CSCI 572 Assignment #5 – Adding Spell Checking, AutoComplete and Snippets to Your Search Engine

STEPS FOLLOWED DURING EXECUTION:

Step 1: Generated big.txt:

• Extracted text from Nypost crawled data sets using Tika parser written in Java and added the text into big.txt, which is the dictionary for spell check.

Step 2: Spell Correction:

- I used Peter Norvig's Spelling corrector php version code to implement spell correction.
- According to Norvig's algorithm, the correct() function generates all the candidates that are at the minimum edit distance from the query term. From all the candidates term the term closest to the query term in the dictionary is selected and results are displayed for that corrected_query(used in Main.php code) term.
- When a user enters an incorrect query, the program checks if it's a correct word or not. If it's a correct word then relevant results are displayed else it gives a suggestion with the correct word.
- Once the user clicks on the keyword, the appropriate results are displayed.

Step 3: Auto complete:

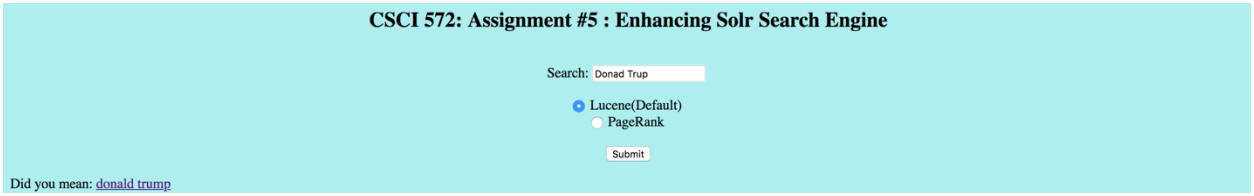
- Changed the Solrconfig.xml file by adding the suggester component as mentioned in homework pdf.
- Autocomplete functionality is added in Main.php(the php code of 4th assignment). This php uses the PREFIX URL: http://localhost:8983/solr/myexample/suggest?q=" and SUFFIX url parameters: "&wt=json&indent=true"; to extract suggestions which is sent back to the home page.
- Further, I used loop of count indicating the number of suggestions you need for autocomplete.
- The results of auto complete are thus displayed in a dropdown list.

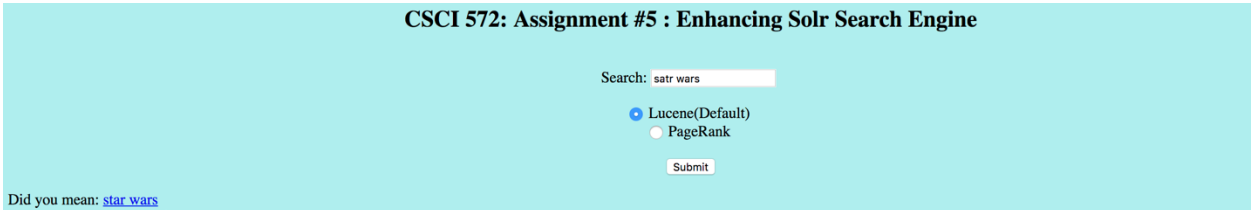
Step 4: Snippet:

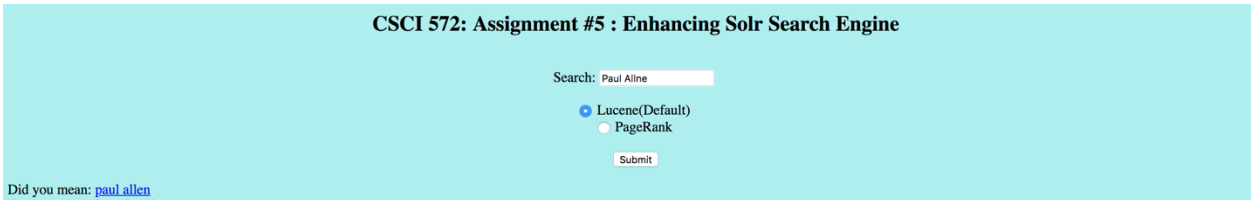
- Used file_get_contents function that obtained the contents of the given filename and divided the content into sentences and words.
- Executed two loops, one for sentences and other for words to find the query terms from the contents of the given filename and find the index of the terms in the string that is used to print that part string along with ellipses (…) either at the beginning or at the end of the snippet.
- If a snippet is found, then it is displayed with bold query terms and if snippet is not found then N/A is displayed.

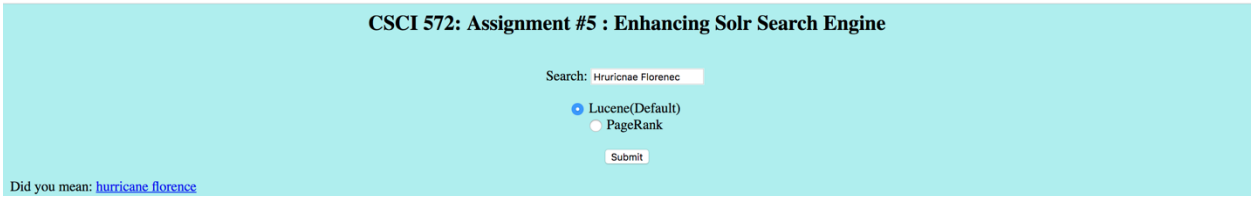
ANALYSIS OF THE RESULTS:

Spell correction:



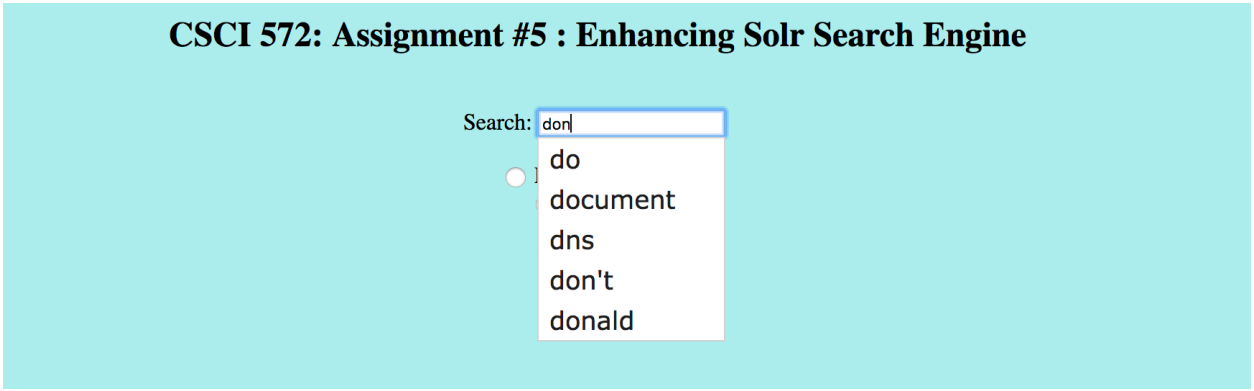


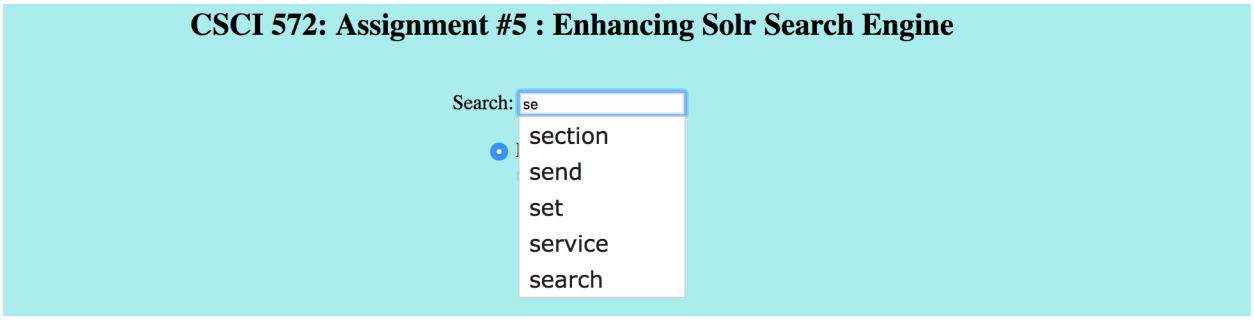


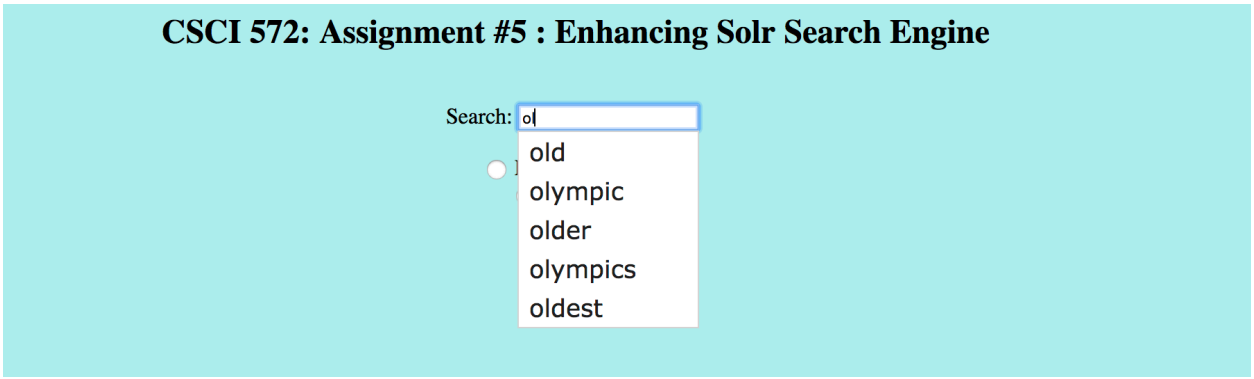


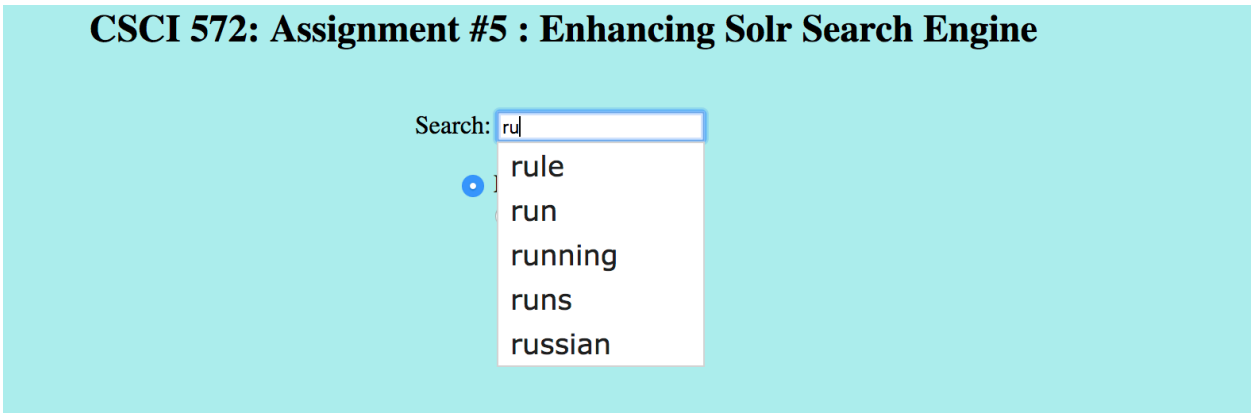
Incorrect Word	Correct Word
Donad Trup	Donald Trump
Satr Wars	Star Wars
Lerbon Jamse	Lebron James
Paul Allne	Paul Allen
Hruricnae Florenec	Hurricane Florence

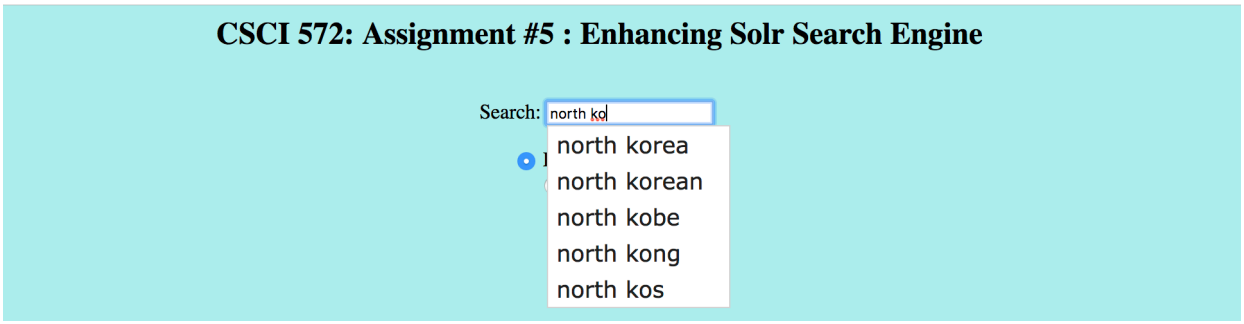
Auto-complete:











Prefix	Autocompletion
don	do document dns don't donald

se	section send set service search
ol	old olympic older olympics oldest
ru	rule run running runs russian
north ko	north korea north korean north kobe north kong north kos