

# Homework-5: Enhancing Your Search Engine

Name: **Sachin Hanumantha Reddy**

USC ID: **7112139958**

News Site: **Mercury News**

## Steps:

### **Step-1: Spelling Correction**

- Peter Norvig's Spelling corrector code was used in PHP language to determine the incorrect terms in the given query.
- It takes big.txt as input which contains all terms in the inverted index of the search engine.
- Generated a big.txt file from the HTML files downloaded from the google drive using htmlparser to parser the downloaded HTMLfiles.
- The correct() function of Norvig's code generate all the candidates that are at a minimum edit distance from the query term. From all the candidates terms, the term closest to the query term in the dictionary is selected and results are displayed for the corrected query.

### **Step-2: Autocomplete Functionality**

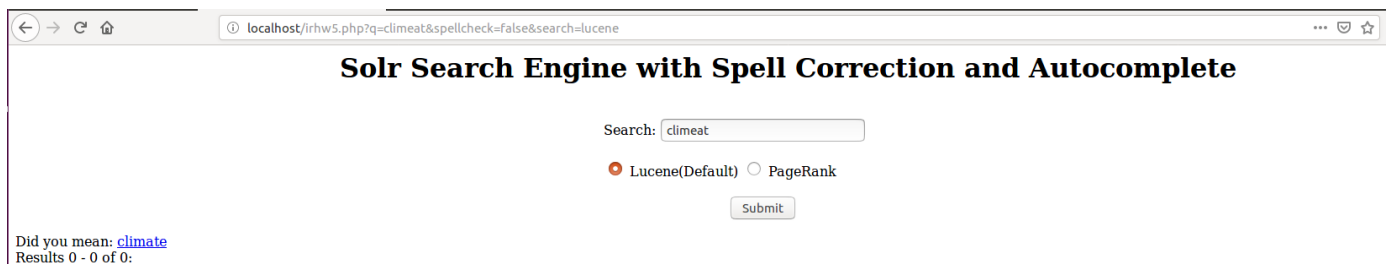
- Included the suggest component as given in the homework description. Edited the SolrConfig.xml accordingly.
- I used loop count of number of suggestions to be shown to show all the suggestions for the query term submitted.

## Step-3: Snippet Generation

- simple\_html\_dom.php file is included to enable this feature where it retrieves a passage which is most relevant to the query in terms of most query terms in it, same sequence of the query terms in the passage etc.
- Divided the contents of extracted data from the given file name into sentences and words. Removed the special characters so that they don't affect in keywords matching.
- There are two iterations, 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 the snippet is displayed with bold query terms and if snippet is not found then N/A is shown.

## Examples of Spell Correction:

### 1. Climate queried as Climeat



localhost/irhw5.php?q=climeat&spellcheck=false&search=lucene

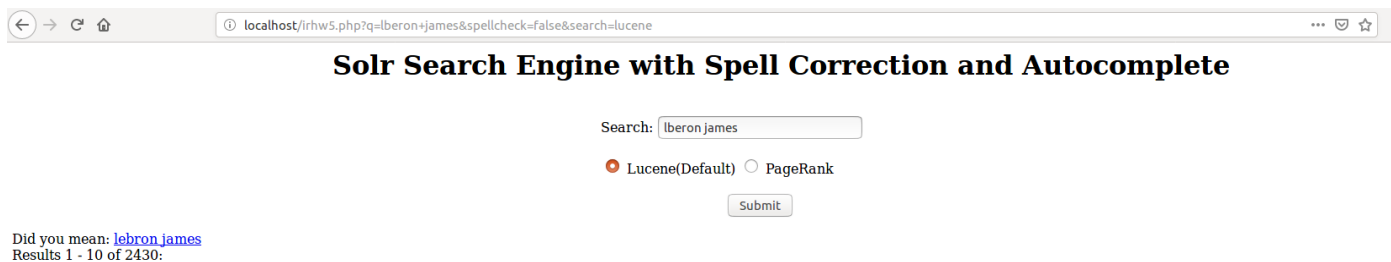
### Solr Search Engine with Spell Correction and Autocomplete

Search:

☒ Lucene(Default) ☐ PageRank

Did you mean: [climate](#)  
Results 0 - 0 of 0:

### 2. Lebron James queried as Lberon James



localhost/irhw5.php?q=lberon+james&spellcheck=false&search=lucene

### Solr Search Engine with Spell Correction and Autocomplete

Search:

☒ Lucene(Default) ☐ PageRank

Did you mean: [lebron james](#)  
Results 1 - 10 of 2430:

### 3. Donald Trump queried as Donald Tramp

localhost/irhw5.php?q=donald+tramp&spellcheck=false&search=ucene

### Solr Search Engine with Spell Correction and Autocomplete

Search:

☒ Lucene(Default) ☐ PageRank

Did you mean: [donald trump](#)  
Results 1 - 10 of 984:

### 4. Google queried as Googel

localhost/irhw5.php?q=googel&spellcheck=false&search=ucene

### Solr Search Engine with Spell Correction and Autocomplete

Search:

☒ Lucene(Default) ☐ PageRank

Did you mean: [google](#)  
Results 0 - 0 of 0:

### 5. La Lakers queried as La Lekers

localhost/irhw5.php?q=la+lekers&spellcheck=false&search=ucene

### Solr Search Engine with Spell Correction and Autocomplete

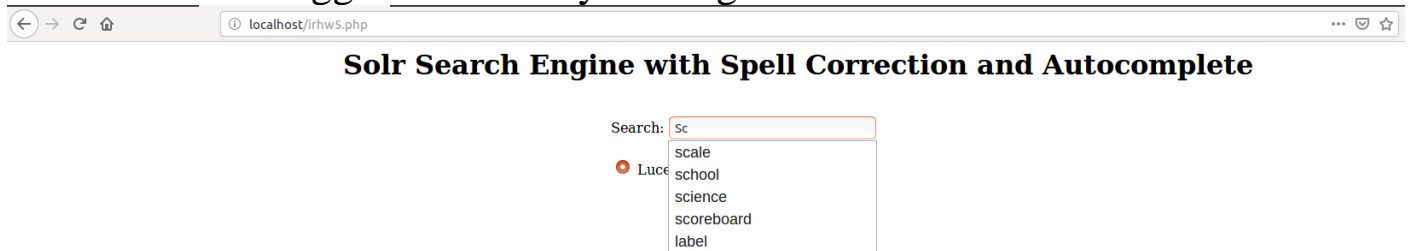
Search:

☒ Lucene(Default) ☐ PageRank

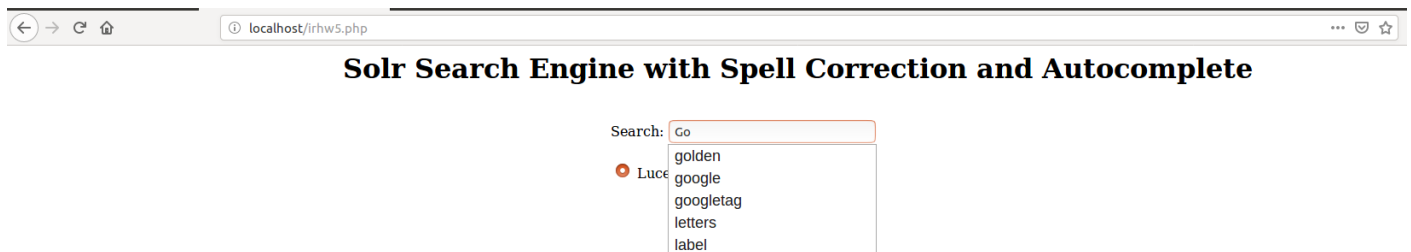
Did you mean: [la lakers](#)  
Results 1 - 10 of 577:

## Examples of Auto-Completion:

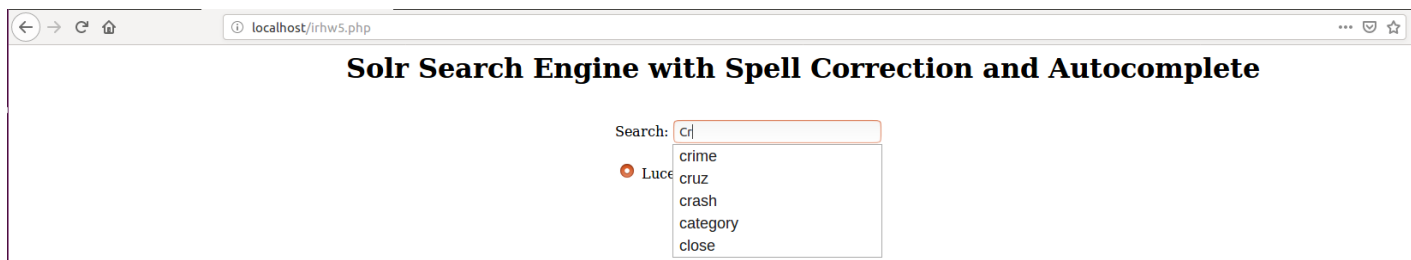
### 1. Sc shows suggestions mostly starting from Sc



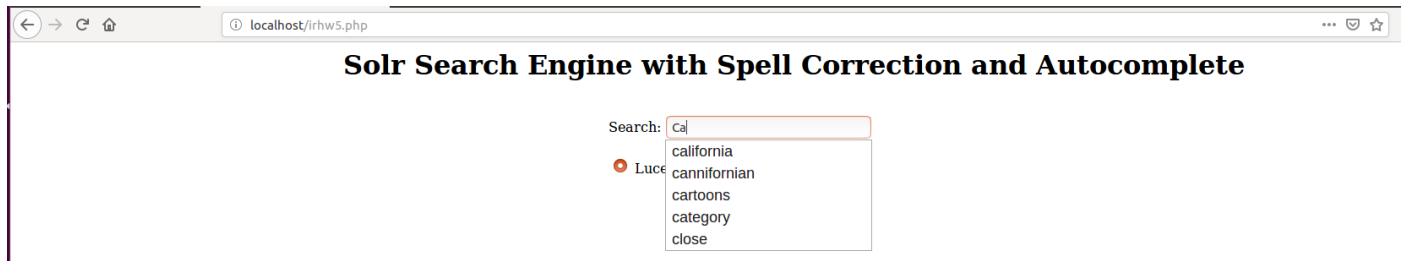
### 2. Go shows suggestions mostly starting from Go



### 3. Cr shows suggestions mostly starting from Cr



4. Ca shows suggestions mostly starting from Ca



5. Li shows suggestions mostly starting from Li

