CSCI572 HW#5 Report

Name: Lingquan Han ID:3919234233

1. Steps to complete the assignment

1.1 Generate "big.txt" file

Firstly, to generate the file of "big.txt" from latimes html files, I wrote the java code and used Tika parser package to implement it. The "big.txt" will be used as dictionary for spelling correction and auto-completion features.

1.2 Spelling correction

With the suggestion in assignment document, I chose the Norvig's spelling program written in PhP to implement spelling correction feature. To do this, I downloaded SpellCorrector.php from https://www.phpclasses.org/package/4859-PHP-Suggest-corrected-spelling-text-in-pure-PHP.html#download and imported it into index.php file. When users input a query, the program will check if the input is correct by looking up in the "big.txt" file. If the input is not correct, the program will return the correction. After the user click the submit button, the search engine will display the message of "Did you mean XXX?" (where XXX means the correction).

1.3 Auto-completion

For the feature of auto-complettion, I use the FuzzyLookupFactory of Solr/Lucene to implement it. For the first character and second character that is entered as query, it will returns a list of completions/suggestions. While typing in the search box, the top suggestions should automatically appear and be updated as the user keeps typing.

1.4 Snippet

In order to implement Snippet feature, for each search query, the program generates it by the web page that is referred to. The specific process is to look for a string match of the query terms with the web page and then return the first sentence that provides a match. If no match is found, then no snippet is returned. For multiple term queries, the search engine will find a sentence with all the terms together. If not, it will return the sentence that has all the terms in it, even if they are not together or in same order. In addition, simple_html_dom.php file was included to ensure only plaintext will be extracted from html files. When meeting matched string, the search engine will add at its two sides to implement highlight. Save modified matched sentence as snippet and later show it to user.

2. Examples

2.1 Five spelling correction examples

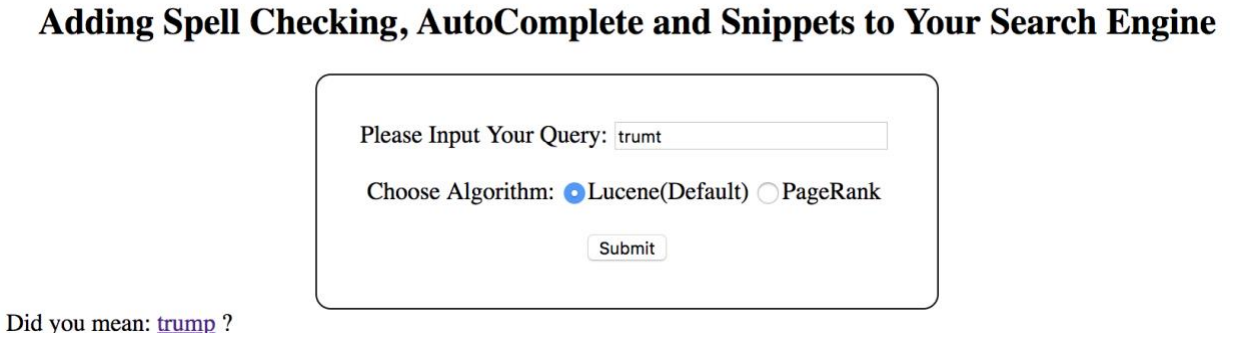
No.	Input	Correction
1	trumt	trump
2	lso angeles	los angeles
3	compuder	computer
4	waetr	water
5	univerisity	university

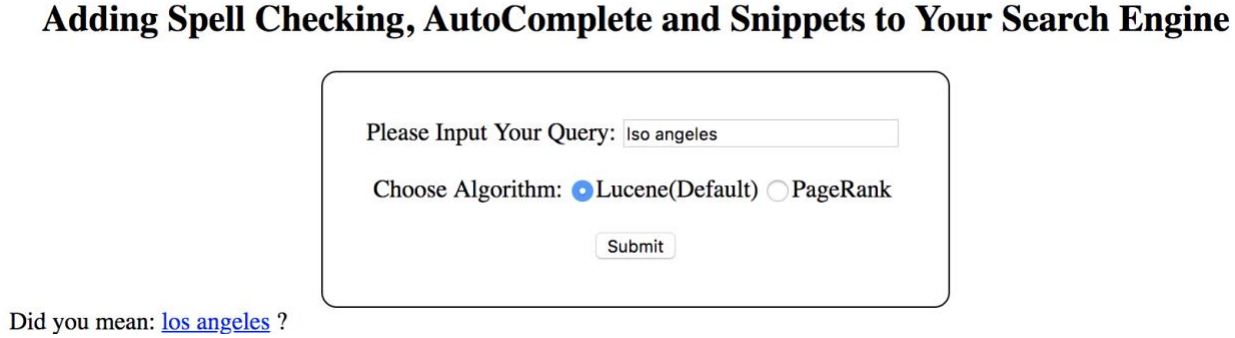
2.2 Five auto-completion examples

No.	Input	Autocomplete
1	autoco	autocomplete, autonomous, automotive, automobile, autonomy
2	oppor	opportunity, opponent, opposition, opponents, opportunities
3	fina	fina, find, final, finally, finds
4	betw	between, better, bet, betty, beth
5	memor	memorable, memoir, memorabilia, memories, memory

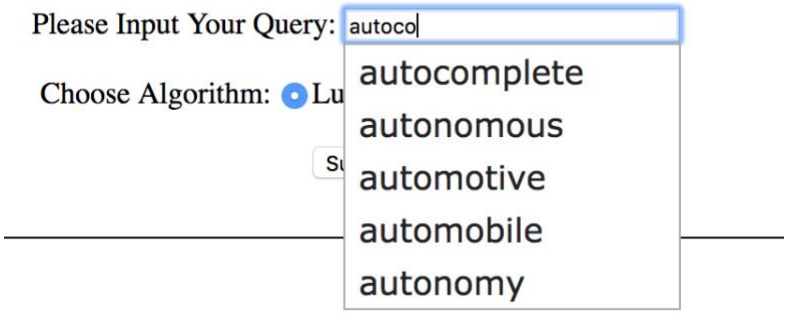
3. Screenshots:

3.1 Spelling correction:





3.2 Auto-completion:



Adding Spell Checking, AutoComplete and Snippets to Your Search Engine