

Report of CSCI 572 Assignment 5

1. Process of finishing the assignment

1) All the steps are based on accomplishment of assignment 4;

2) Add autocomplete suggest function in Solr

➔ Add a search component to solrconfig.xml

➔ Add the request handler in solrconfig.xml

```
856 <searchComponent class="solr.SuggestComponent" name="suggest">
857   <lst name="suggester">
858     <str name="name">suggest</str>
859     <str name="lookupImpl">FuzzyLookupFactory</str>
860     <str name="field">_text_</str>
861     <str name="suggestAnalyzerFieldType">string</str>
862   </lst>
863 </searchComponent>
864 <requestHandler class="solr.SearchHandler" name="/suggest">
865   <lst name="defaults">
866     <str name="suggest">>true</str>
867     <str name="suggest.count">10</str>
868     <str name="suggest.dictionary">suggest</str>
869   </lst>
870   <arr name="components">
871     <str>suggest</str>
872   </arr>
873 </requestHandler>
```

➔ Test in the Solr UI to make sure I can get json results after input a query string.

3) Prepare for spelling correction function.

➔ Download the PHP version of Norvig's spelling corrector.

➔ Use Apache Tika to create "big.txt" as the text file. (Download jar of Tika and create project for to parse content of all WSJ html files.)

At the same time, I created a parsed text file for each html file for later snippet creation.

```
16 Tika tika = new Tika();
17 String outputPath = "/Users/WeiLi/Downloads/solr-7.1.0/WSJ/plaintext/";
18 int i = 0;
19 for(File file : dir.listFiles()) {
20   System.out.println(++i + "finished.");
21   String filename = file.getName();
22   if(filename.equals(".DS_Store"))
23     continue;
24   String text = "";
25   try {
26     text = tika.parseToString(file);
27   } catch (TikaException te) {
28     te.printStackTrace();
29   }
30
31   String[] tokens = text.trim().split("\\s+");
32   System.out.println(tokens.length);
33   for(String token : tokens) {
34     dicWriter.write(token + " ");
35   }
36   filename = outputPath + filename.substring(0, filename.lastIndexOf(".html")) + ".txt";
37   FileOutputStream fileOut = new FileOutputStream(filename);
38   OutputStreamWriter writer = new OutputStreamWriter(fileOut, "UTF-8");
39   writer.write(text);
40   writer.flush();
41   writer.close();
42 }
43
44 dicWriter.flush();
45 dicWriter.close();
```


4) Add all these components to previous PHP file of assignment 4.

For autocomplete, use ajax to send query to Solr server and get json results. Parse json results and override the autocomplete function of input textbox. For CORS problem happened in the process, use jsonp to fetch the response.

For **spell correction**, include SpellCorrector.php to the directory and call correct() function once user submit the query. Compare returned result with original query to determine if there are any misspell words in original query. If yes, display a not like “Showing results for <corrected query> with a clickable link.

For **snippet**, when I get the top 10 results from Solr and check each corresponding txt file (which I generated by Tika before), find first position of the whole query occurs, if not, just find single terms in the query and return several sentences following that. Then call shoeSnippet() function to make each term in the query bold when they display in the result page.

2. Results of the page(Snippet)

 **USC** Search Hurricane Harvey ☒ Solr Lucene ☐ PageRank

Results 1 - 10 of 426:

[Hurricane Harvey Slams Texas With Devastating Force - WSJ](https://www.wsj.com/articles/harvey-slams-texas-with-devastating-force-1503750047)
<https://www.wsj.com/articles/harvey-slams-texas-with-devastating-force-1503750047>
Hurricane Harvey slammed into Texas as a powerful Category 4 **hurricane**, with intense rain and winds of more than 100 miles an hour as it struck land. It had weakened to a Category 1 storm by 5 a.m. and to a tropical storm by mid-afternoon.

[Hurricane Harvey: One Week After Landfall - WSJ](https://www.wsj.com/articles/hurricane-harvey-one-week-after-landfall-1504390092)
<https://www.wsj.com/articles/hurricane-harvey-one-week-after-landfall-1504390092>
Hurricane Harvey: One Week After Landfall One week after **Harvey** blasted into southeast Texas as a Category 4 **hurricane**, some residents have returned to their homes to clean up while rescuers search fo...

[Hurricane Harvey Threatens Largest Flood Insurer: The Government - WSJ](https://www.wsj.com/articles/hurricane-harvey-threatens-largest-flood-insurer-1503771686)
<https://www.wsj.com/articles/hurricane-harvey-threatens-largest-flood-insurer-1503771686>
Hurricane Harvey could inundate the National Flood Insurance Program with billions in new claims shortly before the plan is scheduled to expire on Sept. 30 with just \$5.8 billion left it can borrow from the Treasury to meet obligations.

[Houston's Environmental Threats Come Into Focus - WSJ](https://www.wsj.com/articles/houstons-environmental-threats-come-into-focus-1504554072)
<https://www.wsj.com/articles/houstons-environmental-threats-come-into-focus-1504554072>
Hurricane Harvey or leave for good? Video: Jake Nicol. Photo: Annie Mulligan for The Wall Street Journal Floodwaters also have inundated at least five toxic-waste Superfund sites near Houston, and som...

[Small Businesses Say Federal-Disaster Aid Needs Strengthening - WSJ](https://www.wsj.com/articles/small-businesses-say-federal-disaster-aid-needs-strengthening-1507667045)
<https://www.wsj.com/articles/small-businesses-say-federal-disaster-aid-needs-strengthening-1507667045>
Hurricane Harvey this summer swamped the office he has occupied for nearly 25 years, soaking chiropractic tables and X-ray and ultrasound machines. He said he is reluctant to take on debt and plans in...

[Insuring Coastal Cities Against the Next Hurricane Harvey - WSJ](https://www.wsj.com/articles/insuring-coastal-cities-against-the-next-hurricane-harvey-1504282974)
<https://www.wsj.com/articles/insuring-coastal-cities-against-the-next-hurricane-harvey-1504282974>
Hurricane Harvey Americans don't like being forced to buy flood coverage, but it may be the best way to protect the booming economies most at risk People float belongings out of their flooded neighb...

[Hurricane Harvey Slams Texas With Devastating Force - WSJ](https://www.wsj.com/articles/harvey-slams-texas-with-devastating-force-1503750047?tesla=y)
<https://www.wsj.com/articles/harvey-slams-texas-with-devastating-force-1503750047?tesla=y>
Hurricane Harvey slammed into Texas as a powerful Category 4 **hurricane**, with intense rain and winds of more than 100 miles an hour as it struck land. It had weakened to a Category 1 storm by 5 a.m. and to a tropical storm by mid-afternoon.

[Lower U.S. Oil Prices Are a Shot in the Arm for Crude Exports - WSJ](https://www.wsj.com/articles/lower-u-s-oil-prices-are-a-shot-in-the-arm-for-crude-exports-1505986208)
<https://www.wsj.com/articles/lower-u-s-oil-prices-are-a-shot-in-the-arm-for-crude-exports-1505986208>
Hurricane Harvey has passed, but analysts expect the storm's effects on global crude flows to linger for months Tug boats towed the Hess Corp. Stampede tension leg oil platform past Port Aransas, Te...

[News Article Archive from Sept 06, 2017 - Wsj.com](http://www.wsj.com/public/page/archive-2017-9-06.html)
<http://www.wsj.com/public/page/archive-2017-9-06.html>
Hurricane Harvey aid bill, just hours after House Speaker Paul Ryan sharply criticized the plan. Obamacare Insurer in Virginia to Scale Back Planned Expansion Obamacare Insurer in Virginia to Reduce P...

[Progressive Reports Hurricane Harvey Linked to 90% of Total Catastrophe Losses in August - WSJ](https://www.wsj.com/articles/progressive-reports-hurricane-harvey-linked-to-90-of-total-catastrophe-losses-in-august-1505831379)
<https://www.wsj.com/articles/progressive-reports-hurricane-harvey-linked-to-90-of-total-catastrophe-losses-in-august-1505831379>
Progressive Corp. swung to a loss in August, the company reported Tuesday, with **Hurricane Harvey** more than tripling the insurer's catastrophe losses in the period compared to last year.

3. Results of the 5 spelling corrections

Query Entered	Spelling Correction
gradute	graduate
Southen Califona	Southern California
Snpachat	Snapchat
Olimpic Recods	Olympic Records
Compter Science	Computer Science

1) gradute – graduate

 **USC** Search ☒ Solr Lucene ☐ PageRank

Showing results for [graduate](#)
Results 0 - 0 of 0:

2) Southen Califona – Southern California

 **USC** Search ☒ Solr Lucene ☐ PageRank

Showing results for [southern california](#)
Results 0 - 0 of 0:

3) Snpachat – Snapchat

 **USC** Search ☒ Solr Lucene ☐ PageRank

Showing results for [snapchat](#)
Results 0 - 0 of 0:

4) Olimpic Recods – Olympic Records

 **USC** Search ☒ Solr Lucene ☐ PageRank

Showing results for [olympic records](#)
Results 0 - 0 of 0:

5) Compter Science – Computer Science

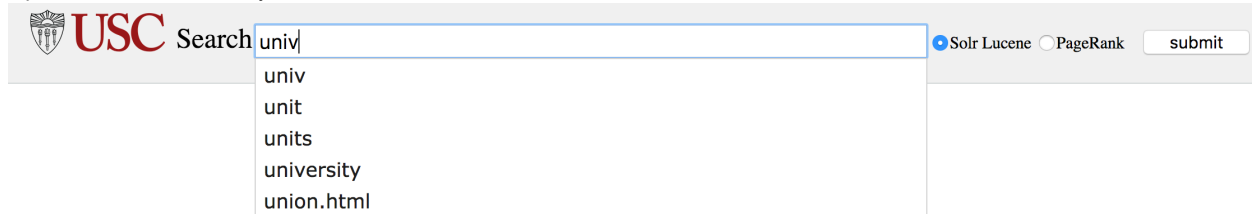
 **USC** Search ☒ Solr Lucene ☐ PageRank

Showing results for [computer science](#)
Results 0 - 0 of 0:

4. Results of the 5 autocomplete suggestions

Query Entered	Expected Terms
univ	university
cali	california
go	google
ent	entertainment
algor	algorithms

1) univ – university

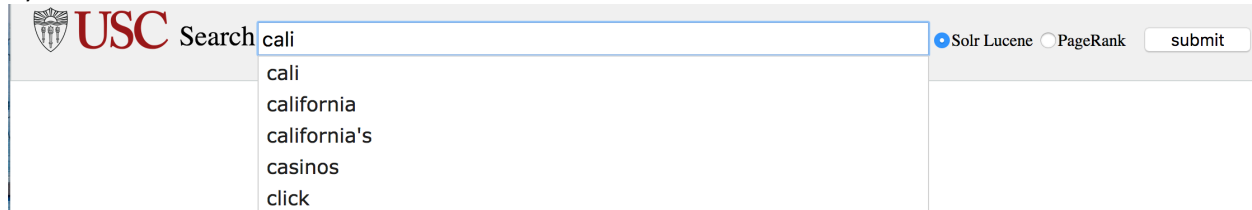


USC Search univ

- univ
- unit
- units
- university
- union.html

☒ Solr Lucene ☐ PageRank

2) cali – california

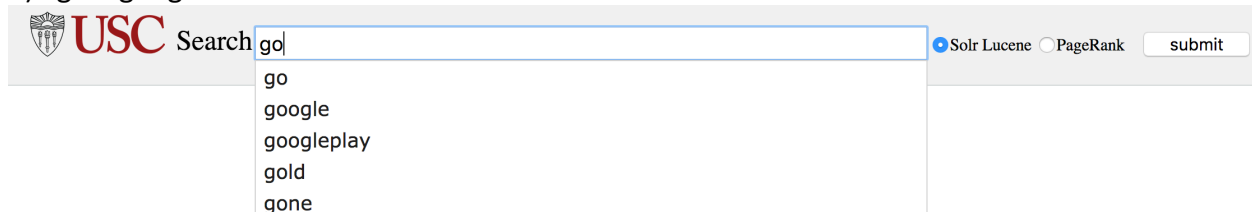


USC Search cali

- cali
- california
- california's
- casinos
- click

☒ Solr Lucene ☐ PageRank

3) go – google

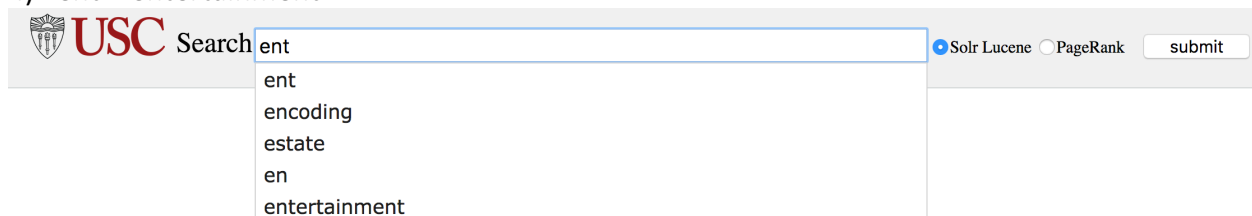


USC Search go

- go
- google
- googleplay
- gold
- gone

☒ Solr Lucene ☐ PageRank

4) ent – entertainment

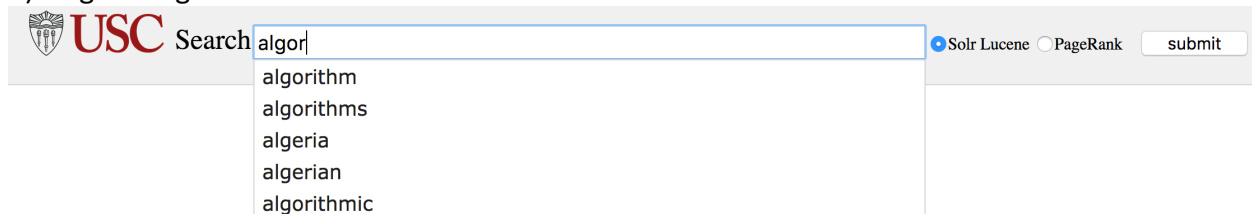


USC Search ent

- ent
- encoding
- estate
- en
- entertainment

☒ Solr Lucene ☐ PageRank

5) algor – algorithms



USC Search algor

- algorithm
- algorithms
- algeria
- algerian
- algorithmic

☒ Solr Lucene ☐ PageRank