Andrew Risse

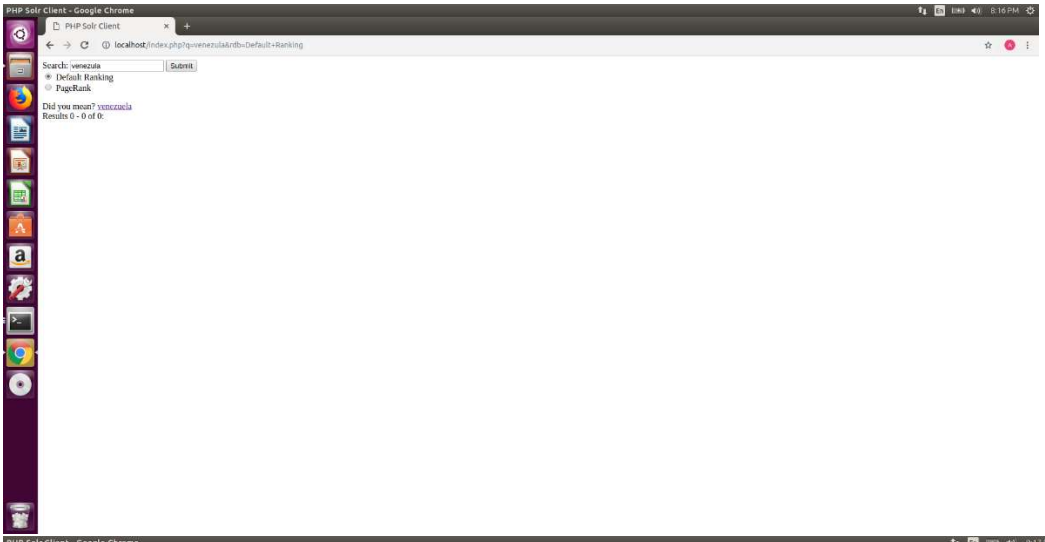
USC ID: 5987029524

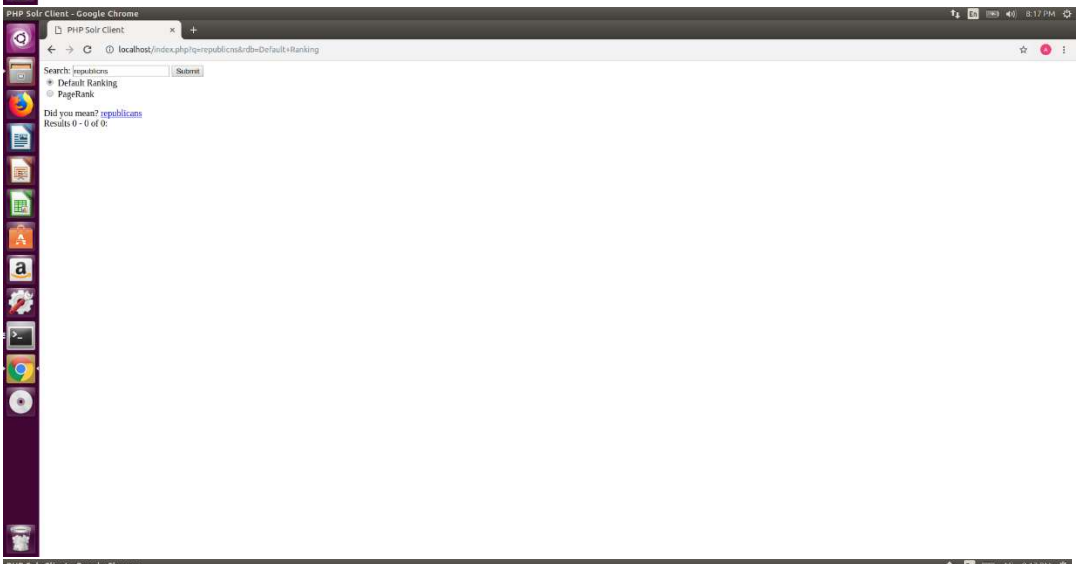
HW 5 Report

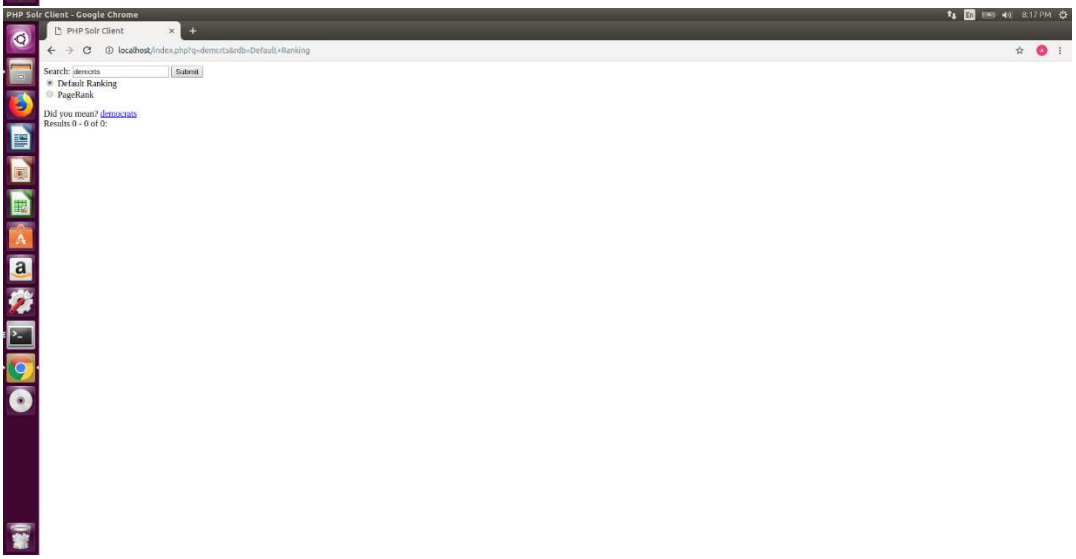
These are the steps I used to complete this project:

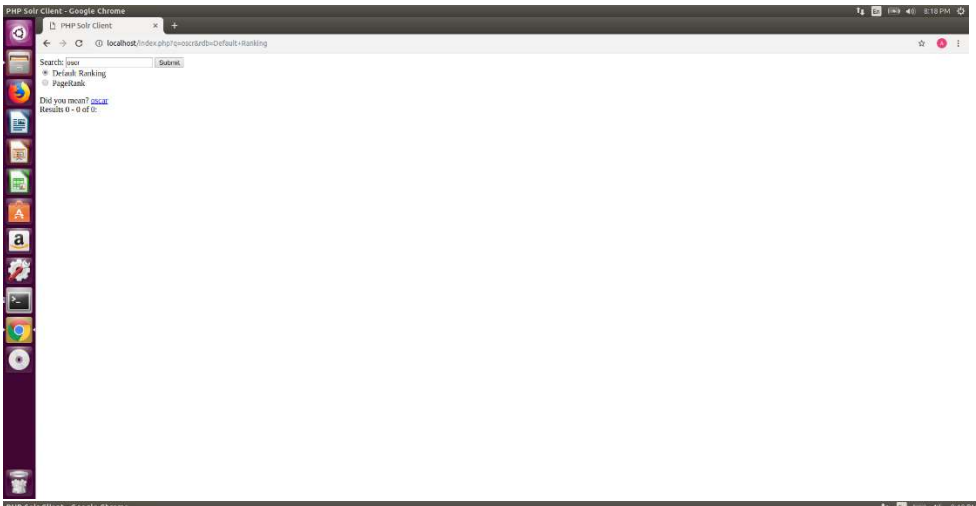
1. Write a script (tikascript.sh) that uses the command:
   java -jar tika-app-1.20.jar --text-main "$file" >> big.txt
   This script executed the command for all the “Yahoo” html files and put it’s results in “big.txt” It took several hours for this to complete.
2. I then followed the instructions for adding SOLR autocomplete. I wrote some PHP code that uses JQuery.ajax() to submit a /suggest query to my core and respond with auto-suggestions in JSON format. This format had to be parsed so only the terms were extracted and displayed in a drop down. I used this website as a reference: https://examples.javacodegeeks.com/enterprise-java/apache-solr/solr-autocomplete-example/
   My version does not handle more than two words for the query when provided auto-complete, I ran out of time to continue messing with the solrconfig.xml.
3. I then tried to use my big.txt file with the PHP version of Norvig SpellCorrector. I was able to get it to spell correct with the Norivg big.txt file, but it would not work with mine. I tried everything recommended on the forums (permissions, looking for a serialized_dictionary.txt file, adding code to handle memory issues, etc) and nothing would work. I elected to add the query terms from HW4 to the beginning of the Norvig.txt file and move on. I had spent way too many hours troubleshooting. I decided to implement the spell correction in my HTML after the user submits a query so that any misspellings appear as a clickable link after the words “Did you mean?” on the results page.
4. The final step was to write PHP code that generates snippets (snippet.php). I tried using Tika to parse the HTML for text, but this caused snippet generation to be very slow even though the results were better. I ended up using file_get_contents and stripping the tags. I had to apply preg_replace several times with different regular expressions to clean up the output. There are still some issues with the snippets that are produced (leftover HTML stuff), but I ran out of time to perfect the regular expressions.

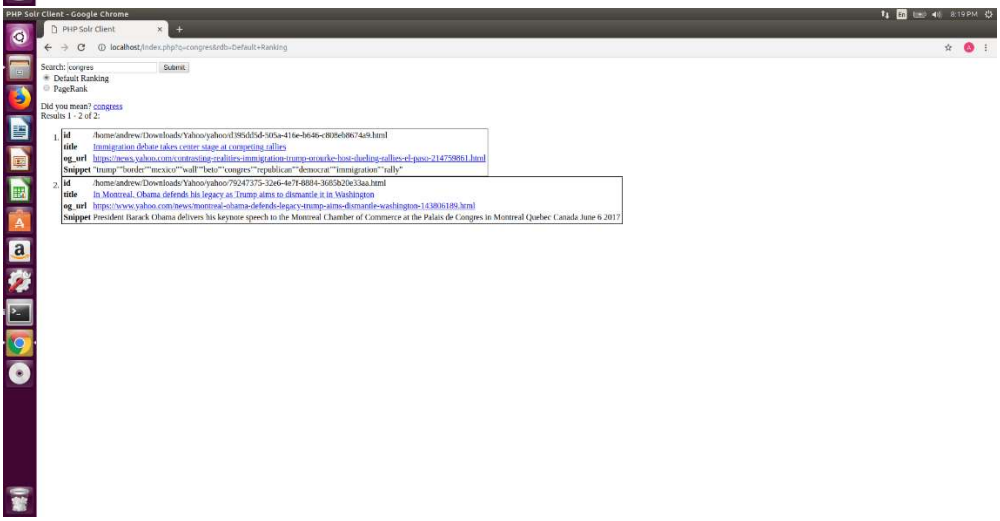
Screenshots of spell correct:



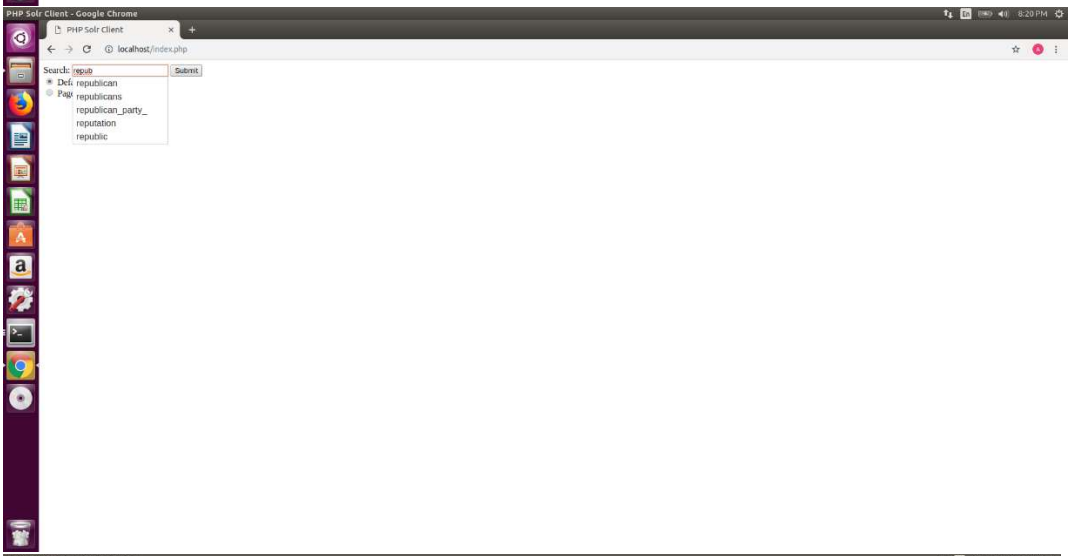


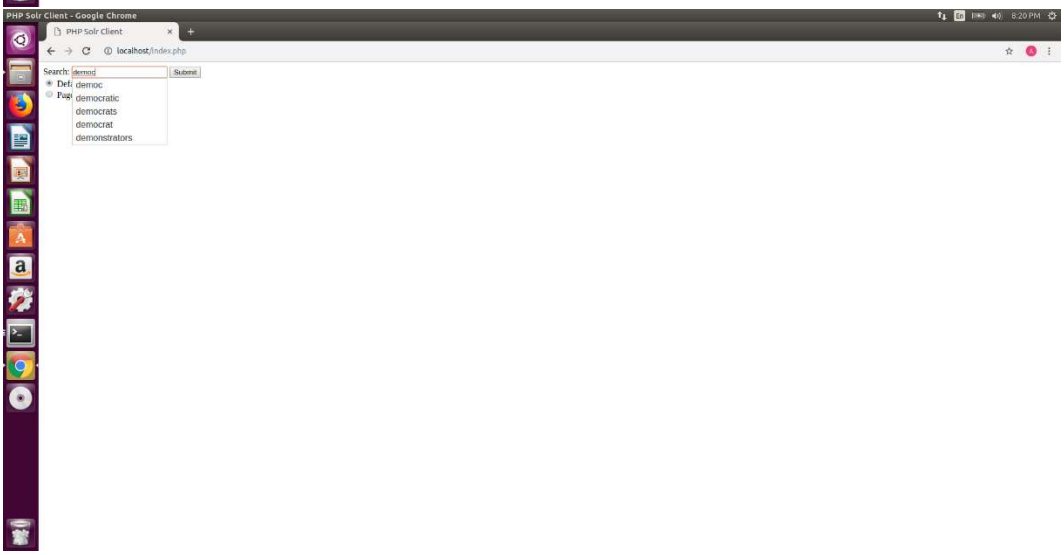


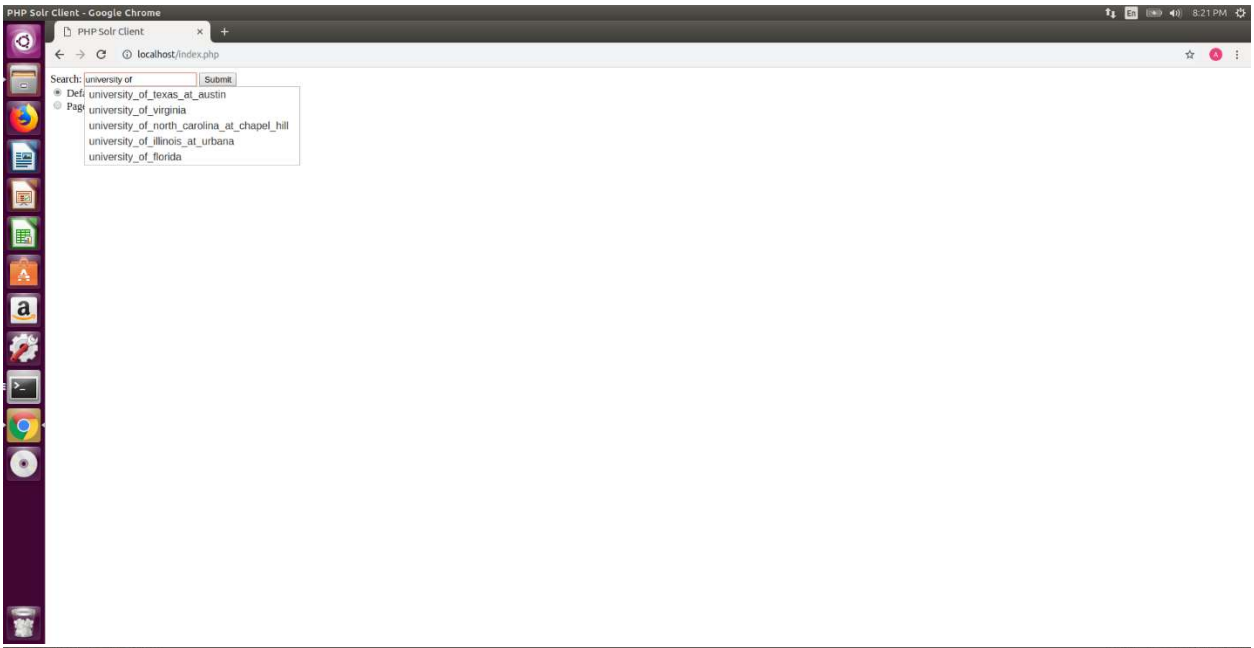

PHP Solr Client - Google Chrome
8:18 PM
PHP Solr Client
localhost/index.php?q=oscr&rdb=Default+Ranking
Search: oscr
Submit
Default Ranking
PageRank
Did you mean? oscar
Results 0 - 0 of 0:

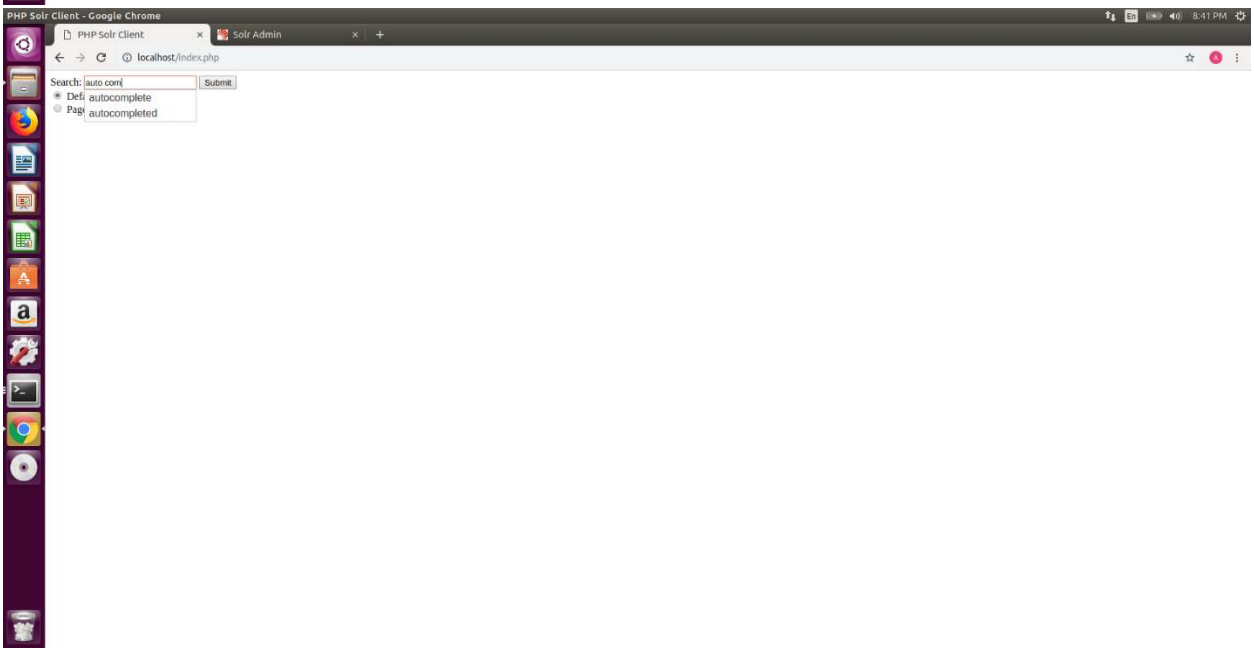

PHP Solr Client - Google Chrome
8:19 PM
PHP Solr Client
localhost/index.php?q=congres&rdb=Default+Ranking
Search: congres
Submit
Default Ranking
PageRank
Did you mean? congress
Results 1 - 2 of 2:
1. id /home/andrew/Downloads/Yahoo/yahoo/1985dd5d-505a-416e-b646-c805eb8674e9.html
title Immigration debate takes center stage at competing rallies
og_url https://news.yahoo.com/contrasting-realities-immigration-trump-orourke-host-dueling-rallies-el-paso-214759861.html
Snippet "trump""border""mexico""wall""beto""congres""republican""democrat""immigration""rally"
2. id /home/andrew/Downloads/Yahoo/yahoo/7924737 5-32e6-4e7f-8884-3685b20e33aa.html
title In Montreal, Obama defends his legacy as Trump aims to dismantle it in Washington
og_url https://www.yahoo.com/news/montreal-obama-defends-legacy-trump-aims-dismantle-washington-143806189.html
Snippet President Barack Obama delivers his keynote speech to the Montreal Chamber of Commerce at the Palais de Congres in Montreal Quebec Canada June 6 2017

Screenshots of auto-complete:






PHP Solr Client - Google Chrome
8:21 PM
PHP Solr Client
localhost/index.php
Search: university of
Submit
university_of_texas_at_austin
university_of_virginia
university_of_north_carolina_at_chapel_hill
university_of_illinois_at_urbana
university_of_florida


PHP Solr Client - Google Chrome
8:41 PM
PHP Solr Client
Solr Admin
localhost/index.php
Search: auto com
Submit
autocomplete
autocompleted