Terry Chiang
3357860428

CSCI572 Homework 5 Report

1. Steps you followed to complete this assignment. Include the details of what tools and techniques you used to implement spelling correction and autocomplete.

Spelling Correction

- Using the tika library by downloading the jar file from the website, I was able to parse all of the html files inside HtmlParse.java and create the big.txt file.

- Peter Norvig's SpellCorrector.php reads big.txt in order to create the serialized_dictionary. The serialized_dictionary is then called by SpellCorrector.php to use for spelling corrections.

- After the dictionary is created, I am able to use the spell corrector call inside my HW5.php file in order to make spelling corrections using include 'SpellCorrector.php' and SpellCorrector::correct($word). If a multi-word query is entered, I have to split the string into separate queries and do a spell check on each query.

Autocomplete

- First I added the 2 blocks of code to enable the use of solr's autocomplete in the solrconfig.xml file.

- Then I added jquery code within my php code to make ajax calls to solr's autocomplete feature. The ajax ink would be formed like so: http://localhost:8983/solr/myexample/suggest?q=word

- Jquery's autocomplete function creates a clean design for displaying autocompletes in a timely manner for the specified tag or class. Within the javascript code, I added an array to display multi-word autocompletes since solr only does single word queries.

- I then made ajax calls whenever the user typed anything in order to obtain the typed query and create suggestions for the autocomplete. The ajax call required extra parameters to allow for CORS on my local machine.

- The ajax call to the link for suggestions come back in json. I would parse out the json for the suggested words in order to create the suggestionList that I would end up displaying in the autocomplete box.

- The last portion of my javascript code included running the query search if the autocomplete suggestion was clicked on using $('#searchForm').submit();

Snippet

- For snippets, I used the single_html_dom_parser.php, found here http://simplehtmldom.sourceforge.net/, to parse the web pages returned by solr. Using the dom_parser I was able to look for paragraph tags with a line such as $html->find(“p”).

- I created a genSnippets function that handled the bulk of the snippets. In particular I searched for query words inside the paragraph tags of each link using regular expressions. The regular expression would return a match if the paragraph includes the query term(s). Within genSnippets I would add bold tags to the any words that matched the query words. Also genSnippets would utilize other helper functions, which I mention below, to create the proper snippets.

- Since some pages could have multiple paragraph tags, I would save the best snippet inside a string during each loop of the paragraph tags. At the end I would call compareReturn function, details below, to choose the best snippet.

- I created a trimSnippets function to trim snippets from the back of the snippet if the paragraph was greater than 160 characters. If a query term is found during the trimming of the back of the snippet, I would break out of the for loop and start trimming from the front of the sentence. I would then add “…” where needed if trimming occurred.

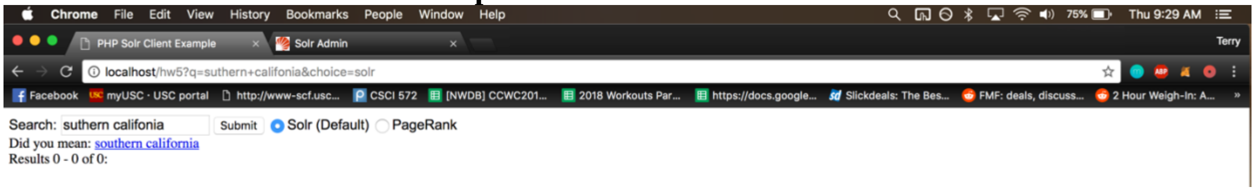
- I created a addSnippet function that would add more words to the snippet if the already generated snippet was under 160 characters.

- I also created a compareReturn function that would compare 2 different snippets to see which would be best to return. The best one would be the one that contained all the query terms.

- For my particular news site, Newsday, I would get some pages with 404 not found. In those cases it had to check if the page was retrieved or not. I displayed “Error: HTTP/1.1 404 Not Found” as the snippet.

2. Analysis of the results: In this you should provide FIVE examples of misspelled terms that are correctly handled by your spelling correction program. You should also provide FIVE examples of auto-completion.

Misspelled Terms

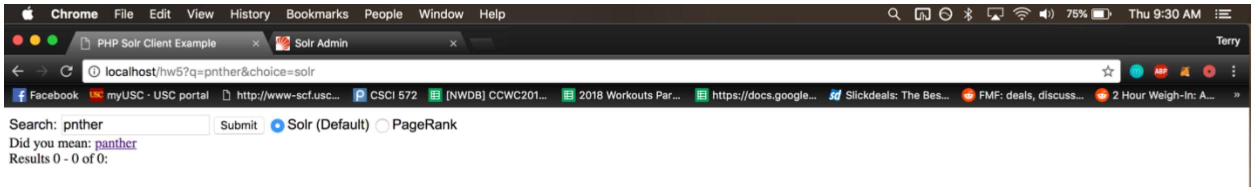
1. 
Chrome File Edit View History Bookmarks People Window Help
75% Thu 9:29 AM
PHP Solr Client Example
Solr Admin
Terry
Facebook myUSC · USC portal http://www-scf.usc... CSCI 572 [NWDB] CCWC201... 2018 Workouts Par... https://docs.google... Slickdeals: The Bes... FMF: deals, discuss... 2 Hour Weigh-In: A...

localhost/hw5?q=suthern+califonia&choice=solr

Search: suthern califonia Submit ● Solr (Default) ○ PageRank

Did you mean: southern california

Results 0 - 0 of 0:

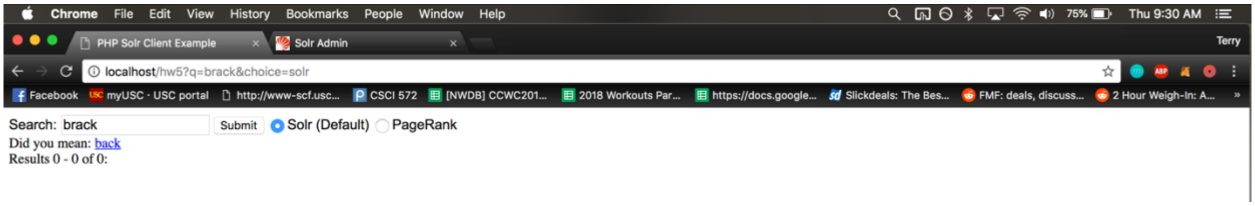
2. 
Chrome File Edit View History Bookmarks People Window Help
75% Thu 9:30 AM
PHP Solr Client Example
Solr Admin
Terry
Facebook myUSC · USC portal http://www-scf.usc... CSCI 572 [NWDB] CCWC201... 2018 Workouts Par... https://docs.google... Slickdeals: The Bes... FMF: deals, discuss... 2 Hour Weigh-In: A...

localhost/hw5?q=pnther&choice=solr

Search: pnther Submit ● Solr (Default) ○ PageRank

Did you mean: panther

Results 0 - 0 of 0:

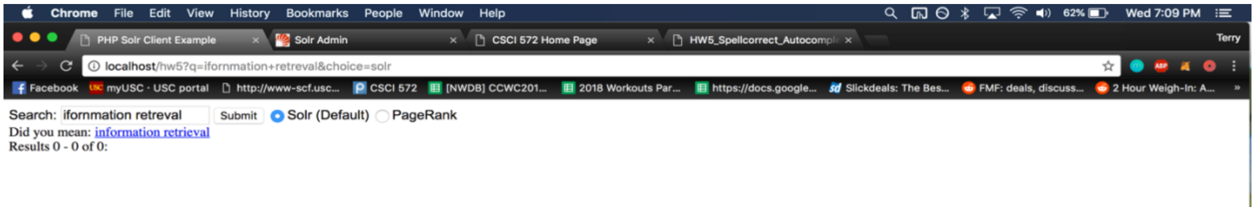
3. 
Chrome File Edit View History Bookmarks People Window Help
75% Thu 9:30 AM
PHP Solr Client Example
Solr Admin
Terry
Facebook myUSC · USC portal http://www-scf.usc... CSCI 572 [NWDB] CCWC201... 2018 Workouts Par... https://docs.google... Slickdeals: The Bes... FMF: deals, discuss... 2 Hour Weigh-In: A...

localhost/hw5?q=brack&choice=solr

Search: brack Submit ● Solr (Default) ○ PageRank

Did you mean: back

Results 0 - 0 of 0:

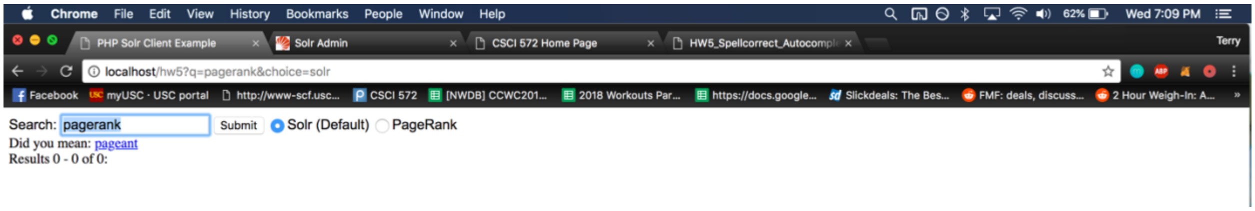
4. 
Chrome File Edit View History Bookmarks People Window Help
62% Wed 7:09 PM
PHP Solr Client Example
Solr Admin
CSCI 572 Home Page
HW5_Spellcorrect_Autocompl
Terry
Facebook myUSC · USC portal http://www-scf.usc... CSCI 572 [NWDB] CCWC201... 2018 Workouts Par... https://docs.google... Slickdeals: The Bes... FMF: deals, discuss... 2 Hour Weigh-In: A...

localhost/hw5?q=ifornmation+retreval&choice=solr

Search: ifornmation retreval Submit ● Solr (Default) ○ PageRank

Did you mean: information retrieval

Results 0 - 0 of 0:

5. 
Chrome File Edit View History Bookmarks People Window Help
62% Wed 7:09 PM
PHP Solr Client Example
Solr Admin
CSCI 572 Home Page
HW5_Spellcorrect_Autocompl
Terry
Facebook myUSC · USC portal http://www-scf.usc... CSCI 572 [NWDB] CCWC201... 2018 Workouts Par... https://docs.google... Slickdeals: The Bes... FMF: deals, discuss... 2 Hour Weigh-In: A...

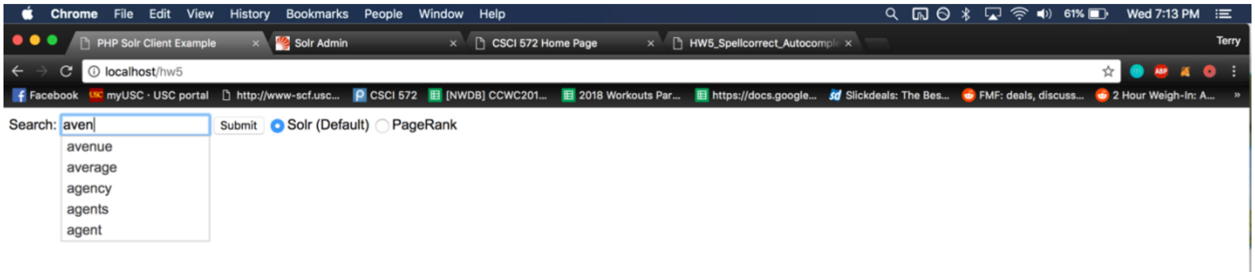
localhost/hw5?q=pagerank&choice=solr

Search: pagerank Submit ● Solr (Default) ○ PageRank

Did you mean: pageant

Results 0 - 0 of 0:

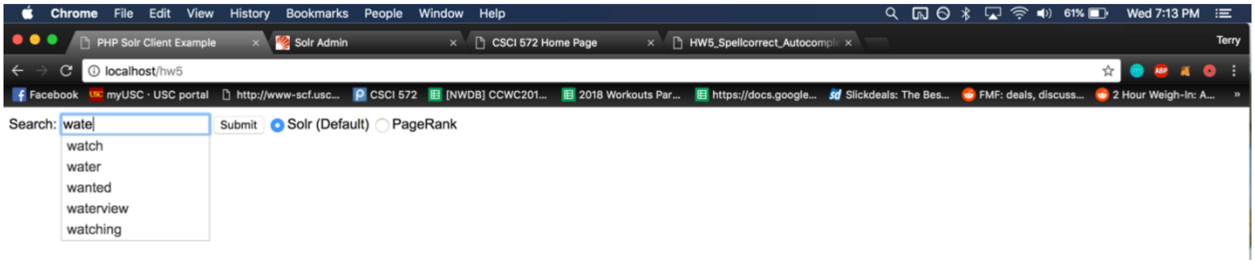
Auto-Completion


Chrome File Edit View History Bookmarks People Window Help
Wed 7:13 PM
PHP Solr Client Example
Solr Admin
CSCI 572 Home Page
HW5_Spellcorrect_Autocompl
localhost/hw5

Search: aven Submit ● Solr (Default) ○ PageRank

avenue
average
agency
agents
agent

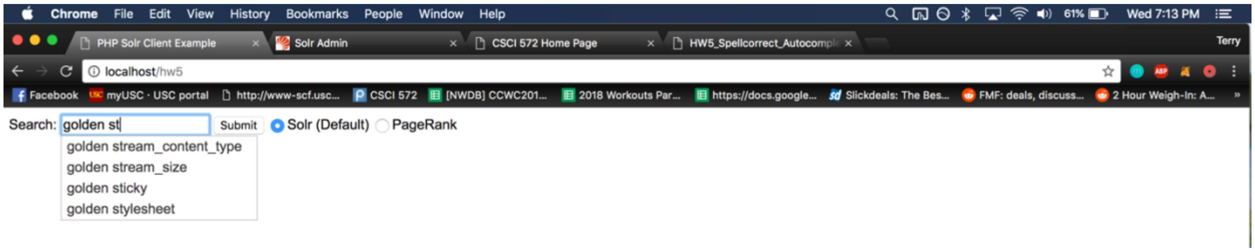
1.


Chrome File Edit View History Bookmarks People Window Help
Wed 7:13 PM
PHP Solr Client Example
Solr Admin
CSCI 572 Home Page
HW5_Spellcorrect_Autocompl
localhost/hw5

Search: wate Submit ● Solr (Default) ○ PageRank

watch
water
wanted
waterview
watching

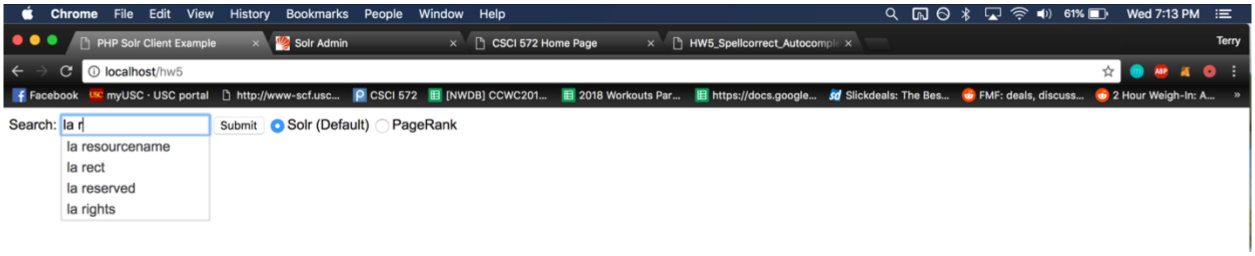
2.


Chrome File Edit View History Bookmarks People Window Help
Wed 7:13 PM
PHP Solr Client Example
Solr Admin
CSCI 572 Home Page
HW5_Spellcorrect_Autocompl
localhost/hw5

Search: golden st Submit ● Solr (Default) ○ PageRank

golden stream_content_type
golden stream_size
golden sticky
golden stylesheet

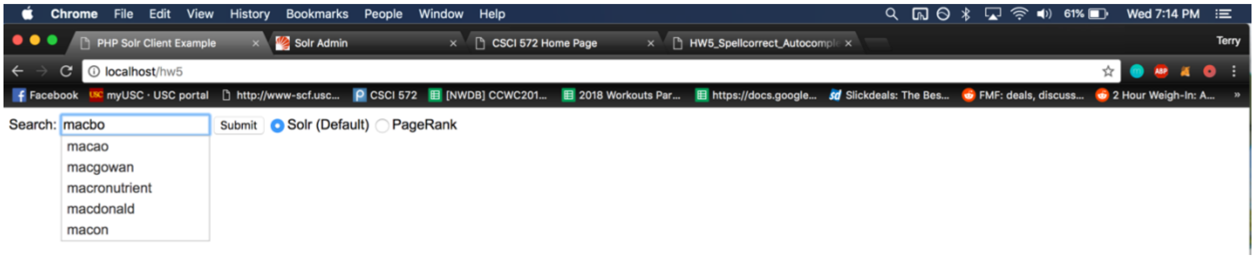
3.


Chrome File Edit View History Bookmarks People Window Help
Wed 7:13 PM
PHP Solr Client Example
Solr Admin
CSCI 572 Home Page
HW5_Spellcorrect_Autocompl
localhost/hw5

Search: la r Submit ● Solr (Default) ○ PageRank

la resourcename
la rect
la reserved
la rights

4.


Chrome File Edit View History Bookmarks People Window Help
Wed 7:14 PM
PHP Solr Client Example
Solr Admin
CSCI 572 Home Page
HW5_Spellcorrect_Autocompl
localhost/hw5

Search: macbo Submit ● Solr (Default) ○ PageRank

macao
macgowan
macronutrient
macdonald
macon

5.