Homework 5

Name: Peng Zhang

USC-ID: 7292183243

Description:

1. Create “big.txt”

Created Java program using Tika parser package to extract text from Boston Globe crawled data ser. And put the text into a file called “big.txt”. This file is used for the vocabulary of spell check and auto-completion.

1. Spell Correction

Used Norvig’s spell check php version downloaded from internet to implement the spell correction.

When a user inputs a query, the program checks I there a word or vocabulary of the same inside the corpus(big.txt). If there is a valid word, then doing nothing. Otherwise, the program considers other similar words that is closest to the query entered. Then display the words after “Do you mean:”. Also add it as a clickable href for searching the corrected words in the server.

1. Auto Complete

Called the autocomplete from main page and send every character individually to implement it.

Call <http://localhost:8983/solr/boston_globe/suggest?wt=json&indent=on&q=> and send it back to the home page.

The result of the auto complete is displayed in a dropdown list from the entering space, user can select the corresponding results that is displayed.

1. Snippet

Added the simple\_dom\_html php file for reading the content of html.

Break the content to sentence array by “.” and break the query to words group.

Match the words in the words group with the sentences, and store it to the snippet.

Analysis:

Auto-Correct:

|  |  |
| --- | --- |
| Incorrect Word | Correct Word |
| chian | china |
| medcine | medicine |
| hilry cliton | hilary clinton |
| gogle phtoo | google photo |
| mechine leaning | machine learning |

Auto-Completion:

|  |  |
| --- | --- |
| Incorrect Word | Correct Word |
| yell | yellow, yelled, yelling, yelp, year’s |
| machi | machine, machado, marching, michigan, years |
| immi | Immediately, immigration, immediate, immigrant, gallary |
| carol | carolina, caroline, carl, cars, crossword |
| cre | center, career, cent, crossword, credit |