**HW 5 – Report**

**Name: Prerana Teligi Harapanahalli Math**

**USCID: 7795750397**

**1. Steps you followed to complete this assignment. Include the details of**

**what tools and techniques used to implement spelling correction and**

**autocomplete:**

Step 1: big.txt Generation:

• Extracted text from NBC News by crawling html data files using Apache Tika parser. Outputted the text to big.txt, which is later used to generate a vocabulary named dictionary.txt for spell check.

Step 2: Spell Checking:

• I used Peter Norvig’s spell check code (PHP version) to implement spell checking and integrated into a file called a SpellCorrector.php.

• When a misspelt or incorrect query is provided as input to the page, this

program checks to see if the word is valid or not. If the word is a valid word,

corresponding results are displayed otherwise phrase saying “Did you

mean?” followed by the suggested word is displayed. Upon clicking the new displayed keyword corresponding results are displayed.

• An incorrect query will be sent to the Norvig’s spell corrector and the correct word is returned as a suggestion for which dictionary.txt is used.

Step 3: Auto Complete:

• I followed the tutorial for adding auto complete feature in the search engine and defined Suggest component and search handler in solrconfig.xml file.

• FuzzyLookupFactory provided by Solr is used to suggest words for autocompleting a query.

• I Reloaded the core after updating the config file.

• I wrote an autocomplete script that hits the suggest component of solr whenever a new character is encountered in the search box which then triggers an ajax call to /suggest, whenever a character is either added or removed.

• Wrote a python program that uses Tika library for generating big.txt, that is used for the auto correction feature of the search engine.

Step 4: Snippet Generation:

• For every result, the snippet.php program searches for the query terms in the meta data first and if found, displays that as the snippet.

• If incase it fails to find relevant terms in meta data, entire file is searched to check if query terms exist in any sentence and the sentence it is shown as snippet. snippet.php and start.php are the files used for snippet generation.

Technologies used:

External Spell corrector code by Peter Norvig. It uses the edit distance algorithm to determine the closest word in dictionary (mostly words that are one or two edit distances away).

Auto Completion was built using internal Solr features such as FuzzyLookupFactory which uses suggest component for getting a list of suggestions based on the given term. The suggestions returned are based on an ajax call to query, and are displayed as a dropdown list in the query box.

Apache Tika was used for content extraction.

**2. Analysis of the results: In this you should provide examples of misspelled**

**terms that are correctly handled by your spelling correction program. You**

**should also provide some examples of auto completion.**

**SPELL CHECKING:**

|  |  |
| --- | --- |
| **Incorrect Word** | **Correct Word** |
| telsa | tesla |
| Aemriac | America |
| naas | nasa |
| Rusia | Russia |
| Donad Trmp | Donald Trump |

**AUTO COMPLETE:**

|  |  |
| --- | --- |
| **Prefix** | **Autocompletion** |
| Ca | Canonical  Careers  Can  Contact  California |
| immi | immigration  immediate  immediately  immigrants  imminent |
| Wh | Who  White  When  Which  With |
| Russia inves | Russia investigation  Russia investigators  Russia investigating  Russia investigates |
| Donald tr | Donald true  Donald trump  Donald triangle  Donald transparent  Donaldjtrumpjr |