HW-5 Report

Steps for AutoComplete:-

* Made changes to solr-config.xml file to enable TSTLookUpFactory

<searchComponent class="solr.SpellCheckComponent" name="suggest">

<lst name="spellchecker">

<str name="name">suggest</str>

<str name="classname">org.apache.solr.spelling.suggest.Suggester</str>

<str name="lookupImpl">org.apache.solr.spelling.suggest.tst.TSTLookupFactory</str>

<str name="field">\_text\_</str> <!-- the indexed field to derive suggestions from -->

<float name="threshold">0.005</float>

<str name="buildOnCommit">true</str>

</lst>

</searchComponent>

<requestHandler class="org.apache.solr.handler.component.SearchHandler" name="/suggest">

<lst name="defaults">

<str name="spellcheck">true</str>

<str name="spellcheck.dictionary">suggest</str>

<str name="spellcheck.onlyMorePopular">true</str>

<str name="spellcheck.count">5</str>

<str name="spellcheck.collate">true</str>

</lst>

<arr name="components">

<str>suggest</str>

</arr>

</requestHandler>

* Relevant solr query

<http://localhost:8983/solr/hwcore/suggest?indent=on&wt=json&q=”--query-string-->“;

Steps to enable Spellcheck:-

* Downloaded the javascript version of the library to ensure easy integration with NodeJS server
* <http://stoi.wordpress.com/2012/12/31/jspell/> (library by Shine Xavier)
* Warning:- There is a problem with the library where some of the corrected spellings will have the first letter removed:- Eg:- Berxit gives rexit instead or Brexit. (Tested with 2000+ occurences of Brexit in the file) Same with Russia. Wrong spelling of Russia returns Ussia.
* In server.js we train the data before the server starts accepting connection requests.
* We put NorvigSpellChecker into the server file. We read big.txt using the fs module and use nsc.train(Text) .
* nsc.correct(“abc”,”cde”,”efg”) will return the correct spellings in a json format

Steps for Snippet Generation:

* Snippet generation is done by a combination of two techniques
* The first half of the snippet is taken from the description returned by solr.
* The second half of the snippet is taken from the html file itself. Fetch the html file locally, extract data using the <https://github.com/werk85/node-html-to-text> library.
* Extract the line with first occurrence of the query word in it.

RESULTS:-

* Auto Complete

|  |  |
| --- | --- |
| ac | action  account\_id  according  across  account |
| bre | breadcrumb  break  breaking  breathing\_space  breakpoint |
| Po | politics  popular  policy  possible  position |
| fra | francisco  fragmentstart  frantically  frank  franchise |
| rest | restaurant  restaurants  restrictions  restoring |

* Spellcheck

|  |  |
| --- | --- |
| asalt | salt |
| asault | assault |
| Redice | reduce |
| iur | our |
| wrase | erase |

I have started a thread on piazza between the instructors detailing the problem with the javascript library written by Shine. For the more popular queries, spellcheck removes the first letter of the most probable answer and returns that to us.

“Berxit” returned “rexit”

“californan” returned “alifornian”

“Rusisa” returned “ussia”

The pattern follows. Words that have a high frequency count are returned without the first character.

Sample Snippet :- With Description and One sentence extracted from html

