Spelling Checking ,Auto Suggestion and Snippet (NBC and WSJ)

Table of Contents :

Basic Steps followed for php application

- Installed Solr in local computer and run the sample core example to index the web pages.
- Start the solr server with **bin/solr start** and create a new core using command **bin/solr create –c IRWSJNBC**
- Change the configuration of "**managed-schema**" file as shown in instruction.
- Downloaded the data and merged data files for both NBC and WSJ dataset.
- Index the data set using command **bin/post –c / IRWSJNBC <file_Path>**
- Created a PHP application with Apache Server that will take user request and send request to local solr server with parameter such as request query and type of algorithm.
- The output from Solr will be parsed by PHP application and return to user.
- The browser output will return top 10 results with each entity have a external link, document ID , URL and size of the webpage.
- If Solr response don't return the Link, then PHP application will search local map.csv file for given document ID and return the external link.
- To support the solr server call, we have used solr php client which will enable call between solr and Apache Server(PHP) running locally.
- For Pagerank Algorithm, added external pageRank file to solr configuration to support the functionality.
- We will use edgeList.txt file to create a directed graph using **NetworkX graph** The output file contain pagerank for each html file and subsequently added to data folder of NewsSitesIndexing core.

- **Steps followed for Spell Check**
    - Utilized Peter Norvig spell correction program to spell correct the search query for WSJ and NBC dataset.
    - Peter Norvig program requires to build serialized dictionary of known words. For WSJ and NBC dataset, used t**ika to extract content and metadata** of each web page and generated Big.txt file .
    - Peter norvig program generates **serialized_dictionary.txt** file containing word-frequency for given big.txt file.
    - Furthermore, Peter Norvig program utilizes edit distance methods which will create variation of given search query with one or two edit distance.
    - For implementation in php client set **ini_set('memory_limit', '-1')** to avoid memory issue loading spell corrector files in memory.
    - For each term in query, we call **SpellCorrector::correct(q)** method to return the correct for given term and then concat all the returned term and send to solr for query and post the result back to browser.
    - To emulate google style functionality ,Every time a corrected query result is displayed on browser, it also shows the **original query in link format** to support querying the original term.

- **Steps followed for Auto Completion**
    - On php client we used ajax call to query solr for autosuggest dropdown using Jquery UI.
    - At Php client , used http://www.ranks.nl/stopwords stop word set to avoid suggesting any stop word.
    - At Php Client we used Stemmer.js along with application javascript to map suggestion to their root word.
    - At the server side, we added below code to managed-schema :

```
<field name="suggest_phrase" type="suggest_phrase"
indexed="true" stored="false" multiValued="false" />
```

    - And further added field definition

```
<fieldType name="suggest_phrase" class="solr.TextField"
positionIncrementGap="100">
<analyzer>
<tokenizer class="solr.KeywordTokenizerFactory" />
<filter class="solr.LowerCaseFilterFactory" />
</analyzer>
</fieldType>
```

- To support the functionality through spellcheck,we added spellCheck component under schema.xml

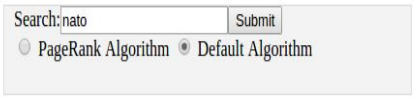
searchComponent name="suggest_phrase"
class="solr.SpellCheckComponent">
<lst name="spellchecker">
<str name="name">suggest_phrase</str>
<str name="classname">org.apache.solr.spelling.suggest.Suggester</str>
<str name="lookupImpl">org.apache.solr.spelling.suggest.fst.FSTLookup</str>
<str name="field">_text_</str>
<str name="buildOnCommit">true</str>
</lst>
</searchComponent>

- Utilized KeywordTokenizerFactory so that string get indexed as whole and spell check component for request handler. Used query of form **spellcheck.q = nato**

- **<u>Steps followed for Snippet Generation</u>**
  - To generate snippet, used PHP simple html dom parser and utilized function file_get_html()->plaintext to get the text from a given webpage.
  - Further set the file_get_html() parameters so it will preserve the delimiters such as " . "
  - Then split the result using explode function and checked if any sentence contain full query or partial and return the result accordingly.
  - Also added a list of stop word such as [Facebook, Google] which are generated by program and posted the result filtering the stop word.

Screenshot for NATO :

1. When misspelled as **“NAOT”**



Showing results for nato

Show results for naot

Results 1 - 10 of 416:

1. News Link
   Title : NATO Rejects Russian Air-Safety Proposal for Planes in Baltic Region - WSJ
   File Name: /home/lastfighter/Dropbox/USC/CSCI572/Assignment 3/data/defbe70d-a13d-4521-b5ec-f67fc5f08476.html
   Link Name : http://www.wsj.com/articles/nato-rejects-russian-air-safety-proposal-for-planes-in-baltic-region-1474391644
   Snippet : All allied planes flying missions for NATO—such as the Baltic air-policing mission—currently fly with their transponders on, meaning they emit an identifying signal in response to other radio signals...

2. News Link
   Title : Ukraine Crisis: U.S., Bulgaria to Begin NATO Military Drills - NBC News
   File Name: /home/lastfighter/Dropbox/USC/CSCI572/Assignment 3/data/3850cb5f-97f3-4169-ad0f-3c313bb2b631.html
   Link Name : http://www.nbcnews.com/storyline/ukraine-crisis/ukraine-crisis-u-s-bulgaria-begin-nato-military-drills-n322946
   Snippet : Bulgaria and other ex-communist countries in eastern Europe that are now inside NATO and the European Union have been rattled by Russia's annexation of Ukraine's Black Sea peninsula of Crimea and its support for separatists in eastern Ukraine...

2. Auto Suggestion for NATO :

Showing results for nato

Show results for naot

Results 1 - 10 of 416:

1. News Link
   Title : NATO Rejects Russian Air-Safety Proposal for Planes in Baltic Region - WSJ
   File Name: /home/lastfighter/Dropbox/USC/CSCI572/Assignment 3/data/defbe70d-a13d-4521-b5ec-f67fc5f08476.html
   Link Name : http://www.wsj.com/articles/nato-rejects-russian-air-safety-proposal-for-planes-in-baltic-region-1474391644
   Snippet : All allied planes flying missions for NATO—such as the Baltic air-policing mission—currently fly with their transponders on, meaning they emit an identifying signal in response to other radio signals...

2. News Link
   Title : Ukraine Crisis: U.S., Bulgaria to Begin NATO Military Drills - NBC News
   File Name: /home/lastfighter/Dropbox/USC/CSCI572/Assignment 3/data/3850cb5f-97f3-4169-ad0f-3c313bb2b631.html
   Link Name : http://www.nbcnews.com/storyline/ukraine-crisis/ukraine-crisis-u-s-bulgaria-begin-nato-military-drills-n322946
   Snippet : Bulgaria and other ex-communist countries in eastern Europe that are now inside NATO and the European Union have been rattled by Russia's annexation of Ukraine's Black Sea peninsula of Crimea and its support for separatists in eastern Ukraine...

All the other misspelling and autosuggest examples are placed under screenshot folder to keep the report 5 pages (As instructed)

- **Query Set:**

Query Name	One Interchange	Two Interchange
NATO	NOAT	NOAT
Dow Jones	Doj wones	Dow njoes
Rio Olympics	Rio Olympics	Rio yOlmpics
PokeMon Go	PoeKMon Go	PokoeMn Go
California Wild Fire	Cailfornia Wild Fire	Caloifrnia Wild Fire
Donald Trump	Doanld Trump	Dondal Trump
Harry Potter	Harry Potter	Harpry otter
Brazil	Brzail	Bralzi

- **Submission Files:**
  - ScreenShot folder :- contain screenshot for all 8 queries for both misspelling and auto suggestion.
  - Data Folder :- contain serialized dictionary generated by Spell Correction program
  - Code/ContentGenerator :- contain java program to create big.txt file
  - Code/PhpCode :- contains frontend javascript and phpclient file to request to solr for data
- **Youtube Video Link :https://youtu.be/8QXdbriY2qg**