**CSCI 572 HW5: Enhancing Your Search Engine**

**Name: Manish S. Dhankani**

**USC: 8605754465**

**Website: Mercury**

**Step 1: Spelling Correction**

* To implement this functionality, I used the Peter Norvig’s Spelling corrector code provided in the PHP to find out the incorrect terms in my query.
* Peter Norvig’s code requires big.txt file as input, this file contains all terms in the inverted index of the search engine.
* I wrote a java program to generate a big.txt file from the HTML files downloaded from the google drive. I used htmlparser to parser the downloaded HTML files to big.txt file.
* I used the Peter Norvig text and code to check a test file whether it is working or not and then I used Peter Norvig’s code to generate serialized\_dictionary.txt file.
* The correct() function of Norvig’s code generate 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 ranking.php code) term.

**Step 2: Autocomplete**

* For including the autocomplete functionality I made modification in Solrconfig.xml file. I included the suggest component as given in exercise pdf.
* In the ranking.php code of 4th Assignment I added some of the functionalities to enable autocomplete on my UI. For this I used jquery code snippet from the link given on piazza
* Autocomplete function is included in ranking.php. In this function, I used two variables prefix and suffix. Prefix store the term suggested by solr corresponding to query term and suffix is for setting wt to json. Then the query terms are stored in lowercase along with character count and space count.
* Then I stored prefix and suffix in a variable called URL along with correct term. After this the suggested data is parsed in json and parsed data suggestions are stored in results variable.
* I used loop of count (how many suggestions you want to show for autocomplete) to show all the suggestions for the query term written by user.

**Step 3: Snippet**

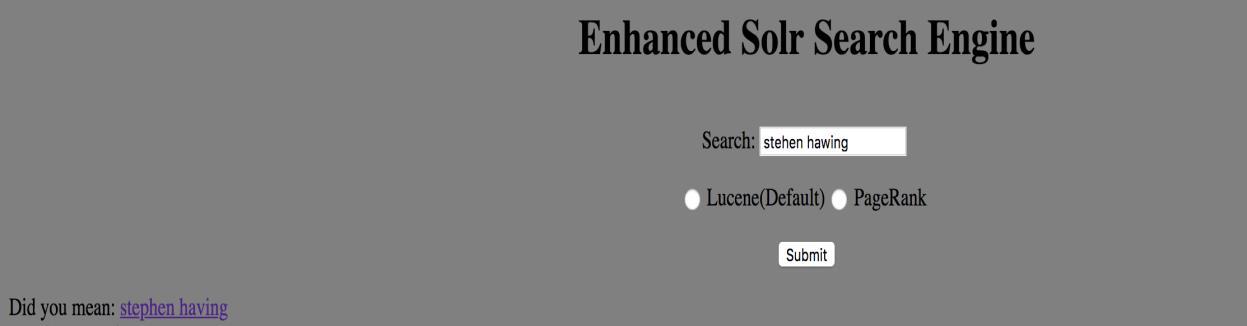
* To enable this functionality, I have included simple\_html\_dom.php file in my ranking.php file. simple\_html\_dom.php is used to parse html pages of NBC\_News and generate the text out of it.
* I used file\_get\_contents function that extracted the contents of the given filename. Then I divided the content into sentences and words. Also removes the special characters so that they don’t affect in keywords matching.
* I run 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 I displayed the snippet with bold query terms and if snippet is not found then I displayed N/A.

**ANALYSIS OF RESULTS**

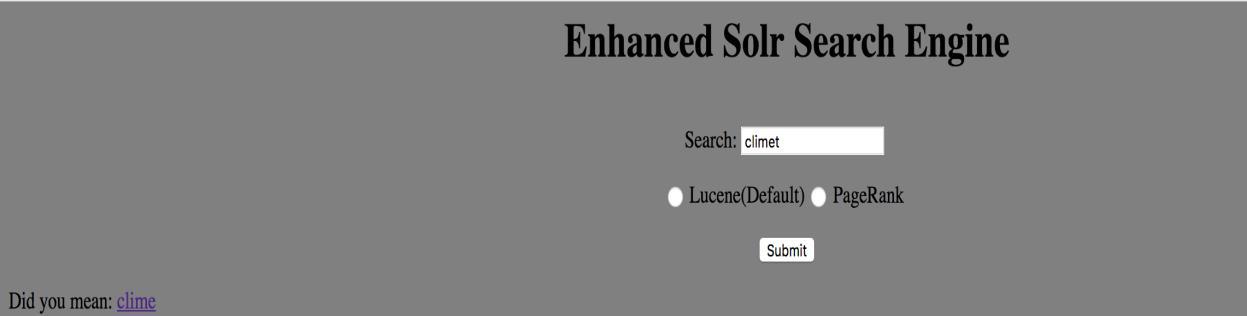
**Examples of Spell Correction**

Misspelled words are taken from demo of this assignment:

**1. Stehen hawing corrected result Stephen having**

****

**2. Climet corrected result clime**

****

**3. Bitcni corrected result bitcoin**

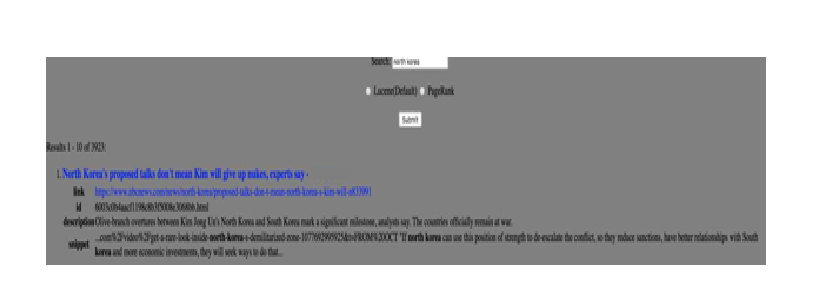
****

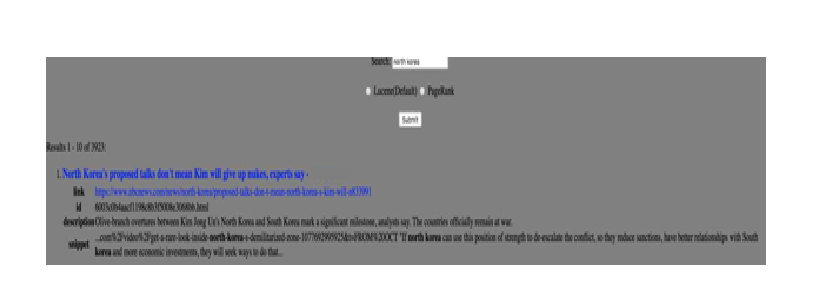
****

1. **Jamse coemy corrected result james comey**
2. **Norh korae corrected result north korea**

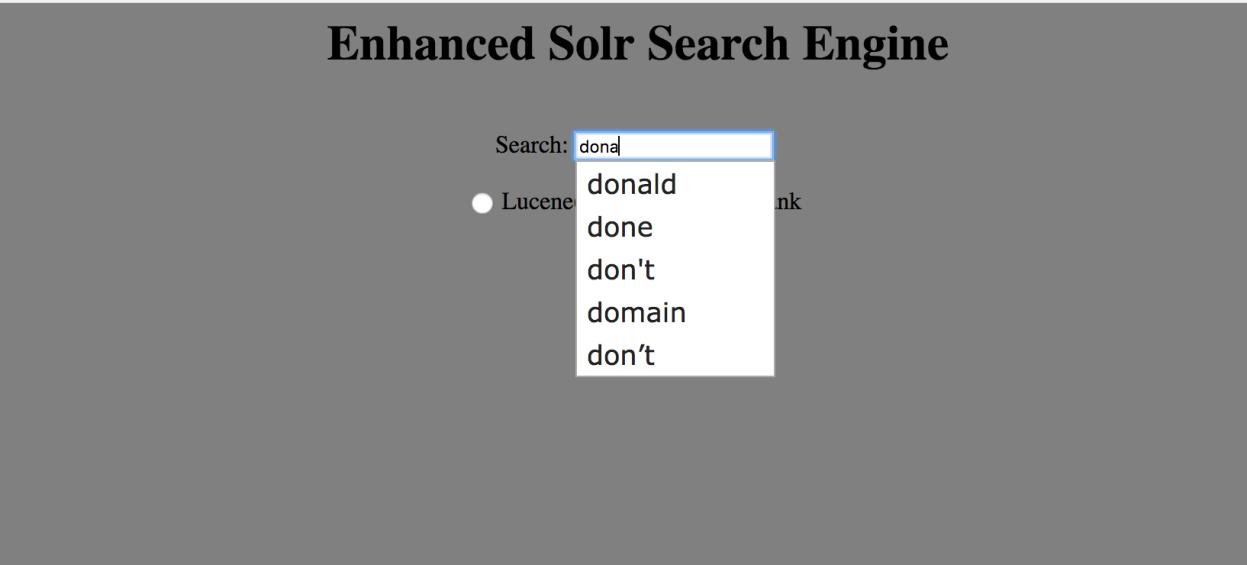
I clicked on the North Korea and the results are shown below along with snippet and highlighted query terms in it. Below is the screen shot of the same.

Snippet:

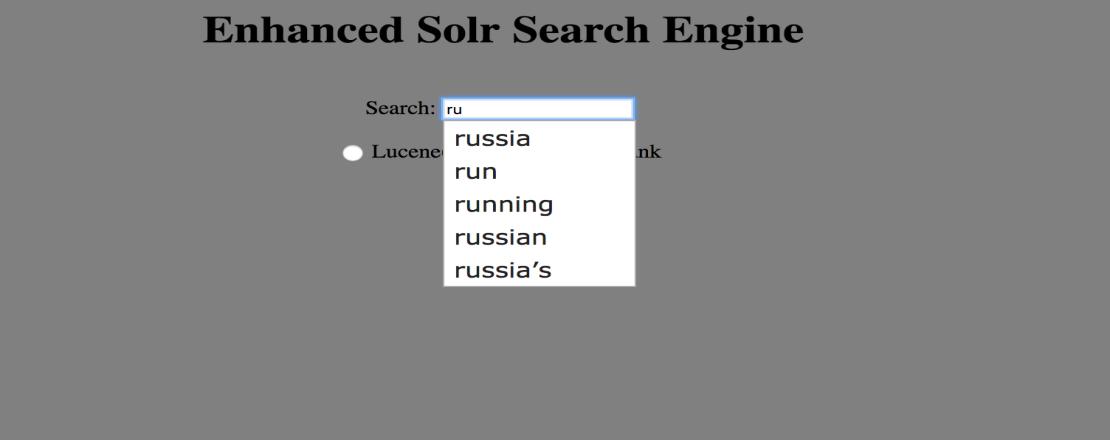




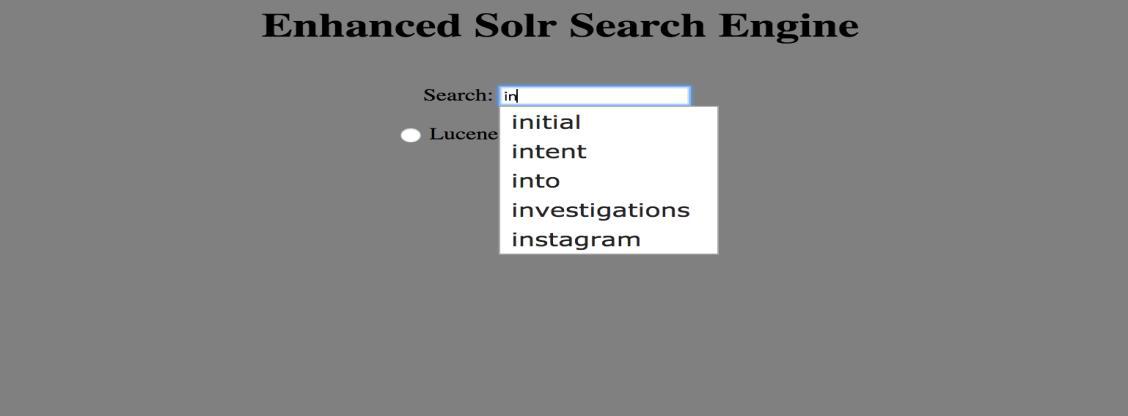
**Examples of Auto-Completion**

**1. dona**

**2. ru**

****

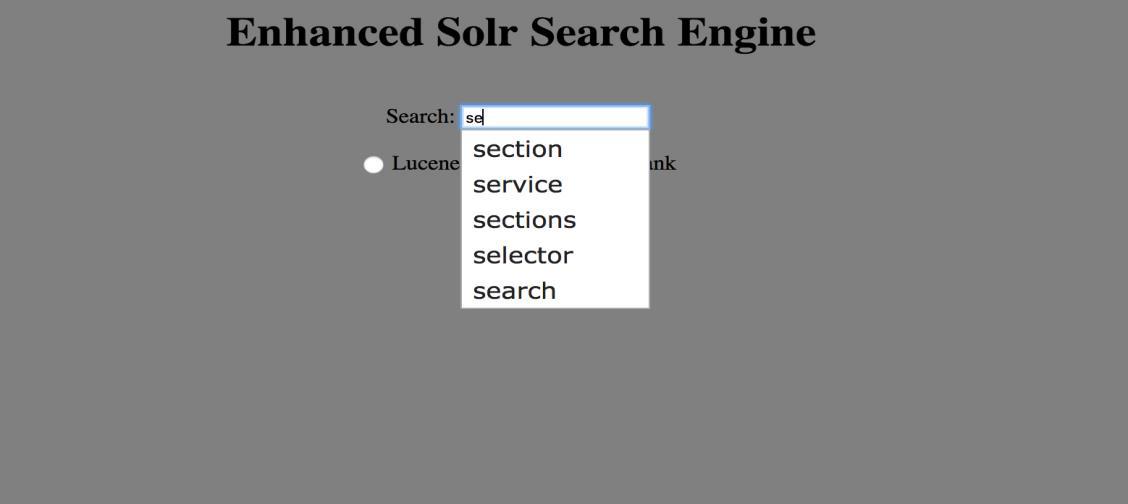
**3. in**

****

**4. ir**

****

**5. se**

****

index.php file contains code of UI, Autocomplete and for implementing spell corrector code.

SpellCorrector.php is an external Peter Norvig’s spell correction code.

big\_file.java contains code to generate big.txt file

simple\_html\_dom.php file contains code to parse html web pages of Mercury News