

Homework 5 – Adding Spell Checking, AutoComplete and Snippets to Your Search Engine

Steps executed during the assignment:

Step 1: Spelling Correction – Creating big.txt

- Download **apache-tika-1.2src.zip** and **tika-app-1.2.jar**
- Set up **CLASSPATH** for jar file tika-app-1.2.jar
- Download m2e plugin to integrate **Tika-Maven** into Eclipse
- Create a project in Eclipse and convert the project into **Maven project**
- Edit the pom.xml file of the project to add Maven dependency of Apache Tika
- Write a java program to extract contents of reuters html files using Tika parser and write the generated results into **big.txt**
- Download Norvig's spelling program written in PHP from the website and run it on my server
- When users submit a query, use the function `SpellCorrector::correct(query)` to get correct term from Norvig's spelling program. If the users' entered query is different from the correct term, then show a message of "Show results for [correct term]" and make the [correct term] clickable so users can directly get the results of correct term

Step 2: Autocomplete

- Edit solrconfig.xml to add a search component "suggest component"
- Edit solrconfig.xml to add a request handler "/suggest"
- Reload the core on Solr admin dashboard and check the functionality of suggestion by typing some simple query using /suggest handler
- On my server side, using ajax and jQuery to get suggestion results from Solr when users are typing queries through the below url:
For example, when users type "ca", the request Url will be
http://localhost:8983/solr/myexample/suggest?wt=json&indent=on&q=ca
- In order to also provide spelling check in the suggest list, I will also use the users' input query to check in the Norvig's spelling program and check if the users' input words is correctly spelled. If not, then I will place the correct query in the **top** of the suggestion list.

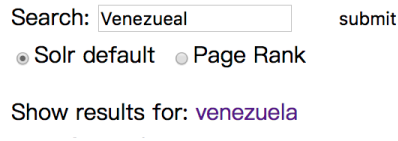
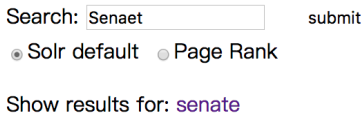
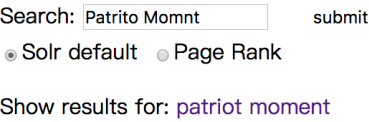
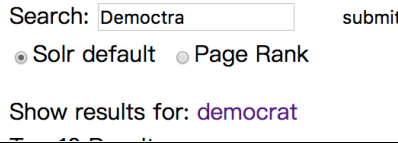
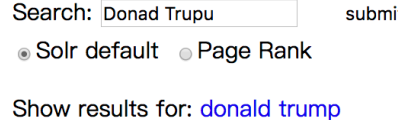
- Use jQuery autocomplete attribute to detect users' input and show the suggest list as a drop down list

Step 3: Create Snippets

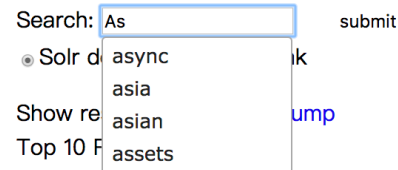
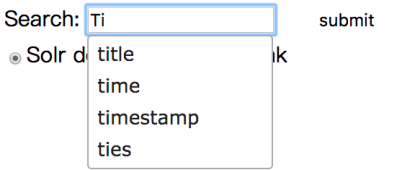
- Download PHP Simple HTML DOM parser file "simple_html_dom.php" for and run it on server for parsing html files
- For each returned results, use **file_get_contents()** to get contents of html file, then use **str_get_html()** (which is from simple_html_dom.php) to convert the content to string
- Loop through each line of html contents, find the first line that contain all the query terms, I used the following comparing methods:
 - a) For single query term, I only need to find the first line that contain this query term
 - b) For multiple query terms, I will first try to find the first line that contain all the query terms together (in the original order). If such a line is found, then break the loop. If not, I will count how many terms from the query the line contains, then compare with other lines in the contents, and get the line that contain the most terms in the query
- In order to restrict the snippets to around 160 words, I used the following trimming methods:
 - a) If the sentence has less than 160 words, use the original sentence as the snippet for this result
 - b) If the sentence has more than 160 words, in order to show the most query terms in the snippets for the users, I will calculate the distance between the first occurrence and the last occurrence of the query terms in the sentence.
 - c) If the distance is less than 160: I will use the substring between first occurrence position, and the start position+160 position of the sentence as the final snippet, and add "..." in the beginning and the end of the sentence if applicable
 - d) If the distance is more than 160: I will also use the substring between first occurrence position, and the start position+160 position of the sentence as the final snippet, and add "..." in the beginning and the end of the sentence if applicable

Analysis of the Results:

SPELL CORRECTION:

Incorrect Word	Correct Word	Screenshot
Venezueal	Venezuela	 <p>Search: Venezueal submit • Solr default • Page Rank Show results for: venezuela</p>
Senaet	Senate	 <p>Search: Senaet submit • Solr default • Page Rank Show results for: senate</p>
Patrito Momnt	Patriot Moment	 <p>Search: Patrito Momnt submit • Solr default • Page Rank Show results for: patriot moment</p>
Democtra	Democrat	 <p>Search: Democtra submit • Solr default • Page Rank Show results for: democrat</p>
Donad Trupu	Donald Trump	 <p>Search: Donad Trupu submit • Solr default • Page Rank Show results for: donald trump</p>

AUTO-COMPLETE:

Prefix	Autocompletion	Screenshot
As	Async Asia Asian Assests	 <p>Search: As submit • Solr default • Page Rank Show results for: as Top 10 Results: assets</p>
Ti	Title Time Timestamp Ties	 <p>Search: Ti submit • Solr default • Page Rank Show results for: ti</p>

Patrito	Patriot Patriots Patriotism Patriotic Patrimoine	Search: <input type="text" value="Patrito"/> submit ● Solr d <input type="text" value="Patrito"/> k patriot patriots patriotism patriotic patrimoine
Massi	Mass Masscre Missing Massive Mission	Search: <input type="text" value="Massi"/> submit ● Solr d <input type="text" value="Massi"/> k mass massacre missing massive mission
Legal Eleci	Legal Elect Legal Electric Legal Election Legal Eleki Legal Elections Legal Electronic	Search: <input type="text" value="Legal Eleci"/> submit ● Solr d <input type="text" value="Legal Eleci"/> k Legal elect Legal electric Legal election Legal eleki Legal elections Legal electronic