Varnit Sah   vsah@usc.edu   9083888704

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.

Varnit Sah   vsah@usc.edu   9083888704

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.

Varnit Sah   vsah@usc.edu   9083888704

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

Title: Brexit Articles, Photos, and Videos - Los Angeles Times
Link: None
Desc: News, Photos and Information about Brexit | [/topic/politics-government/11000000-topic.html]Brexit
/home/vsah/Downloads/solr-6.5.0/hw4data/39a7634a-199b-45a7-8ec0-4cd68eb36a1c.html

Title: Brexit leads IMF to cut forecast for global economy - LA Times
Link: http://www.latimes.com/business/la-1468882806-snap-photo-photo.html
Desc: Pro-European Union supporters and pro-"Brexit" supporters demonstrate in London's Green Park on July 9. | Brexit leads IMF to cut forecast for global economy Brexit leads IMF to cut
/home/vsah/Downloads/solr-6.5.0/hw4data/ebd91d9c-1e87-4c3f-8e77-2b665c3824e3.html

Title: British scientists are freaking out about 'Brexit' too - LA Times
Link: http://www.latimes.com/science/sciencenow/la-sci-sn-brexit-science-20160624-snap-story.html
Desc: Scientists in the U.K. were overwhelmingly opposed to 'Brexit.' Now that it's happening, here's what they want politicians to keep in mind. | are concerned about how "Brexit" will affect their work with the ESA and other
/home/vsah/Downloads/solr-6.5.0/hw4data/78ebd410-efa6-4a45-a4c4-b5eaba56dfde.html

Title: 'Brexit' roils currency markets - LA Times
Link: http://www.latimes.com/business/la-1467824818-snap-photo-photo.html
Desc: The dollar looks to gain strength against the euro and pound in the wake of Britain's vote to leave the European Union. | 'Brexit' roils currency markets 'Brexit' roils currency markets Joseph De Sciose
/home/vsah/Downloads/solr-6.5.0/hw4data/19793351-63e0-4dea-8ced-1a56ea4c70f1.html

Title: Brexit - LA Times
Link: http://www.latimes.com/travel/deals/la-1467744419-snap-photo-photo.html
Desc: The vote to leave the European Union has weakened the British pound against the U.S. dollar, giving travelers more for their money. | Brexit Brexit Michael Kappeler/DPA/Zuma Press/TNS The vote to leave the European
/home/vsah/Downloads/solr-6.5.0/hw4data/e1d8349d-303b-4650-b62f-9da84d7a9498.html

Title: Brexit sign - LA Times
Link: http://www.latimes.com/world/la-brdelossantos-1485251826-snap-photo-photo.html
Desc: Britain's Supreme Court ruled Tuesday that the government must receive Parliament's approval to trigger the process of taking Britain out of the European Union. | Brexit sign Brexit sign Frank Augstein / Associated Press Britain's Supreme
/home/vsah/Downloads/solr-6.5.0/hw4data/bf11ea70-7d9a-458d-881b-48be6f3deb51.html

Title: Secretary of State John Kerry attempts to soothe 'Brexit' fears in Europe - LA Times
Link: http://www.latimes.com/world/la-fg-brexit-updates-secretary-of-state-john-kerry-attempts-1467127537-htmlstory.html
Desc: America's top diplomat sought to soothe fears Monday on both sides of the Atlantic as aftershocks of Great Britain's vote to withdraw f... | * Brexit
/home/vsah/Downloads/solr-6.5.0/hw4data/c4e94f55-e2e5-4704-af20-cda2bcfe8a1f.html