Steps followed to complete the assignment (Summary):-

The Solr program is enhanced with autocomplete and spell checking functionality. An external spelling correction program (Norvig's spellchecker algorithm) in conjunction with Solr's autocomplete functionality is used.

Spell Checking:-

To accomplish the task of spell checking, Norwig's spell checking algorithm has been used. The implementation of this algorithm in PHP was downloaded from: http://www.phpclasses.org/package/4859-PHP-Suggest-corrected-spelling-text-in-pure-PHP.html#download.

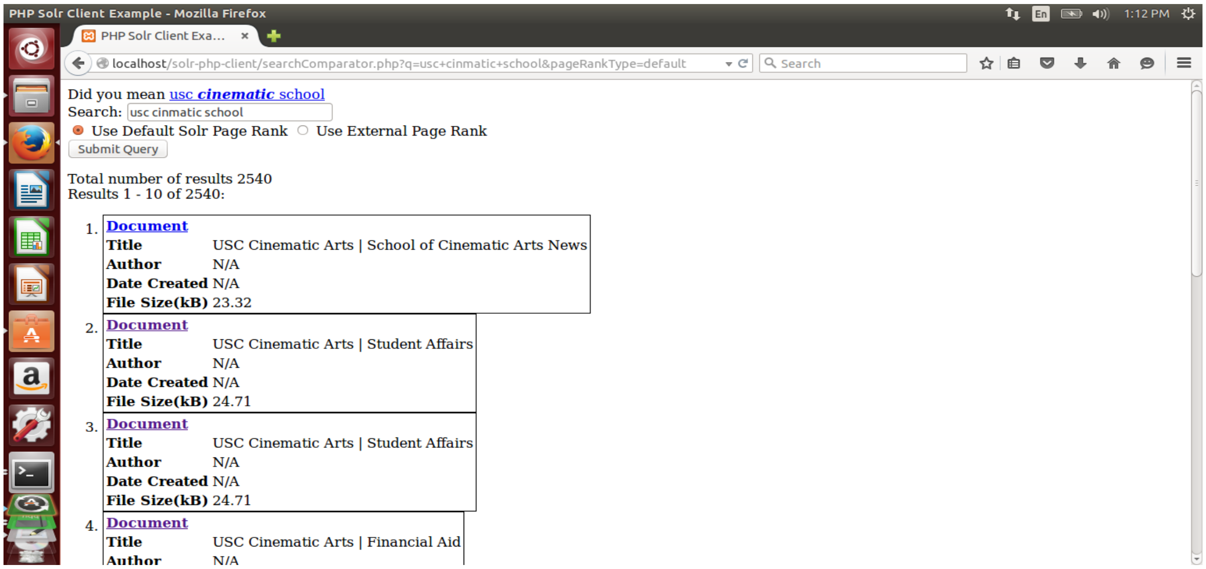
Building the corpus:

The big.txt file, which the program needs is built from the saved files, which were indexed in the previous exercise. A python program was written to build the big.txt file. In this program, corpusBuilder.py, all the saved files are read and html tags and other irrelevant information are removed to obtain the corpus (big.txt).

Relevance feedback of the misspelt words (Google style):

All the words in the query are checked for spelling errors using the above algorithm and the misspelt words (bold and italicized) along with the words which are spelt right are provided to the user as part of relevance feedback. The results are displayed by querying with the user's unaltered query and a relevance feedback of "Did you mean <*altered query*>" is presented to the user as a clickable link. When this link is clicked by the user, the results related the altered query are displayed to the user. This feature is case insensitive.

The following snapshots show the spell check feature for single misspelt word, single misspelt word among multiple words and multiple misspelt words.



localhost/solr-php-client/searchComparator.php?q=usc+cinmatic+school&pageRankType=default

Did you mean usc ***cinematic*** school

Search: usc cinmatic school

● Use Default Solr Page Rank ○ Use External Page Rank

Submit Query

Total number of results 2540
Results 1 - 10 of 2540:

1.

Document	
Title	USC Cinematic Arts | School of Cinematic Arts News
Author	N/A
Date Created	N/A
File Size(kB)	23.32

2.

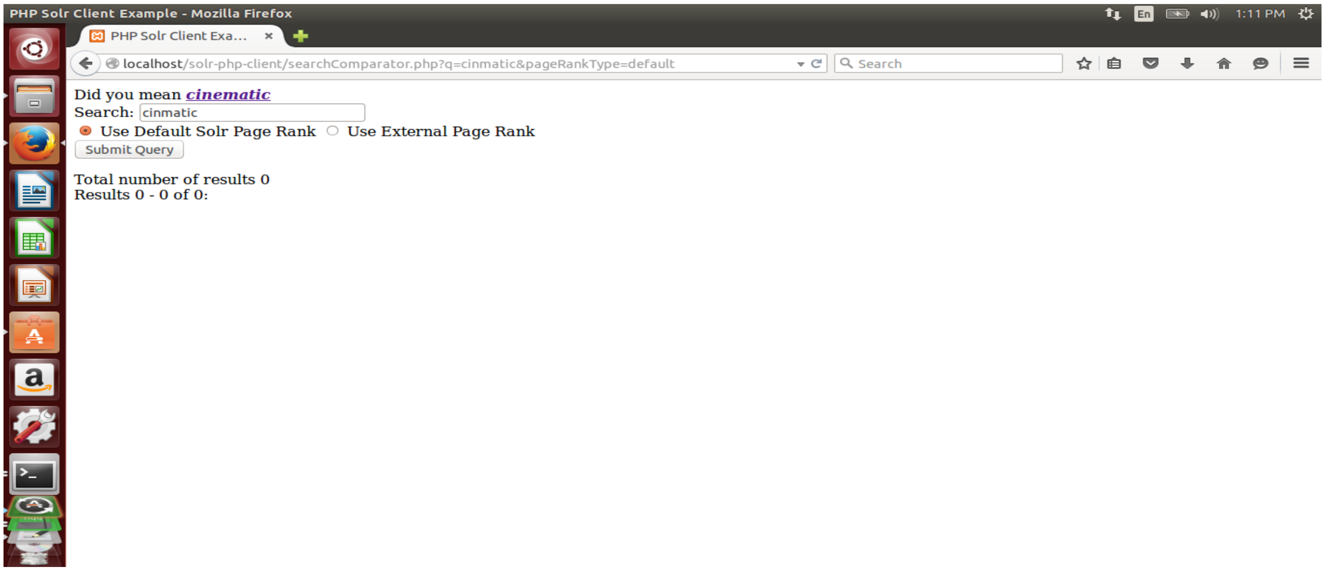
Document	
Title	USC Cinematic Arts | Student Affairs
Author	N/A
Date Created	N/A
File Size(kB)	24.71

3.

Document	
Title	USC Cinematic Arts | Student Affairs
Author	N/A
Date Created	N/A
File Size(kB)	24.71

4.

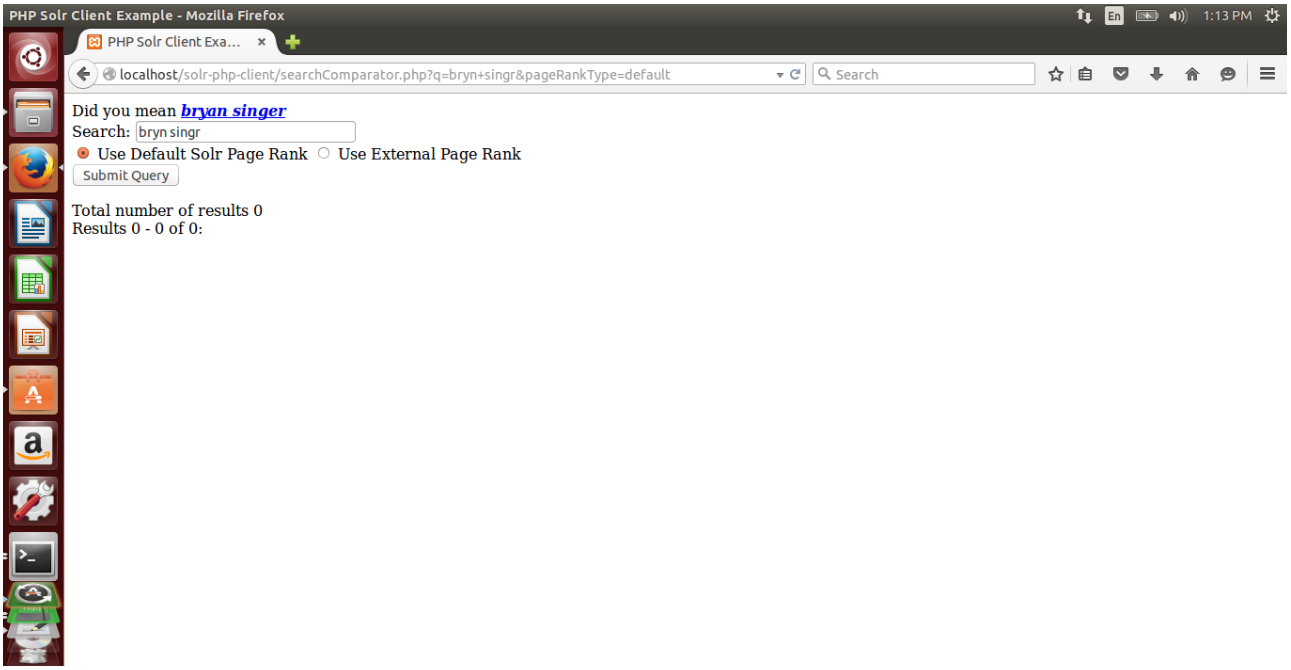
Document	
Title	USC Cinematic Arts | Financial Aid
Author	N/A



localhost/solr-php-client/searchComparator.php?q=cinmatic&pageRankType=default

Did you mean *cinematic*
Search: cinmatic
● Use Default Solr Page Rank ○ Use External Page Rank
Submit Query

Total number of results 0
Results 0 - 0 of 0:



localhost/solr-php-client/searchComparator.php?q=bryn+singr&pageRankType=default

Did you mean *bryan singer*
Search: bryn singr
● Use Default Solr Page Rank ○ Use External Page Rank
Submit Query

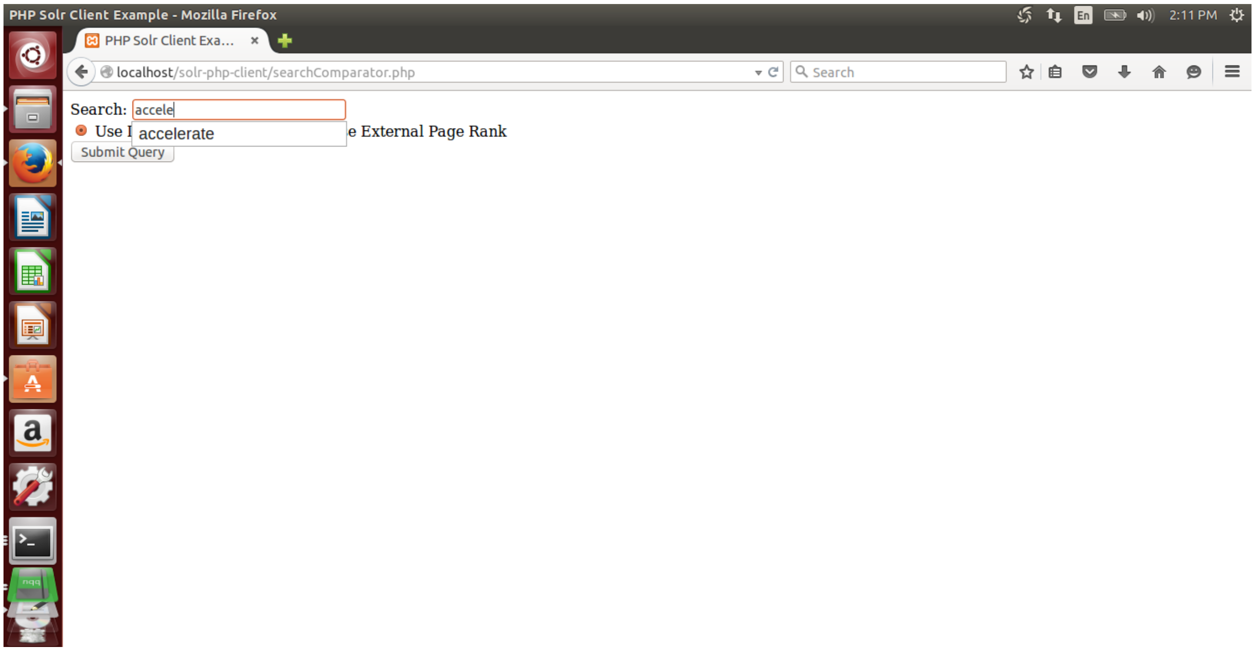
Total number of results 0
Results 0 - 0 of 0:

Auto Suggest feature:

As mentioned earlier, autocomplete was accomplished using Solr's functionality of FuzzyLookupFactory. Every character input by the user in the text box is retrieved with the help of jQuery and an Ajax call is made to the Solr suggest url. The results retrieved from this call is used to feed as suggestions to the user with the help of jQuery autocomplete. The autocomplete feature has been provided for characters, single words and multiple words. The autocomplete feature is case-sensitive. Stemmer and Efficient Stop word removal for effective auto suggestion is also incorporated.

Stemming:

The auto suggest feature is designed to not suggest words that come from the same stem using Porter's stemming algorithm. If the user's intent is to search for "accelerate" and his input is "accele", then "accelerate" is only shown as a suggestion instead of showing "accelerated", "accelerate" and "accelerating" since all of the them have the same stem. The snapshot portrays this.



localhost/solr-php-client/searchComparator.php

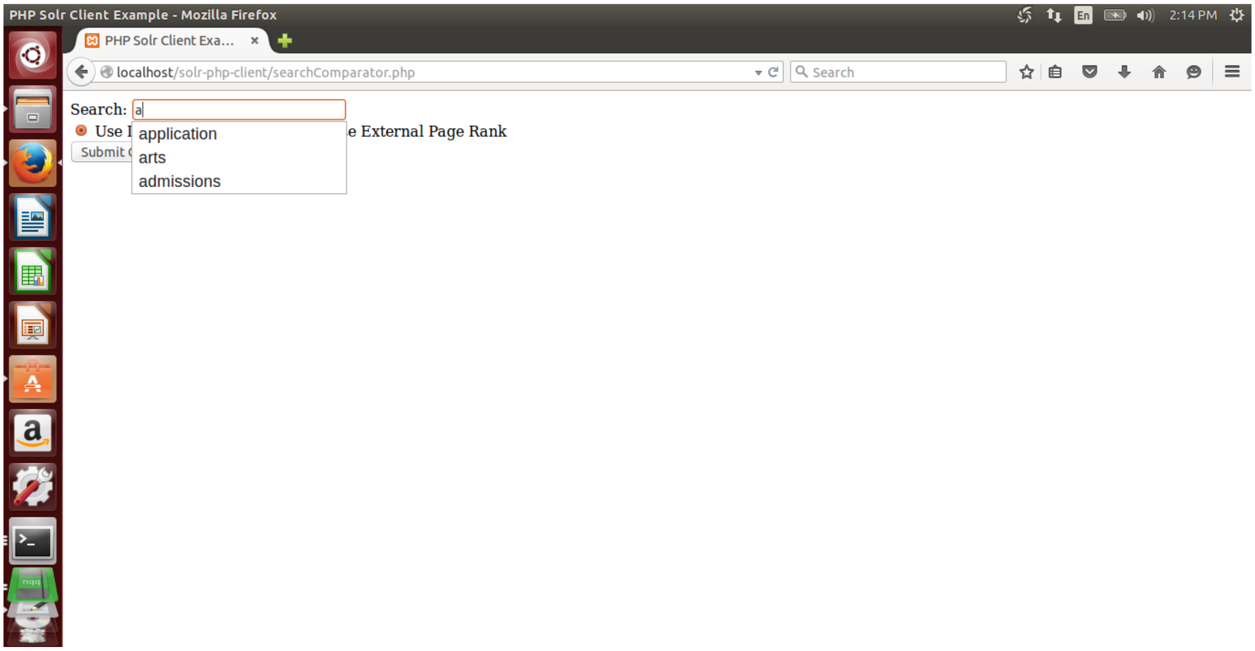
Search: accele

accelerate

● Use I e External Page Rank

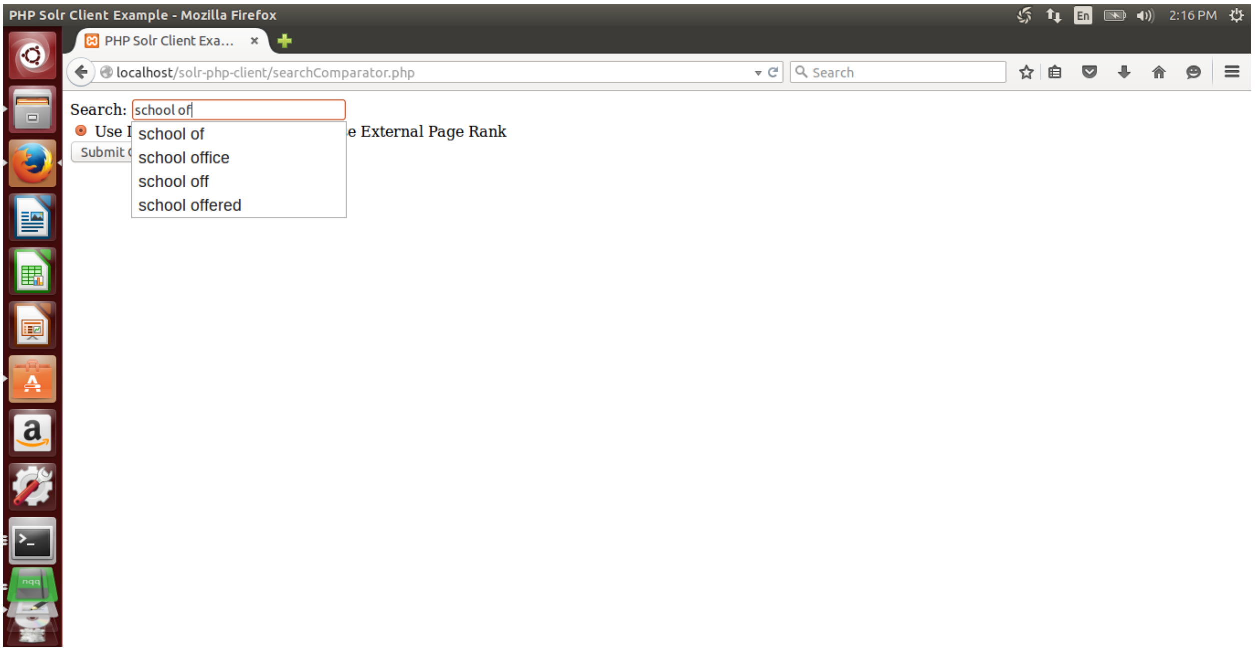
Efficient Stop Words Removal:

Stop words are not provided as suggestions to the user if they match with the user's query and appear in the beginning of the query with higher preference than any other suggestions. An example is that, if the user's intent is to search for "application" and the user has entered "a" in the text box, then the stop word "a" would be removed from the list of suggestions provided to the user. Depicted in the snapshot.


PHP Solr Client Example - Mozilla Firefox
PHP Solr Client Exa...
localhost/solr-php-client/searchComparator.php
Search
2:14 PM
Use I
e External Page Rank
Submit

Search: a

However, the stop words are not removed from the suggestions, if they appear as a one of the word in the query other than the first word and the user's intent is to search for a query along with the stop word. An example would be the suggestions provided to the query "school of" as shown in the snapshot.


PHP Solr Client Example - Mozilla Firefox
PHP Solr Client Exa...
localhost/solr-php-client/searchComparator.php
Search
2:16 PM
Use I
e External Page Rank
Submit

Search: school of