Assignment-4

Steps followed

Spell Corrector

Firstly I downloaded "Peter – Norvig's" spellCorrector.php file from link given in the homework description http://www.phpclasses.org/package/4859-PHP-Suggest-corrected-spelling-text-in-pure-PHP.html#download and included the file within the Solr-Client.php file which was used for the graphical user interface of the search engine. Now whenever the query is typed within the query box, the words are exploded by space (tokenized) and then sent to the correct function of Peter-Norvig spellCorrector.php. After the words are corrected I stored it in a variable and then compared against the original query. If both of them were the same, the query was given as the input and the results were shown corresponding to the query else a "Did you mean:" statement was shown with the corrected query hyperlink. Clicking on the link shows the result for the corrected query.

Implementation of "big.txt" file

The source of the "Peter-Norvig's" spellCorrector.php was "big.txt" file. This big.txt file was made by parsing the content of the web pages, pdf's and word files which were downloaded during the crawling of the website. The files were parsed using Tika parser and the content was stored in big.txt.

Auto-Suggest

For the Auto-Suggest functionality, I first enabled the Auto-suggest feature in Solr by following the steps given in the assignment. After that I used AJAX call in collaboration with jQuery UI to call the solr suggest functionality enabled and that gave me suggestions for the first word typed. Later on to suggest for the other words typed after space I trimmed the query by taking the last index of the space and then passing to the suggestor the word after the space, by following the above method. This gave me suggestions for the other words as well.

Analysis:

Below is the analysis of the results for the Spell Corrector and Auto-Suggest:

1. Auto Suggest worked correctly for most of the queries:
   Example: "derma" gave the suggestions as: "dermatology", "dermal", "dermapathology", "dermatitis", "dermatologist"
   Example: "neu" gave the suggestions as: "neurology", "neuroimaging", "neurobiology", "neurogenetic", "neurorestoration

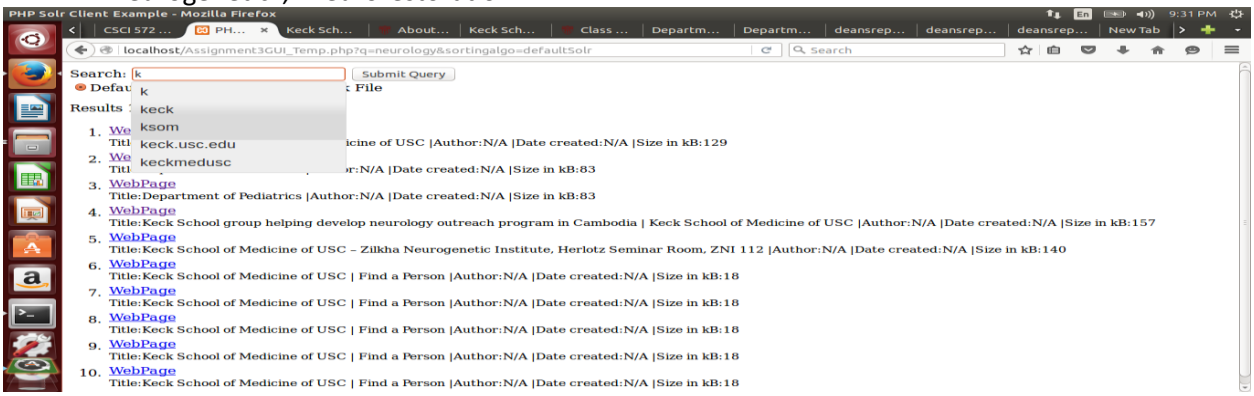

Fig 3: Auto Suggest for keck

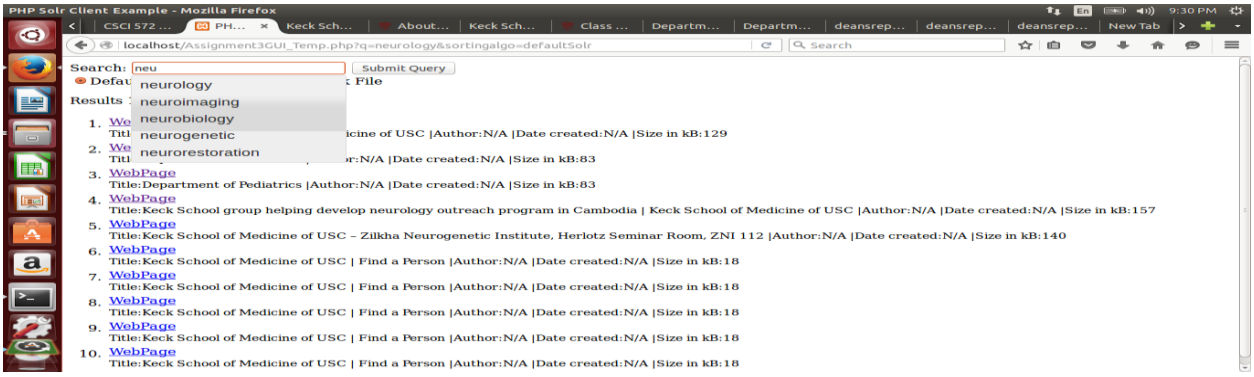

Fig 4: Auto Suggest for Neurology

So, as you can see the Auto suggest worked pretty well for all the queries. The auto suggestor in Solr is based on the term weights and are suggested by the Solr on the basis of the ordering of terms sorted in descending order by the weights.

Spell Corrector:

Spell corrector as mentioned above is incorporated in the code using SpellCorrector.php which is a Peter-Norvig code which works on edit distance of 2.

Below are the snapshots for the Spell Corrector

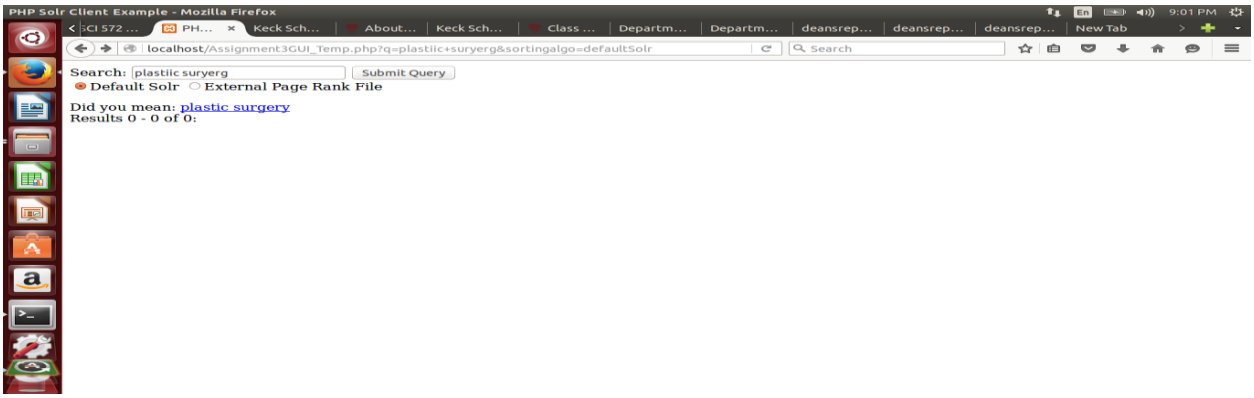

Fig 3: Spell Correction for “Plastiic suryerg”

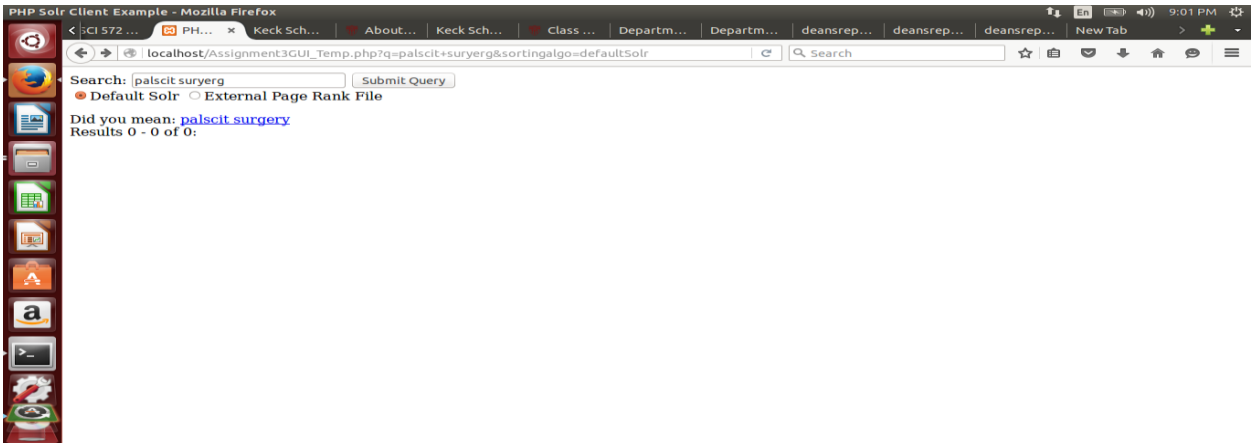

Fig 4: Spell Correction for “Palscit suryerg”

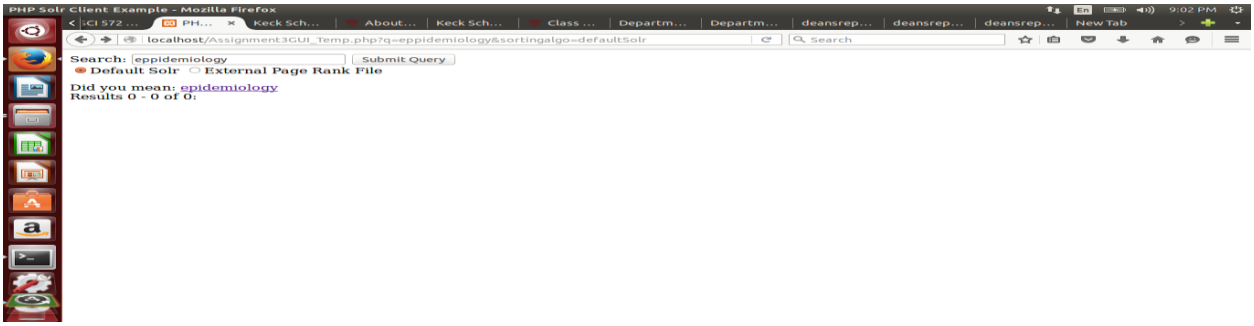

Fig 5: Spell Correction for “eppidemiology”

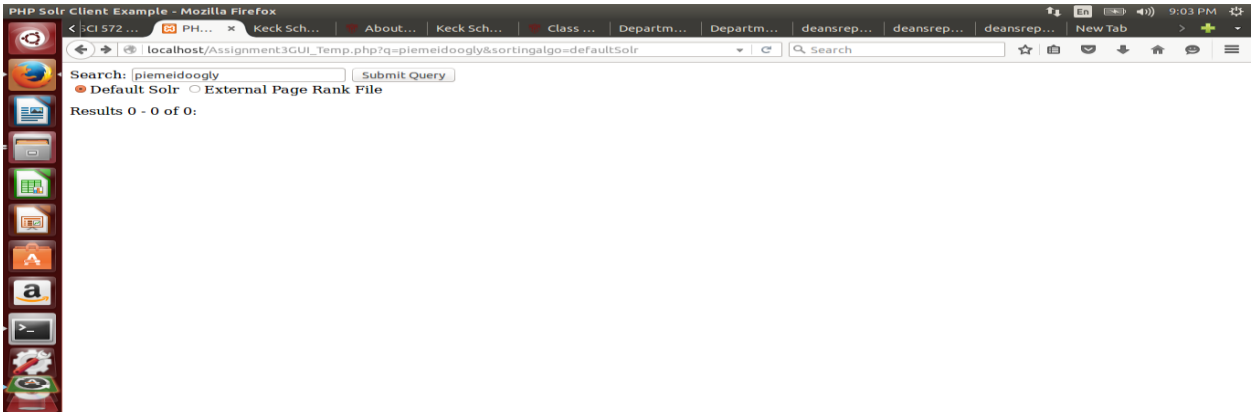

Fig 6: Spell Correction for “piemeidoogly”

So, as you can see, the Spell Correction worked for some of the queries and did not work for some of them. This is because the Peter Norvig algorithm works on the edit distance of 2 because of which “palscit” could not get changed to “plastic” but “plastic” did change to “plastic”. So there can be only two reasons for the spell correction not functioning as expected.

1. It works on the edit distance of two(limitation of the Algorithm)

2. The data is not present in the big.txt and hence got transformed to something which had an edit distance of two with the query entered.

Analysis for the relevant results not being in top 3:

When "usc keck" was searched as the query in the search Engine, the results that turned up in top 3 were all pdf's and the homepage was somewhere below top 5. This is because keck has around 43 occurrences in the pdf whereas if you see the homepage keck is mentioned only eight times throughout the page. So going by the working of Solr/Lucene which works on tf*idf weighting we can easily deduce that the term frequency of the term keck is so huge in the pdf's that it is outweighing the homepage where keck occurs just once and hence the discrepancy.

Links of both the pages are:

www.keck.usc.edu

www.keck.usc.edu/wp-content/uploads/sites/9/2015/02/deansreport17-jul2011.pdf