

# **CSCI 572 Assignment 5: Enhancing Your Search Engine**

**Name: Prakhar Sethi**

**USC ID: 6386461387**

**Website Used: NBC\_NEWS**

## **Steps followed to complete this assignment**

### **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 4<sup>th</sup> 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 <http://api.jqueryui.com/autocomplete/>.
- 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

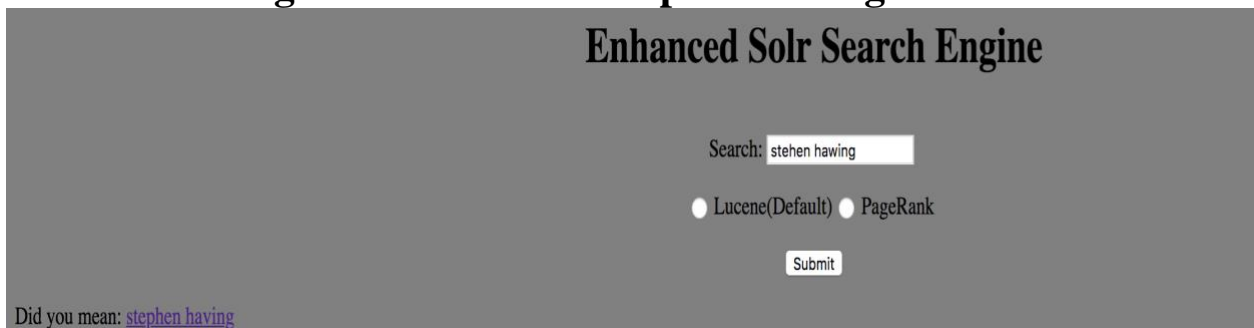
- To enable this functionality, I have included simple\_html\_dom.php file in my ranking.php file. simple\_html\_dom.php is used to parse html pages of NBC\_News and generate the text out of it.
- I used file\_get\_contents function that extracted the contents of the given filename. Then I divided the 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

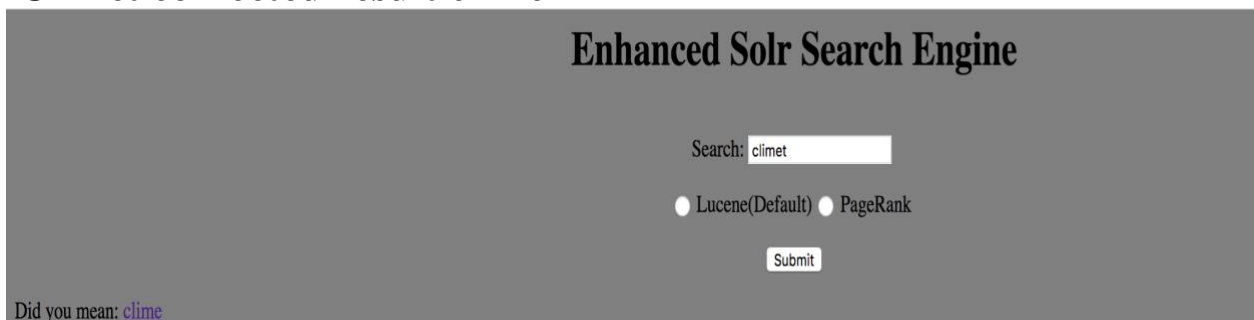
Misspelled words are taken from demo of this assignment:

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



The screenshot shows the 'Enhanced Solr Search Engine' interface. At the top, the title 'Enhanced Solr Search Engine' is displayed. Below it, there is a search bar with the text 'Search: stehen hawing'. Under the search bar, there are two radio buttons: 'Lucene(Default)' and 'PageRank'. Below the radio buttons is a 'Submit' button. At the bottom of the interface, it says 'Did you mean: [stephen having](#)'.

#### 2. Climet corrected result clime



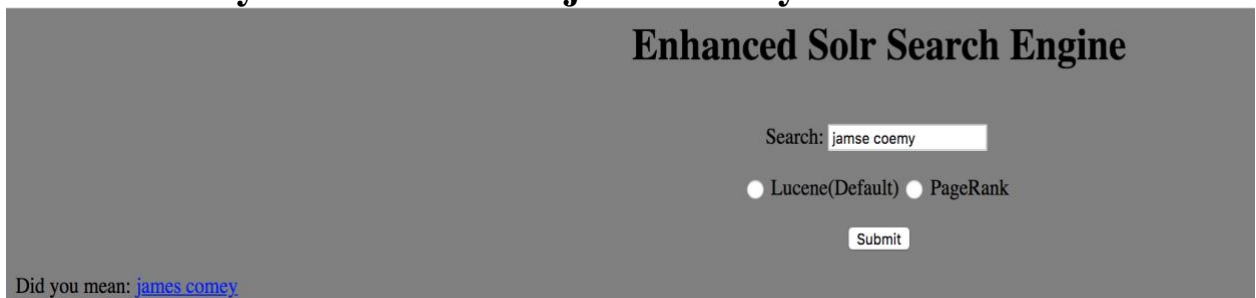
The screenshot shows the 'Enhanced Solr Search Engine' interface. At the top, the title 'Enhanced Solr Search Engine' is displayed. Below it, there is a search bar with the text 'Search: climet'. Under the search bar, there are two radio buttons: 'Lucene(Default)' and 'PageRank'. Below the radio buttons is a 'Submit' button. At the bottom of the interface, it says 'Did you mean: [clime](#)'.

### 3. Bitcni corrected result bitcoin



The screenshot shows the 'Enhanced Solr Search Engine' interface. At the top right, the title 'Enhanced Solr Search Engine' is displayed in a large, bold, black serif font. Below the title, on the right side, is a search input field containing the text 'bitcni'. To the left of the input field is the label 'Search:'. Below the input field are two radio buttons: 'Lucene(Default)' and 'PageRank'. Below the radio buttons is a 'Submit' button. At the bottom left of the interface, the text 'Did you mean: [bitcoin](#)' is displayed, where 'bitcoin' is a blue hyperlink.

### 4. Jamse coemy corrected result james comey



The screenshot shows the 'Enhanced Solr Search Engine' interface. At the top right, the title 'Enhanced Solr Search Engine' is displayed in a large, bold, black serif font. Below the title, on the right side, is a search input field containing the text 'jamse coemy'. To the left of the input field is the label 'Search:'. Below the input field are two radio buttons: 'Lucene(Default)' and 'PageRank'. Below the radio buttons is a 'Submit' button. At the bottom left of the interface, the text 'Did you mean: [james comey](#)' is displayed, where 'james comey' is a blue hyperlink.

### 5. Norh korae corrected result north korea



The screenshot shows the 'Enhanced Solr Search Engine' interface. At the top right, the title 'Enhanced Solr Search Engine' is displayed in a large, bold, black serif font. Below the title, on the right side, is a search input field containing the text 'norh korae'. To the left of the input field is the label 'Search:'. Below the input field are two radio buttons: 'Lucene(Default)' and 'PageRank'. Below the radio buttons is a 'Submit' button. At the bottom left of the interface, the text 'Did you mean: [north korea](#)' is displayed, where 'north korea' is a blue hyperlink.

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.

Search: north korea

● Lucene(Default) ● PageRank

Submit

Results 1 - 10 of 3923:

- North Korea's proposed talks don't mean Kim will give up nukes, experts say - NBC News**  
**link** <https://www.nbcnews.com/news/north-korea/proposed-talks-don-t-mean-north-korea-s-kim-will-n833991>  
**id** 6003c0b4aacf1198c8b3f5008e3068bb.html  
**description** Olive-branch overtures between Kim Jong Un's North Korea and South Korea mark a significant milestone, analysts say. The countries officially remain at war.  
**snippet** ...com%2Fvideo%2Fget-a-rare-look-inside-north-korea-s-demilitarized-zone-1077692995925&t=FROM%20OCT "If **north korea** can use this position of strength to de-escalate the conflict, so they reduce sanctions, have better relationships with South korea and more economic investments, they will seek ways to do that...
- North Korea Says It Has Tested Hydrogen Bomb That Can Fit on ICBM - NBC News**  
**link** <https://www.nbcnews.com/news/north-korea/5-1-magnitude-tremor-recorded-north-korea-could-be-new-n798376>  
**id** edb8caf082e2660a40dc75992cha2dc9.html  
**description** North Korea on Sunday claimed to have successfully tested a hydrogen bomb meant for an intercontinental ballistic missile.  
**snippet** ... official said Sunday that experts were reviewing the data but "we are highly confident this was a test of an advanced nuclear device and what we've seen so far is not inconsistent with **north korea's** claims...
- North Korea Says It Has Tested Hydrogen Bomb That Can Fit on ICBM - NBC News**  
**link** <https://www.nbcnews.com/news/north-korea/north-korea-says-it-tested-hydrogen-bomb-can-fit-icbm-n798376>  
**id** da37bc80a8d29fc1c21799720ee22be.html  
**description** North Korea on Sunday claimed to have successfully tested a hydrogen bomb meant for an intercontinental ballistic missile.  
**snippet** ... official said Sunday that experts were reviewing the data but "we are highly confident this was a test of an advanced nuclear device and what we've seen so far is not inconsistent with **north korea's** claims...
- North Korea, South Korea could meet for talks ahead of Olympics - NBC News**  
**link** <https://www.nbcnews.com/news/north-korea/north-korea-south-korea-could-meet-talks-ahead-olympics-n833941>  
**id** 756b20a8dce5eae1275d485d61ed993.html  
**description** Delegations from North and South Korea could meet for the first official discussions between the neighbors since 2015 ahead of the PyeongChang Winter Olympics.  
**snippet** ...com%2Fnightly-news%2Fvideo%2Fkim-jong-un-entire-u-s-is-within-range-of-north-korea-s-nuclear-weapons-1127330371663&t=Kim%20Jong%20Un%3A%20Entire%20U However, analysts say that based on the current evidence it's hard to prove or debunk **north korea's** claim that it can now hit faraway American targets such as New York or Washington...
- North Korea crisis: How events have unfolded under Trump - NBC News**  
**link** <https://www.nbcnews.com/news/world/north-korea-crisis-how-events-have-unfolded-under-trump-n753996>  
**id** 546001d092d2e32915ee4ee70128f61.html  
**description** North Korea has thumbed its nose at U.N. resolutions — conducting several ballistic missile tests this year and five nuclear tests since 2006  
**snippet** ... Nonetheless, **north korea** thumbed its nose at the resolutions — conducting several ballistic missile **north korea** called it a provocation can take with **north korea**, and how nearly all of them cannot force Kim Jong-un's hand on American policy priorities...
- North Korea crisis: How events have unfolded under Trump - NBC News**  
**link** <https://www.nbcnews.com/news/world/ump/north-korea-crisis-how-events-have-unfolded-under-trump-n753996>  
**id** 6c05c1dc750503997735736fa6f8aa3.html  
**description** North Korea has thumbed its nose at U.N. resolutions — conducting several ballistic missile tests this year and five nuclear tests since 2006  
**snippet** ... Nonetheless, **north korea** thumbed its nose at the resolutions — conducting several ballistic missile **north korea** called it a provocation military officials said it was intended to "send a clear message" to **north korea** following the ICBM test...
- U.S. drills in South Korea trigger 'nuclear war' warning from North - NBC News**  
**link** <https://www.nbcnews.com/news/north-korea/u-s-drills-south-korea-trigger-nuclear-war-warning-north-n826176>  
**id** 4ace0074c481b07644f4e3993b2a7ae6.html  
**description** Two dozen U.S. stealth jets were among hundreds of aircraft involved in war games intended as a show of strength to neighboring North Korea on Monday.

## Examples of Auto-Completion

### 1. dona

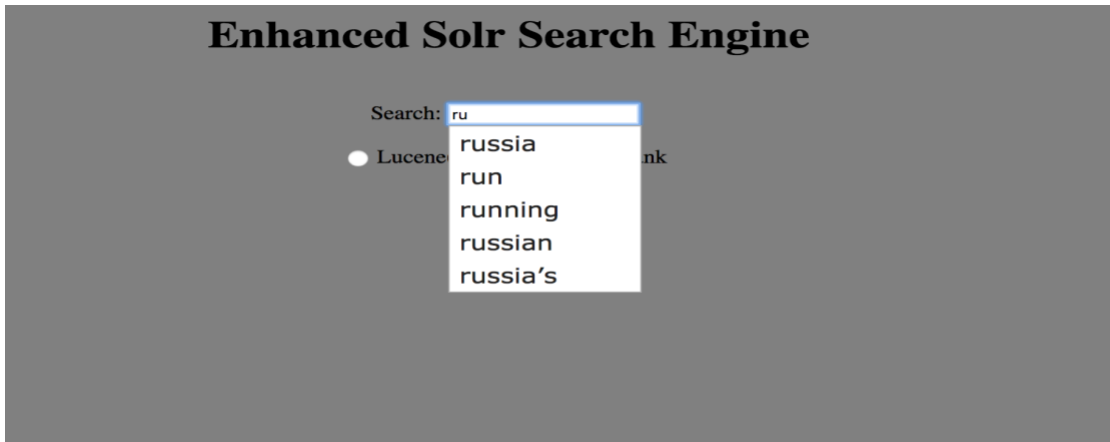
## Enhanced Solr Search Engine

Search: dona

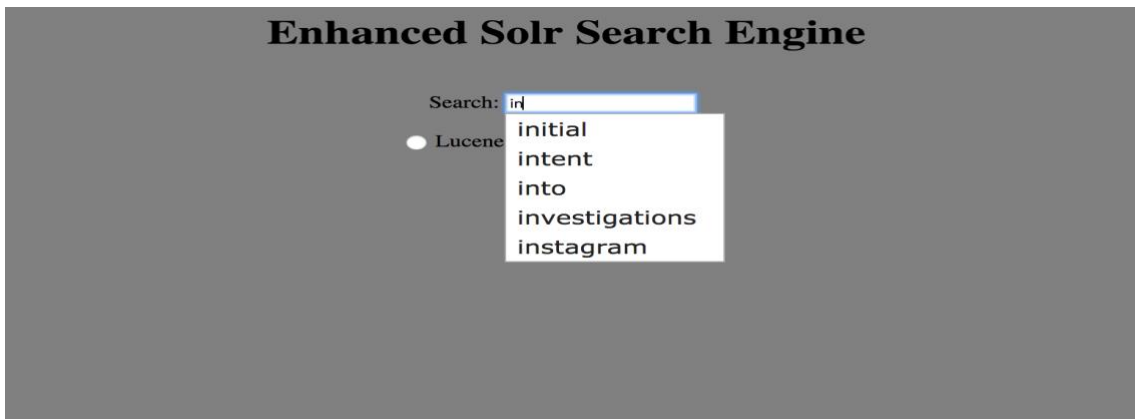
● Lucene

- donald
- done
- don't
- domain
- don't

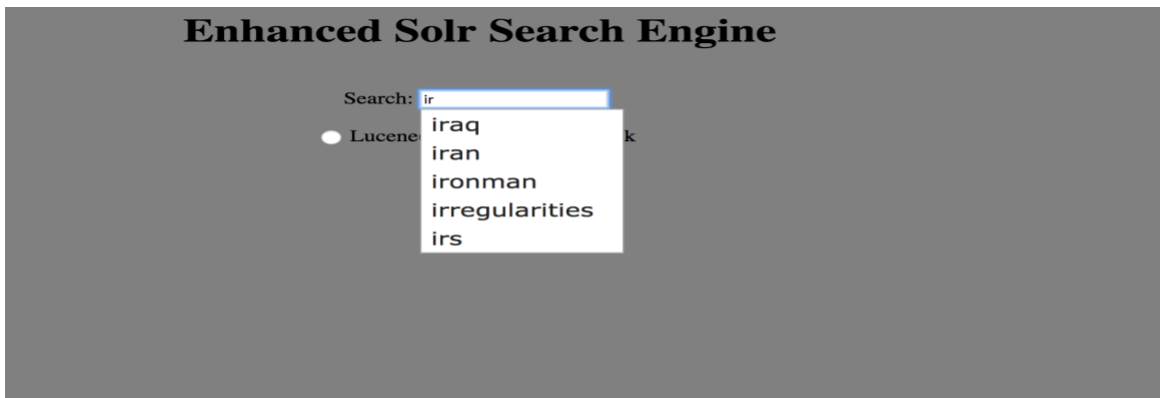
2. ru



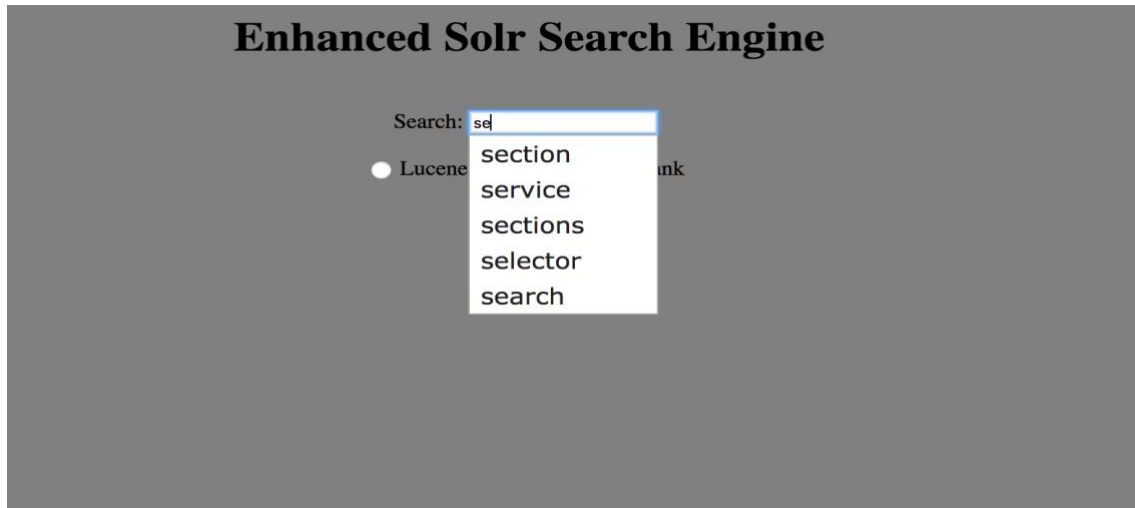
3. in



4. ir



## 5. se



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

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

**Bigtxt.java** contains code to generate big.txt file

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