CS572 Informational Retrieval and Web Search Engines
Assignment 4 – Spellcheck and Autocomplete

Varun Mittal
USC ID :1904-2446-54

In this assignment 4 I had to currently the enhance the search engine interface which I created in Assignment 3 with the following features:

1.Spell Checking
2.Autocomplete

Steps Followed to Complete the Spell Checking:

Step 1: Download SpellCorrector.php from: http://www.phpclasses.org/package/4859-PHP-Suggest-corrected-spelling-text-in-pure-PHP.html#download.
To be used as an external third party program for search engine interface for doing spell checking.

Used the statement to " **include 'SpellCorrector.php'** " to incorporate it into the client code.

Step 2: I downloaded the Apache Tika parser jars and included them in Eclipse IDE for content extraction of PDF, HTML and DOC files. I downloaded these files and indexed the same in Solr in Assignment 3.

Step 3: The content extracted from all the PDF, HTML and DOC files is compiled together into one text file called big.txt which is to be used as a dictionary for Peter Norvig spell corrector program.

Step 4: Separated the query given by the user into an array of words. Iterated over it and used the statement **"SpellCorrector::correct(word)"** for correcting each word on the basis of big.txt file created earlier in step 2 and then combined all the corrected words on the basis of whitespaces and gave a hyper-link as
"Did you mean: the corrected query" for which when the user clicks on the link he will get the results for "the corrected query".

Tools used for Spell Check feature:

1.Apache Tika Parser
2.SpellCorrector.php file
3. Eclipse IDE

Steps Followed to Complete the Autocomplete:

Step 1: Followed the tutorial given on the website http://www-scf.usc.edu/~csci572/ for configuring the Suggester component in Solr and made the following changes:

The first step is to add a search component to solrconfig.xml and tell it to use the Suggest Component. Wrote the code as follows:

```
<searchComponent name="suggest" class="solr.SuggestComponent">
<lst name="suggester">
<str name="name">suggest</str>
<str name="lookupImpl">FuzzyLookupFactory</str>
<str name="nonFuzzyPrefix">3</str>
<str name="field">_text_</str>
<str name="suggestAnalyzerFieldType">string</str>
</lst>
</searchComponent>
```

After adding the search component, a request handler must be added to solrconfig.xml.

```
<requestHandler name="/suggest" class = "solr.SearchHandler">
<lst name= "defaults">
<str name="suggest">true</str>
<str name="suggest.count">5</str>
<str name="suggest.dictionary">suggest</str>
</lst>
<arr name="components">
<str>suggest</str>
</arr>
</requestHandler>
```

Step 2: For the autocomplete I also include the jquery libraries and the scripts required for making the java script program work and also for making the ajax call to the Solr Suggester configured earlier in Step 1. The jquery libraries can be downloaded from https://jqueryui.com/autocomplete/.

Step 3: Write a java script for the autocomplete feature in which we are passing the textbox value which contains the character and we are also returned the list of suggestions that is specific for the character that the user enters each and every time. I have also used porter stemmer algorithm for stemming the words like calculate, calculations, calculator etc. to root word ‘calculate’.

Tools Used for autocomplete feature:

1. Solr Suggester
2. Jquery Autocomplete Plugin
3 Porter Stemmer program

ANALYSIS OF RESULTS

Examples of misspelled words being handled correctly by my Spell Check Program

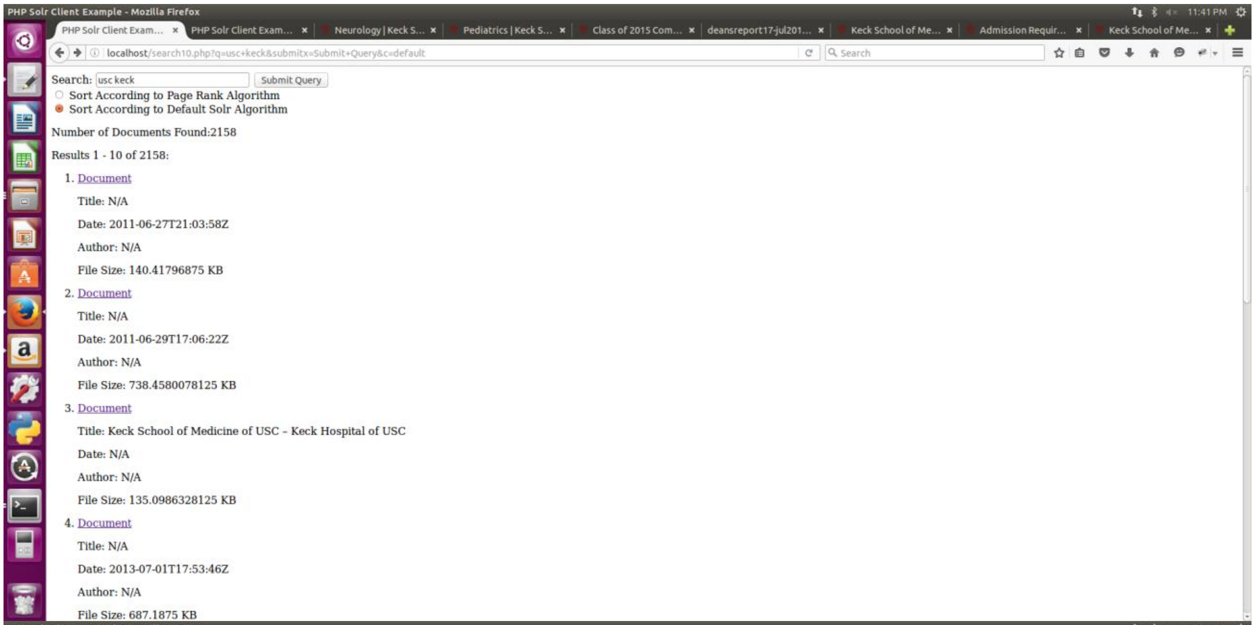
1. kck school of medisine -keck school of medicine
2. opthalmolgy -ophthalmology
3. neurogy- neurology
4. diseze -disease
5. electroic -electric

Examples of autocomplete words being handled correctly by my Autocomplete Program

1. characters me returns m, medicine ,medical, more and md
2. characters neu returns neurology, neuroimaging, neurobiology and neurogenetic
3. characters ra returns radiology, rankings and radiation
4. characters yo returns youtube ,youth, you, your and you're
5. characters vi returns visiting, viewport, vision and Vietnamese

Screenshots and doubts for graded queries presented in demonstration today :

1. Using USC keck did not give the homepage for usc keck school of medicine today in the top three results today because Solr computes the results based on tf-idf and cosine similarity scores which ranks the pages accordingly and we cannot override the default method to get the required pages.



Search: usc keck Submit Query
○ Sort According to Page Rank Algorithm
● Sort According to Default Solr Algorithm

Number of Documents Found:2158

Results 1 - 10 of 2158:

1. Document

   Title: N/A

   Date: 2011-06-27T21:03:58Z

   Author: N/A

   File Size: 140.41796875 KB

2. Document

   Title: N/A

   Date: 2011-06-29T17:06:22Z

   Author: N/A

   File Size: 738.4580078125 KB

3. Document

   Title: Keck School of Medicine of USC – Keck Hospital of USC

   Date: N/A

   Author: N/A

   File Size: 135.0986328125 KB

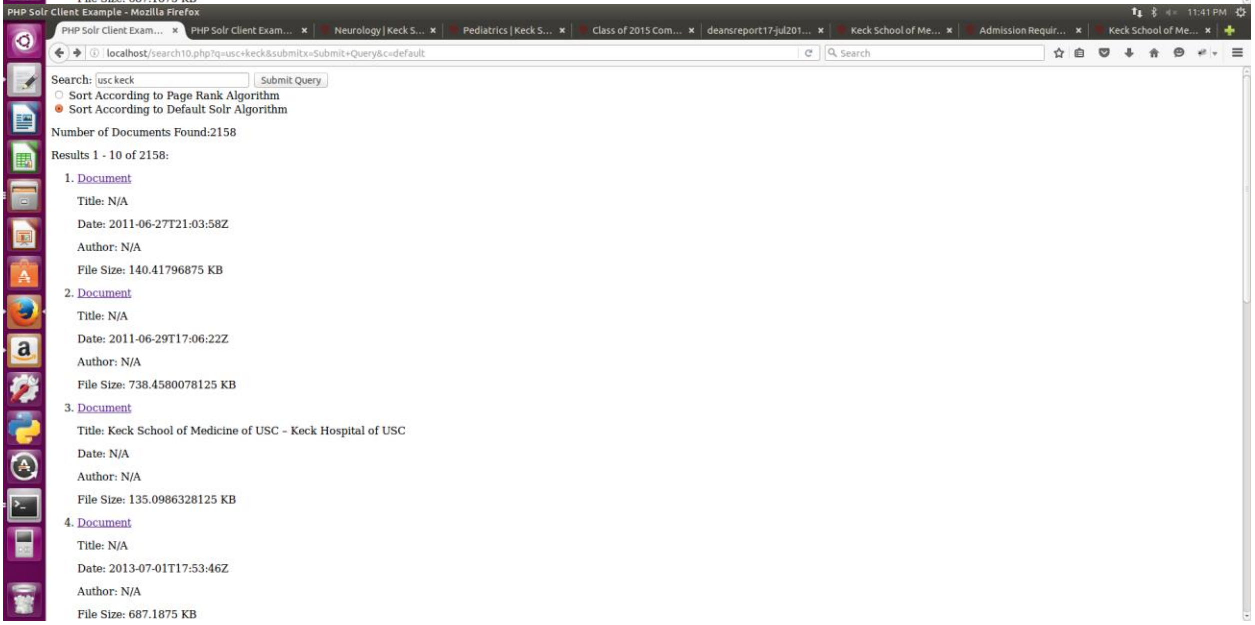
4. Document

   Title: N/A

   Date: 2013-07-01T17:53:46Z

   Author: N/A

   File Size: 687.1875 KB


PHP Solr Client Example - Mozilla Firefox

Search: usc keck Submit Query
○ Sort According to Page Rank Algorithm
● Sort According to Default Solr Algorithm

Number of Documents Found:2158

Results 1 - 10 of 2158:

1. Document

   Title: N/A

   Date: 2011-06-27T21:03:58Z

   Author: N/A

   File Size: 140.41796875 KB

2. Document

   Title: N/A

   Date: 2011-06-29T17:06:22Z

   Author: N/A

   File Size: 738.4580078125 KB

3. Document

   Title: Keck School of Medicine of USC – Keck Hospital of USC

   Date: N/A

   Author: N/A

   File Size: 135.0986328125 KB

4. Document

   Title: N/A

   Date: 2013-07-01T17:53:46Z

   Author: N/A

   File Size: 687.1875 KB

Screenshot for query USC keck

2. Also the misspelled word plascit suryerg will not return plastic surgery as the word is more than 2 edit distance misspelled from the correct word which exceeds the desired quality performance of Peter Norvig spell corrector program

3. Also the misspelled word allzimer disise will not return alzheimer disease as the word is more than 2 edit distance misspelled from the correct word which exceeds the desired quality performance of Peter Norvig spell corrector program

4. Also since I have used stemming I cannot get neurology in the first list of suggestions as I have used the connection to solr suggester and we don’t have full control over how the solr suggester works. We can only modify the solrconfig.xml file to make the solr Suggester give better suggestions .

   I have these issues which occurred when I gave the in class demonstration today and kindly look into these discrepancies and resolve the same while grading my submission.