Assignment 5 Report

Steps taken to complete the assignment:

1. Followed the tutorial to add auto complete feature in the search engine and defined Suggest component and search handler on the solrconfig.xml file the core.
2. For suggesting words to autocomplete query FuzzyLookupFactory provided by solr was used.
3. Reloaded the core after updating the config file.
4. Wrote an autocomplete jQuery script that hits the suggest component of solr whenever a new character is encountered in the querybox.
5. The script triggers an ajax call to …/suggest, whenever a character is either added or removed.
6. Wrote a python program that uses tika library for generating big.txt, that is used for the autocorrection feature of the search engine.
7. Used Norvig's Spell correction program that uses big.txt as a dictionary and edit distance algorithm for suggesting the next closest word in the dictionary if the user enters an invalid query.
8. Wrote a php program using simple_html_dom library for getting the snippet for the results.
9. For each result the snippet program searches for the query terms in the meta data first and if found, displays it as the snippet. If it fails to find relevant terms in meta data, the file is searched for the query terms and the sentence containing the relevant query terms is shown as snippet.

Tools used for Spelling correction mainly included the external Spell corrector code by peter norvig. It leverages the edit distance algorithm to determine the closest word in the dictionary that is one or two edit distance away. If the user searches for a wrong term, a line stating "Do you mean <suggested query term>" is shown to the user.

Auto Completion was built leveraging the internal Solr features such as FuzzyLookupFactory that uses the suggest component for getting a list of suggestions based on the given phrase. Once these suggestions are returned which is based on a ajax call to the query, it is displayed as a drop down on the query box.

Steps to execute :

1. Up and run apache server.
2. Run the solr on port 8983
3. A core with the name “myexample” should be created.
4. All the NBC news files should be indexed in “myexample” core.
5. Create big.txt file using the python code.
6. The file directory names should be updated according to the path of data
7. The generated big.txt and snippets folder should be placed in the solr/ directory which has the assignment5.php file
8. Place the simple_html_dom.php file in the same directory. Also place test2.php file in the same folder. This test2.php file contains the snippet program.
9. Make sure that peter norvig spell corrector program too is in the same directory.
10. On running assignment5.php in localhost, the screens mentioned in the screenshots section should appear.

Sample Misspelled words :

Input : trmp
Word Suggestion : trump

Input : telsa
Word Suggestion : Tesla

Input : north korae
Word Suggestion : north korea

Input : mistkea
Word Suggestion : mistake

Input : alppe
Word Suggestion : apple

Sample Auto Completion words:

Typed : apple
Auto Complete Suggestions : apple , application, appear, appears, appearance

Typed : Tesla
Auto Complete Suggestions : tesla , testament, tesla's, tessa, tesla’s

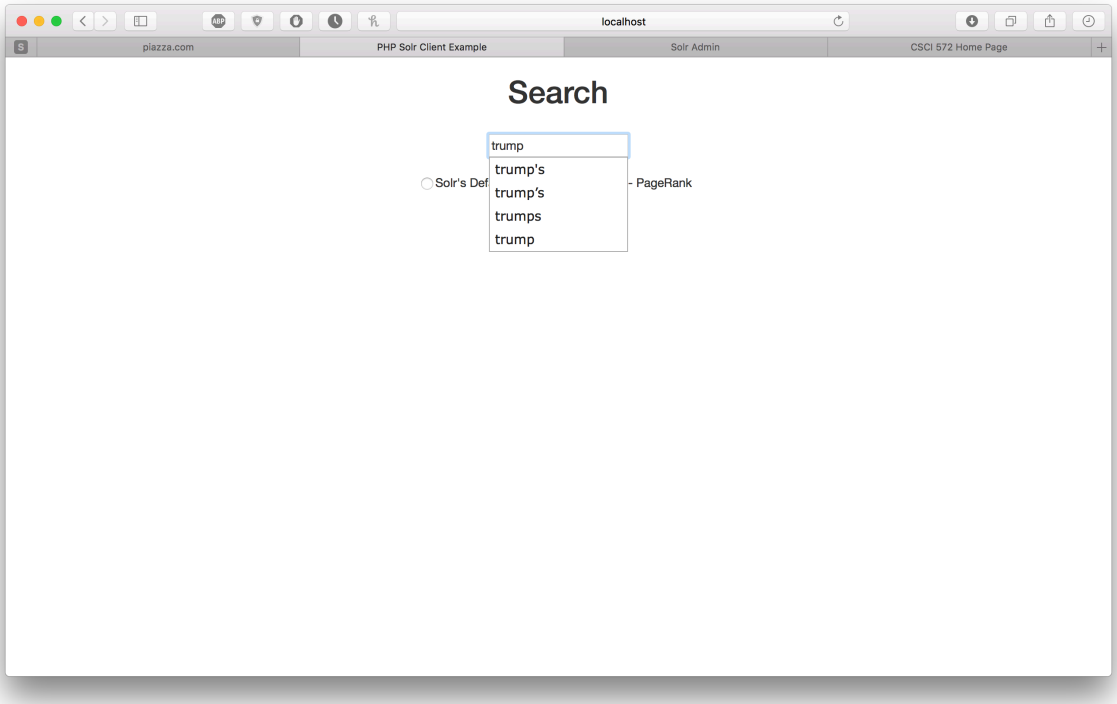
Typed : coder
Auto Complete Suggestions : coder, code, covering, coverimage, covered

Typed : smile
Auto Complete Suggestions : smile, silence, silent, smiles, smiling

Typed : grump
Auto Complete Suggestions : grumpy , grumbling, grumble, gumption, Grumman

Screenshots :

1) Showing Autocomplete



Search

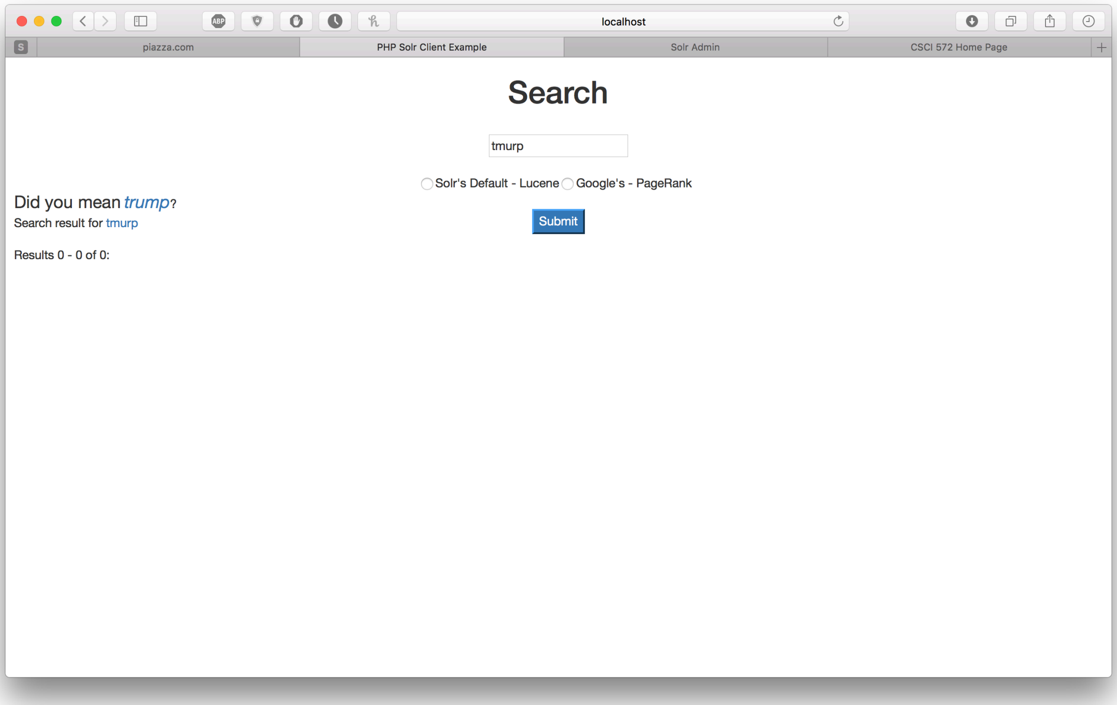
trump

○ Solr's Def

- PageRank

trump's
trump’s
trumps
trump

2) Screenshot showing autocorrect



Search

tmurp

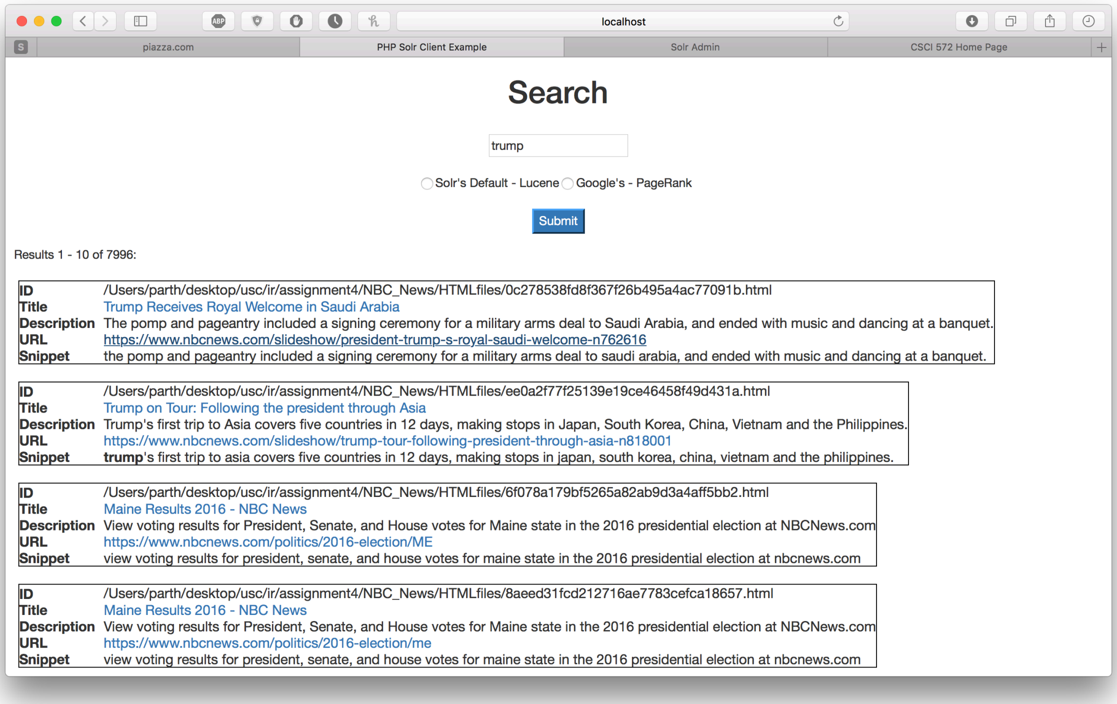
○ Solr's Default - Lucene ○ Google's - PageRank

Submit

Did you mean *trump*?
Search result for tmurp

Results 0 - 0 of 0:

3) Showing results after clicking the suggested spelling



Search

trump

○ Solr's Default - Lucene ○ Google's - PageRank

Submit

Results 1 - 10 of 7996:

ID	/Users/parth/desktop/usc/ir/assignment4/NBC_News/HTMLfiles/0c278538fd8f367f26b495a4ac77091b.html
Title	Trump Receives Royal Welcome in Saudi Arabia
Description	The pomp and pageantry included a signing ceremony for a military arms deal to Saudi Arabia, and ended with music and dancing at a banquet.
URL	https://www.nbcnews.com/slideshow/president-trump-s-royal-saudi-welcome-n762616
Snippet	the pomp and pageantry included a signing ceremony for a military arms deal to saudi arabia, and ended with music and dancing at a banquet.

ID	/Users/parth/desktop/usc/ir/assignment4/NBC_News/HTMLfiles/ee0a2f77f25139e19ce46458f49d431a.html
Title	Trump on Tour: Following the president through Asia
Description	Trump's first trip to Asia covers five countries in 12 days, making stops in Japan, South Korea, China, Vietnam and the Philippines.
URL	https://www.nbcnews.com/slideshow/trump-tour-following-president-through-asia-n818001
Snippet	**trump**'s first trip to asia covers five countries in 12 days, making stops in japan, south korea, china, vietnam and the philippines.

ID	/Users/parth/desktop/usc/ir/assignment4/NBC_News/HTMLfiles/6f078a179bf5265a82ab9d3a4aff5bb2.html
Title	Maine Results 2016 - NBC News
Description	View voting results for President, Senate, and House votes for Maine state in the 2016 presidential election at NBCNews.com
URL	https://www.nbcnews.com/politics/2016-election/ME
Snippet	view voting results for president, senate, and house votes for maine state in the 2016 presidential election at nbcnews.com

ID	/Users/parth/desktop/usc/ir/assignment4/NBC_News/HTMLfiles/8aeed31fcd212716ae7783cefca18657.html
Title	Maine Results 2016 - NBC News
Description	View voting results for President, Senate, and House votes for Maine state in the 2016 presidential election at NBCNews.com
URL	https://www.nbcnews.com/politics/2016-election/me
Snippet	view voting results for president, senate, and house votes for maine state in the 2016 presidential election at nbcnews.com