Name: Shobhit Agarwal

USC ID: 6473476393

CSCI 572 Homework 5

News Set: NBC News

1. Steps you followed to complete this assignment. Include the details of what tools and techniques you used to implement spelling correction and autocomplete.

Steps followed in the homework:

- Carried out the entire assignment on a MAC machine running OS X
- Continued on the PHP script from HW4

Generating Big.txt - GenerateDictionary.java:

- First I wrote a JAVA program (GenerateDictionary.java) using Apache TIKA jar to get all the distinct words from all the downloaded HTML pages for the news website set.
- Once all the keywords were extracted, I saved them in a text file 'big.txt' to be used as a dictionary by Peter Norvig's spelling correction program

Using Peter Norvig's Spelling correction PHP code - SpellCorrector.php:

- Downloaded Norvig's spelling correction PHP code, and used the SpellCorrector:correct() function to get the spelling corrections.
- The PHP code internally created a serialized_dictionary.txt file to be used for spell correction from the big.txt file.
- Made the required changes in my PHP script to mimic Google's behaviour of showing results for the correct spelling and offering a link to search for the misspelled word as well.
- The results are displayed when the user writes a query string and hits submit.

Autocomplete feature:

- Next, I used JQuery autocomplete feature to get the 5 word suggestions.
- Suggestions are made word by word, where words are defined as a continuous array of string after a space or new word.
- Made changes in Solr to add a suggest function that would give 5 suggestions for the current typed in phrase.
- To avoid issues with CORS, wrote another PHP script that would instead call the Solr suggest function to get the suggestions.
- The results are displayed as soon as 1 letter is typed in. URL used: http://localhost/suggestions.php?q=califo

Which in turn calls http://localhost:8983/solr/hw4/suggest?wt=json&indent=true&q=califo

- The results are displayed in the jquery autosuggest bar and on click the highlighted word is selected and on submit it is sent to the server.

Snippets:

- Lastly, I worked on getting snippets from the downloaded webpages.
- The order of snippets is to first try to look for query words in meta tags, if found use it, otherwise look for the first string with query words in the body of the HTML. If even then no such string is found, 'No Snippet Available' is shown, otherwise the first sentence with query words is shown as snippet.

2. Analysis of the results: In this you should provide FIVE examples of misspelled terms that are correctly handled by your spelling correction program. You should also provide FIVE examples of auto-completion.

Examples of Spelling errors:

Misspelled word	Corrected word
donad trup	donald trump
snapcht	snapchat
nsdaq	Nasdaq
illegl immigrition	illegal immigration
califrna	california

Examples of Auto Suggest:

Prefix	Suggestions
califo	california calif californians california's californias californian
donal	donal donald donate donations donating donated
ille	illegal ill illegally illinois illness illegals

immigr	immigr immigration immigrants immigrant immigrated immigrate
snpc	snapchat snack snacks snuck snickers snacking