Radhika Hitesh Bhatt
USC ID: 6169382505

ASSIGNMENT 4

Assigned websites: LA Times and Huffington Post Websites

1] Link to youtube URL: http://www.youtube.com/watch?v=TOCTAypr3lM

The demo or the video uploaded has :

1) All the 8 correct spelling queries with auto suggestions and snippets displayed.

2) Next, spelling correction feature for 1 letter exchange is shown. The corrected spelling displayed is clicked to find its results.

3) Next, spelling correction feature for 2 letter exchange is shown. The corrected spelling displayed is clicked to find its results.

2] Steps followed to complete this assignment:

A] Creation of big.txt file:

I wrote a java program that iterates through all the files(previously given to us for homework3) using java program. I am extracting all the text between the tags using JSoup library and added the text to big.txt file.

B] Spell Correction:

i) I used Norvig's spell check for PHP version to implement spell correction.
ii) When a user inputs an invalid query and submits the page, I am checking if the query is a single word query or a multiple word query. If the query is a multiple word query I am splitting the word and passing each word to the Norvig's code.
iii) I am then displaying the results of the corrected words returned by Norvig's code as the suggested correct word.
iv) On the UI if the user enters the correct terms, the results are directly displayed to the user. If the user enters the wrong term, a phrase "Did you mean?" is displayed followed by the corrected word. On clicking the corrected word, the user is redirected to corresponding word's results.

C] Autocomplete:

i) I am using JQuery Autocomplete for implementing the auto compete feature.
ii) Using Ajax calls for each word entered, autocomplete will show relevant suggestions.
iii) For displaying the result I am showing results returned from FuzzyLookupFactory feature of Solr.
iv) I made changes in the solrconfig.xml file as given in the assignment description.
v) I implemented the logic such that for the first character that is entered by the user, 5 – 10 autocomplete suggestions will appear. For the second character that is entered, 3 – 7 autocomplete suggestions will appear. For the third and subsequent characters between 1 and 4 suggestions will appear.
vi) The PHP calls the following Solr URL:

http://localhost:8983/solr/myexample/suggest?indent=on&wt=json

Sample URL hit from my PHP code for the query term 'cali'

http://localhost:8983/solr/myexample/select?indent=on&q=cali&wt=json

vii) I have implemented logic to remove stop words from the auto-complete suggestions.

D] Snippet Creation:

i) I am using SimpleHTMLDom library to extract text data from the html files(excluding HTML tags). These files include just the result files(the files that has the query terms).
ii) Then I am parsing the text by splitting the text based on '.' .
iii) Further I am searching each of the the query terms in the extracted text from step (ii).
iv) If the query term is present in the sentence I am displaying the sentence as the snippet to the user.
v) If the word is not present as it is, that is, just the word preceded and followed by spaces and NOT as a part of another word only those sentences I am considering as a match.
vi) For the results that do not have snippets, I am displaying the title.

3] Analysis of the results:

In this I am providing FIVE examples of misspelled terms that are correctly handled by my spelling correction program also FIVE examples of auto-completion:

A] Sample of misspelled terms and their corrections provided by my code output:

Misspelled Terms	Corrected Terms
Poekmon Go	Pokemon Go
Doanld Trump	Donald Trump
Brzail	Brazil

Rio yolmpics	Rio Olympics
Caloifrnia Wild Fires	California Wild Fires

B] Sample of autocomplete for different characters entered:

Prefix	**Auto Completetions**
California Wild	California Wild California Wildlife California Wildly
riot	Riot Riots Rioting
Polit	Politics Political Politicians
Donald T	Donald Text Donald Type Donald Twitter
Ki	Kingdom Killed Kind King Kids