. The results section shows 'Results 1 - 10 of 77:'. The first result is a 'News Link' with the title 'ObamaCare As We Know it May be Done For - NBC News'. The file name is '/Users/Lohith/Documents/NBCNewsData/NBCNewsDownloadData/13f7c747-eca0-4efd-b5e1-56bbb199a2b8.html'. The link name is 'http://www.nbcnews.com/storyline/2016-election-day/obamacare-we-know-it-may-be-done-n681441'. The snippet is 'The i s Nasdaq Biotechnology ETF (IBB) skyrocketed, trading 7...'. The second result is also a 'News Link'.

Solr Search x Homework5.pdf x IRHW5 - Google Docs x CSCI 572 Home Page x Lohith

localhost:8888/LohithSearchEngineV4.php?q=nasa&hid=no&algo=PageRank+Algorithm

Search: nasa Submit
☒ PageRank Algorithm ☐ Default Algorithm

Showing results for nasa

Do you want to search: [nasa](#)

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 x Homework5.pdf x IRHW5 - Google Docs x CSCI 572 Home Page x Lohith

localhost:8888/LohithSearchEngineV4.php?q=dnald+trump&hid=no&algo=PageRank+Al...

Search: dnald trump Submit
☒ PageRank Algorithm ☐ Default Algorithm

Showing results for donald trump

Do you want to search: [dnald trump](#)

Results 1 - 10 of 9582:

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-mc-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...