REPORT

CSCI 572 – Information Retrieval and Web Search Engines – Spring 2017

Homework 5 – Addition of Spell Checking, Autocomplete and Snippets to the Search Engine

Steps Followed to Complete Home Work Assignment #5

1. Initial Setup

- Changes were incorporated into the solrconfig.xml file to accommodate the suggest/autocomplete features as per the specifications mentioned in the assignment description.
- The searchComponent was added along with the FuzzyLookUpFactory suggestComponent.
- This was followed by the addition of a request handler to configure the parameters for the suggest component. The count of suggestions was given as 5 indicating that 5 suggestions are shown at a time.
- The core solr/myexample was reloaded for the changes to take effect.

2. Spell Checker

- Peter Norvig's SpellCorrector program was used to implement the spell check functionality.
- The text files from the downloaded web pages of nbcnews.com were parsed using Java's JSOUP library. Eliminated the punctuations and tokenized the contents to get all the words in the data.
- Generated a serialized dictionary (serialized_dictionary.txt) containing the set of words and their combined frequency of occurrence in the entire dataset.
- The php version of *Peter Norvig's* spell corrector code was used to correct the word spellings based on the edit distance of the query word and words in the serialized dictionary. Valid spell correct suggestions are computed up to an edit distance of 2 from the query words.
- The user query is tokenized and sent iteratively to the spell check script that matches the query entered to the nearest valid word with a maximum edit distance of 2 and returns it.
- If the user query words were found to be valid, the user is shown search results for the entered query. Otherwise, the spell corrector suggests a possible correction from the dictionary and the user can view the results for the corrected query by clicking on it in a similar fashion to most modern search engines (like Google).
- The accuracy of the spell correction depends on the Norvig's Spell corrector and the dictionary collected from the NBC news data. It works quite well for the given data but can always be improved by adding more data to generate a more comprehensive dictionary to enhance the accuracy.

3. Auto Complete

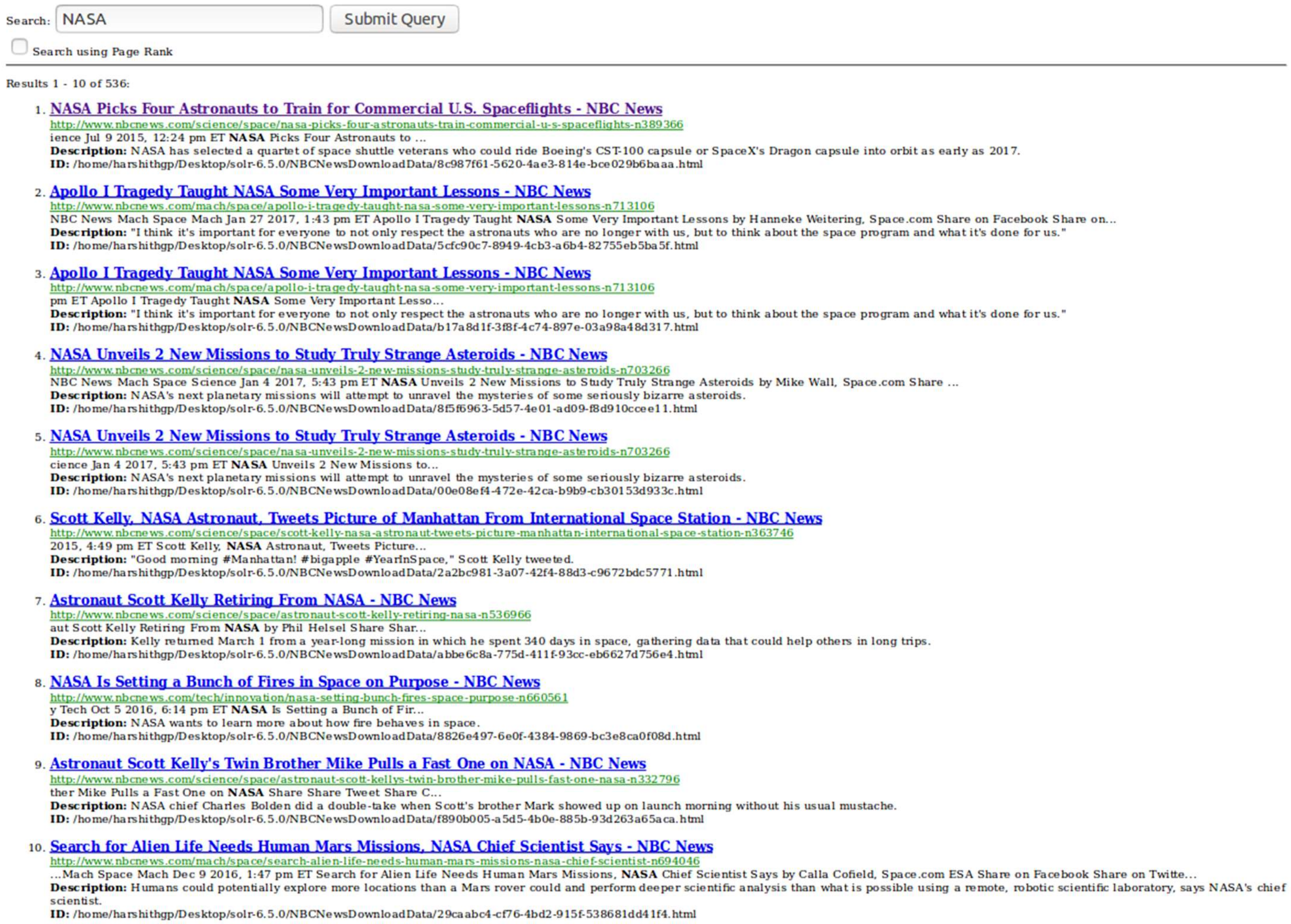
- Implemented the autocomplete functionality by enabling the Fuzzy Lookup Factory feature of the Solr Lucene component which creates suggestions for misspelled words in fields. The suggest component was enabled in Solr as specified in the tutorial.
- Used the AJAX and jQuery to implement live auto completion and spell suggestions and performed a pattern match for various stopwords.
- Used the jQuery API to get the suggestions. The API makes an AJAX call to Solr to retrieve the suggestions. The AJAX call is triggered each time the user types a letter in the search box. This makes a call to http://localhost:8983/solr/myexample/suggest

- The data obtained from the solr server is in JSON format which includes the weight of each suggestion. The terms were extracted from the JSON and pushed into an array, which is used to display up to 5 suggestions for the given input.

4. Snippet Generation

- Snippets were generated using a regex pattern which breaks the html content from the url obtained for each result into sentences which are matched against the query terms.
- Generated a snippet folder with .txt files containing text from the body of given NBC news articles using a JAVA program. The Snippets are generated from the previously generated text files.
- A search performed on each sentence of the snippet text returns the best match for the query terms. The sentence that contains all or most of the query terms is given priority over a sentence that contains just one of the query terms.
- The matched query words are highlighted in the snippet as done on Google.

Results with Snippets for NASA



Search: NASA

□ Search using Page Rank

Results 1 - 10 of 536:

1. NASA Picks Four Astronauts to Train for Commercial U.S. Spaceflights - NBC News
http://www.nbcnews.com/science/space/nasa-picks-four-astronauts-train-commercial-u-s-spaceflights-n389366
ience Jul 9 2015, 12:24 pm ET **NASA** Picks Four Astronauts to ...
Description: NASA has selected a quartet of space shuttle veterans who could ride Boeing's CST-100 capsule or SpaceX's Dragon capsule into orbit as early as 2017.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/8c987f61-5620-4ae3-814e-bce029b6baaa.html

2. Apollo I Tragedy Taught NASA Some Very Important Lessons - NBC News
http://www.nbcnews.com/mach/space/apollo-i-tragedy-taught-nasa-some-very-important-lessons-n713106
NBC News Mach Space Mach Jan 27 2017, 1:43 pm ET Apollo I Tragedy Taught **NASA** Some Very Important Lessons by Hanneke Weitering, Space.com Share on Facebook Share on...
Description: "I think it's important for everyone to not only respect the astronauts who are no longer with us, but to think about the space program and what it's done for us."
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/5cfc90c7-8949-4cb3-a6b4-82755eb5ba5f.html

3. Apollo I Tragedy Taught NASA Some Very Important Lessons - NBC News
http://www.nbcnews.com/mach/space/apollo-i-tragedy-taught-nasa-some-very-important-lessons-n713106
pm ET Apollo I Tragedy Taught **NASA** Some Very Important Lesso...
Description: "I think it's important for everyone to not only respect the astronauts who are no longer with us, but to think about the space program and what it's done for us."
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/b17a8d1f-3f8f-4c74-897e-03a98a48d317.html

4. NASA Unveils 2 New Missions to Study Truly Strange Asteroids - NBC News
http://www.nbcnews.com/science/space/nasa-unveils-2-new-missions-study-truly-strange-asteroids-n703266
NBC News Mach Space Science Jan 4 2017, 5:43 pm ET **NASA** Unveils 2 New Missions to Study Truly Strange Asteroids by Mike Wall, Space.com Share ...
Description: NASA's next planetary missions will attempt to unravel the mysteries of some seriously bizarre asteroids.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/8f5f6963-5d57-4e01-ad09-f8d910ccee11.html

5. NASA Unveils 2 New Missions to Study Truly Strange Asteroids - NBC News
http://www.nbcnews.com/science/space/nasa-unveils-2-new-missions-study-truly-strange-asteroids-n703266
cience Jan 4 2017, 5:43 pm ET **NASA** Unveils 2 New Missions to...
Description: NASA's next planetary missions will attempt to unravel the mysteries of some seriously bizarre asteroids.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/00e08ef4-472e-42ca-b9b9-cb30153d933c.html

6. Scott Kelly, NASA Astronaut, Tweets Picture of Manhattan From International Space Station - NBC News
http://www.nbcnews.com/science/space/scott-kelly-nasa-astronaut-tweets-picture-manhattan-international-space-station-n363746
2015, 4:49 pm ET Scott Kelly, **NASA** Astronaut, Tweets Picture...
Description: "Good morning #Manhattan! #bigapple #YearInSpace," Scott Kelly tweeted.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/2a2bc981-3a07-42f4-88d3-c9672bdc5771.html

7. Astronaut Scott Kelly Retiring From NASA - NBC News
http://www.nbcnews.com/science/space/astronaut-scott-kelly-retiring-nasa-n536966
aut Scott Kelly Retiring From **NASA** by Phil Helsel Share Shar...
Description: Kelly returned March 1 from a year-long mission in which he spent 340 days in space, gathering data that could help others in long trips.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/abbe6c8a-775d-411f-93cc-eb6627d756e4.html

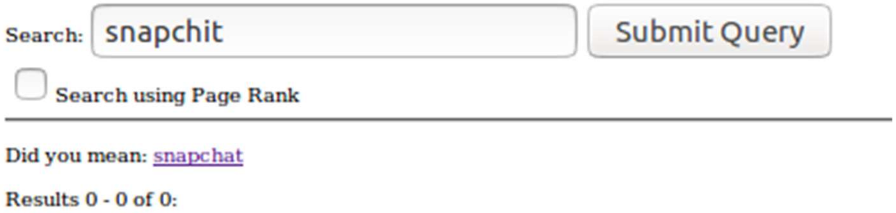
8. NASA Is Setting a Bunch of Fires in Space on Purpose - NBC News
http://www.nbcnews.com/tech/innovation/nasa-setting-bunch-fires-space-purpose-n660561
y Tech Oct 5 2016, 6:14 pm ET **NASA** Is Setting a Bunch of Fir...
Description: NASA wants to learn more about how fire behaves in space.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/8826e497-6e0f-4384-9869-bc3e8ca0f08d.html

9. Astronaut Scott Kelly's Twin Brother Mike Pulls a Fast One on NASA - NBC News
http://www.nbcnews.com/science/space/astronaut-scott-kellys-twin-brother-mike-pulls-fast-one-nasa-n332796
ther Mike Pulls a Fast One on **NASA** Share Share Tweet Share C...
Description: NASA chief Charles Bolden did a double-take when Scott's brother Mark showed up on launch morning without his usual mustache.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/f890b005-a5d5-4b0e-885b-93d263a65aca.html

10. Search for Alien Life Needs Human Mars Missions, NASA Chief Scientist Says - NBC News
http://www.nbcnews.com/mach/space/search-alien-life-needs-human-mars-missions-nasa-chief-scientist-n694046
...Mach Space Mach Dec 9 2016, 1:47 pm ET Search for Alien Life Needs Human Mars Missions, **NASA** Chief Scientist Says by Calla Cofield, Space.com ESA Share on Facebook Share on Twitte...
Description: Humans could potentially explore more locations than a Mars rover could and perform deeper scientific analysis than what is possible using a remote, robotic scientific laboratory, says NASA's chief scientist.
ID: /home/harshithgp/Desktop/solr-6.5.0/NBCNewsDownloadData/29caabc4-cf76-4bd2-915f-538681dd41f4.html

Screenshots – Autocorrect / Spell Check

1. **snapchit to snapchat**



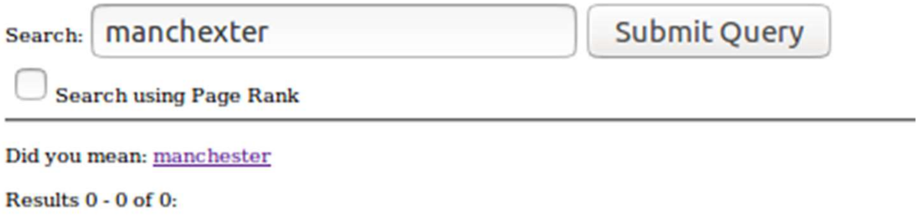
Search: snapchit

□ Search using Page Rank

Did you mean: snapchat

Results 0 - 0 of 0:

2. **manchexter to manchester**



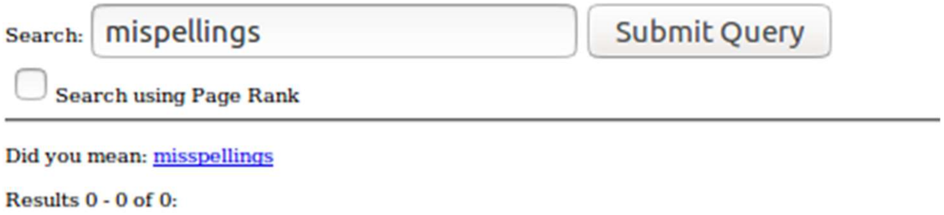
Search: manchexter

□ Search using Page Rank

Did you mean: manchester

Results 0 - 0 of 0:

3. **mispellings to misspellings**



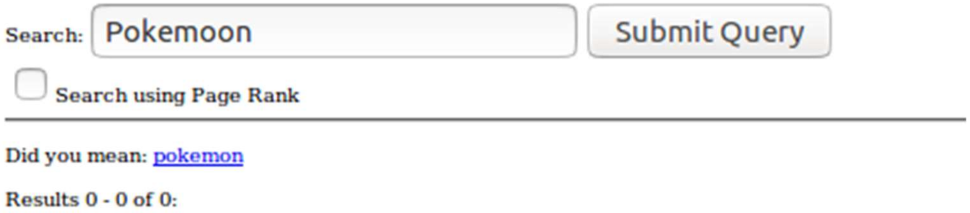
Search: mispellings

□ Search using Page Rank

Did you mean: misspellings

Results 0 - 0 of 0:

4. **Pokemoon to pokemon**



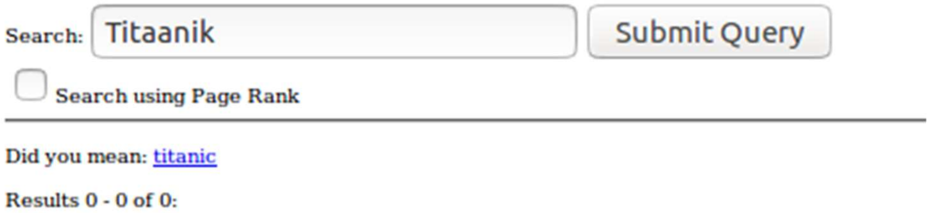
Search: Pokemoon

□ Search using Page Rank

Did you mean: pokemon

Results 0 - 0 of 0:

5. **Titaanik to titanic**



Search: Titaanik

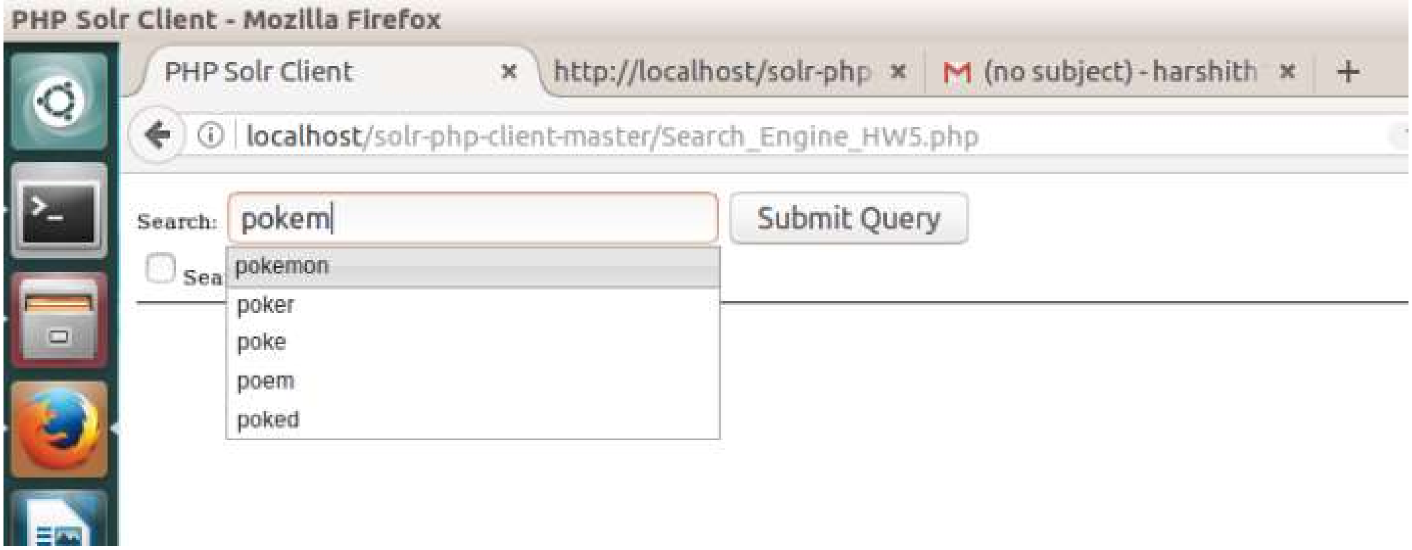
□ Search using Page Rank

Did you mean: titanic

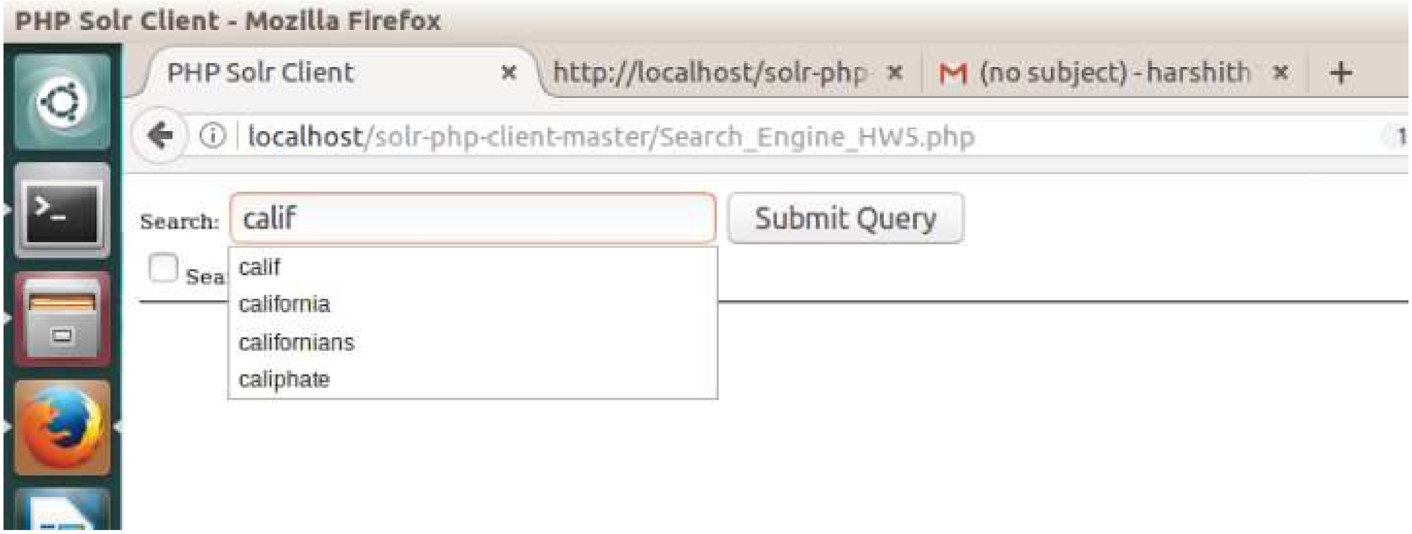
Results 0 - 0 of 0:

Screen Shots - Autocomplete / Suggestions

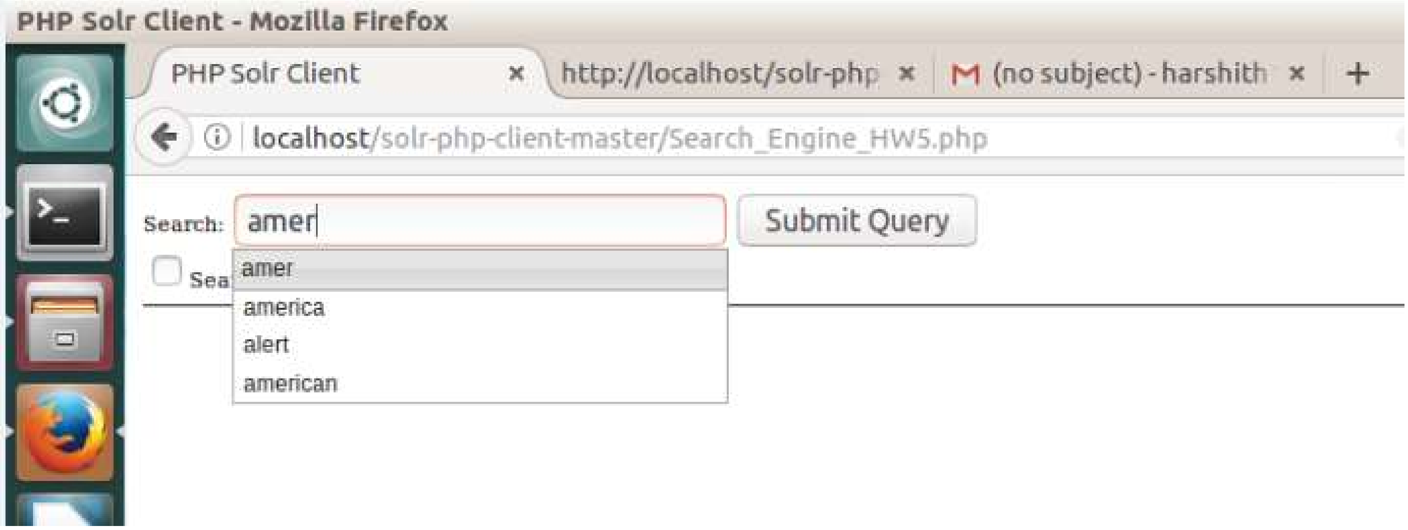
1. pokem


PHP Solr Client - Mozilla Firefox
PHP Solr Client
http://localhost/solr-php
(no subject) - harshith
localhost/solr-php-client-master/Search_Engine_HW5.php
Search:
pokem
Submit Query
Sea
pokemon
poker
poke
poem
poked

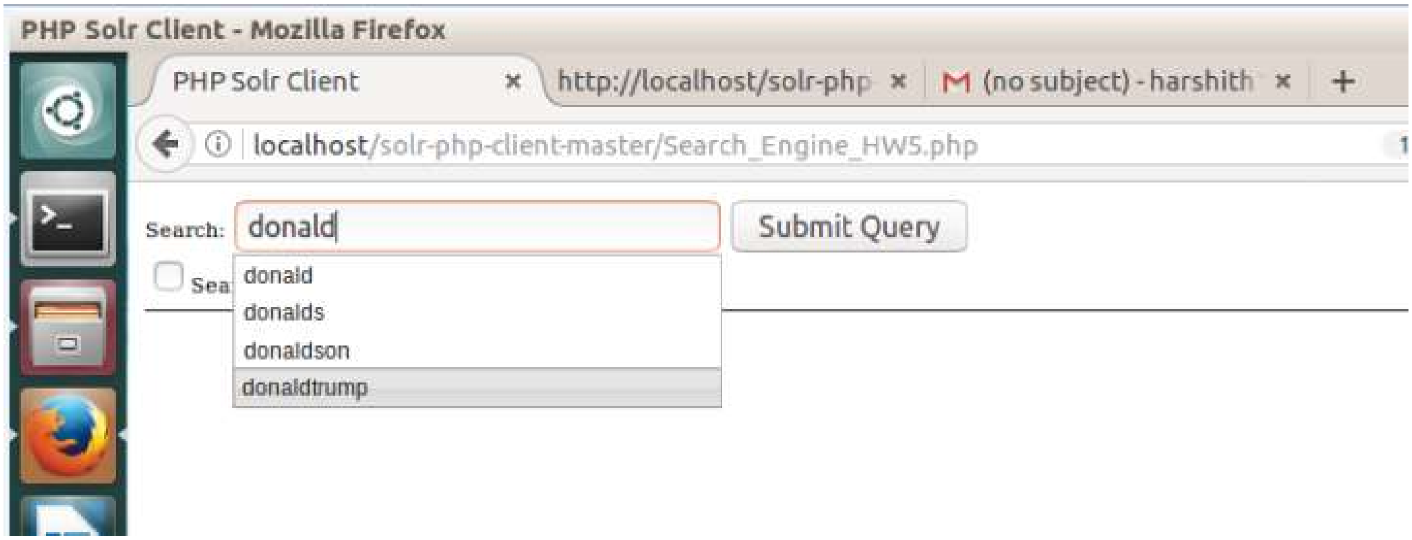
2. calif


PHP Solr Client - Mozilla Firefox
PHP Solr Client
http://localhost/solr-php
(no subject) - harshith
localhost/solr-php-client-master/Search_Engine_HW5.php
Search:
calif
Submit Query
Sea
calif
california
californians
caliphate

3. amer

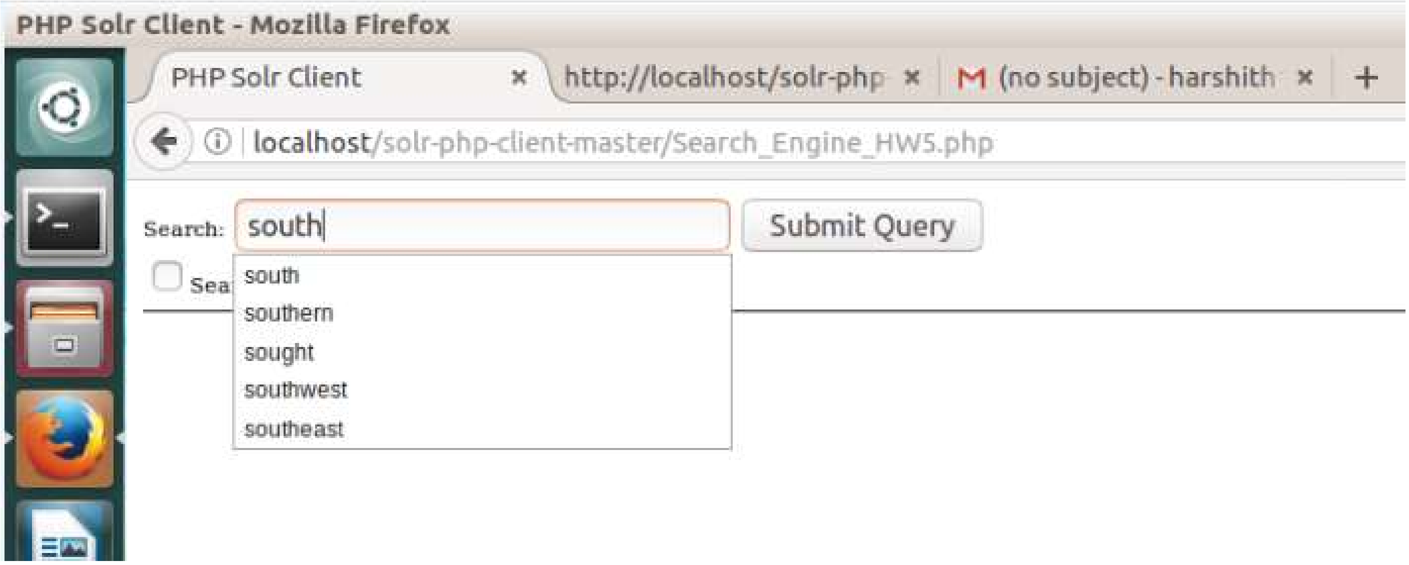

PHP Solr Client - Mozilla Firefox
PHP Solr Client
http://localhost/solr-php
(no subject) - harshith
localhost/solr-php-client-master/Search_Engine_HW5.php
Search:
amer
Submit Query
Sea
amer
america
alert
american

4. donald


PHP Solr Client - Mozilla Firefox
PHP Solr Client
http://localhost/solr-php
(no subject) - harshith
localhost/solr-php-client-master/Search_Engine_HW5.php
Submit Query
Sea
donald
donalds
donaldson
donaldtrump

Search: donald

5. south


PHP Solr Client - Mozilla Firefox
PHP Solr Client
http://localhost/solr-php
(no subject) - harshith
localhost/solr-php-client-master/Search_Engine_HW5.php
Submit Query
Sea
south
southern
sought
southwest
southeast

Search: south

Conclusion

- The autosuggest and spell correct features can be enhanced by using a more comprehensive dictionary of words from various sources.
- The front end of the search engine was designed keeping in mind Google's user interface.
- CSCI 572 – Information Retrieval and Web Search Engines Completed.