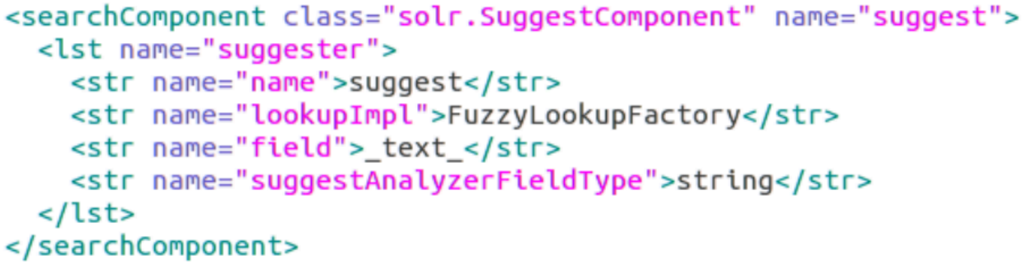
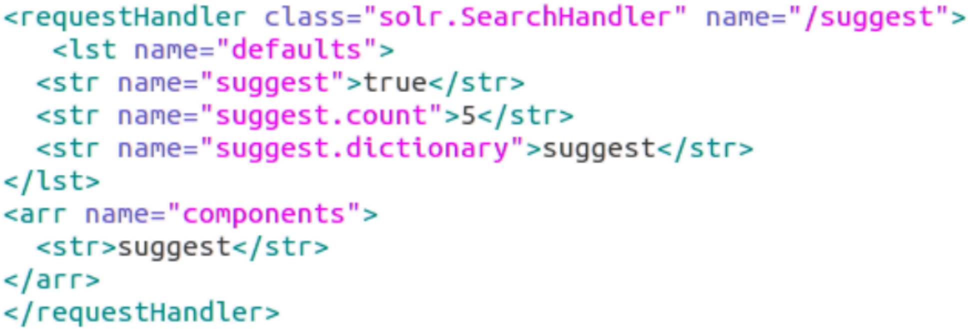
Name: Dixita Patel
USCID: 2014898085
CSCI 572: Assignment 5 - Adding Spell Checking, AutoComplete and Snippets to Your Search Engine

Steps followed to complete the assignment:

1. As the first step for auto suggest & spell correction to work, I updated the solrconfig.xml file inside the core and added a search component in it using solr's FuzzyLookupFactory for the suggest feature



```
<searchComponent class="solr.SuggestComponent" name="suggest">
  <lst name="suggester">
    <str name="name">suggest</str>
    <str name="lookupImpl">FuzzyLookupFactory</str>
    <str name="field">_text_</str>
    <str name="suggestAnalyzerFieldType">string</str>
  </lst>
</searchComponent>
```



```
<requestHandler class="solr.SearchHandler" name="/suggest">
   <lst name="defaults">
  <str name="suggest">true</str>
  <str name="suggest.count">5</str>
  <str name="suggest.dictionary">suggest</str>
</lst>
<arr name="components">
  <str>suggest</str>
</arr>
</requestHandler>
```

2. I first used a java code using Apache Tika's libraries for html parsing, to parse all the indexed html files for my news website. This code generates a file called "big.txt" which contains words from all the html files.

3. This text files acts as an input to the spell corrector code.

4. I used Peter Norvig's SpellCorrector.php code for implementing spell correction. This code takes previously generated "big.txt" as input and performs spell correction.

5. The "correct" function of the spell corrector code is invoked from my main index.php file which contains the code for accepting queries and displaying search results.

6. The first time this function receives a misspelt word from the query, it generates an inverted index of words called serialzed_dictionary.txt

7. For any further spelling corrections, this inverted index is used to find the correct spelling.

8. After this, to handle the autosuggest feature, I used jquery. I used the json feature from solr and made an ajax call

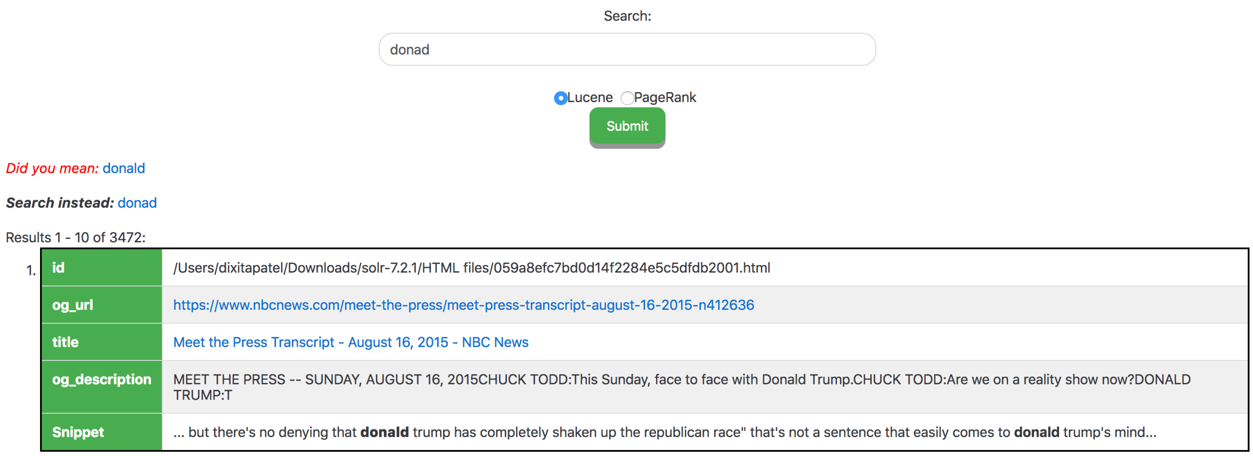
to it using it’s URL. This displays top five autosuggested words for every word if they exist. Removed stop words.

9. I have imported simple_html_dom.php to extract information from a given webpage. This is used for generating snippets.
10. Used the function file_get_contents to get contents of the html files and concatenated it with the description. Used jquery on this to find a snippet of length not greater than 156 characters which contains all the terms in the provided.
11. I used regex to eliminate special characters and the function strip_tags to discard any snippet candidate consisting of html tags.

Analysis of the results:

1. Spell checking

a) <u>Donad</u>



Search: donad

● Lucene ○ PageRank

Did you mean: donald

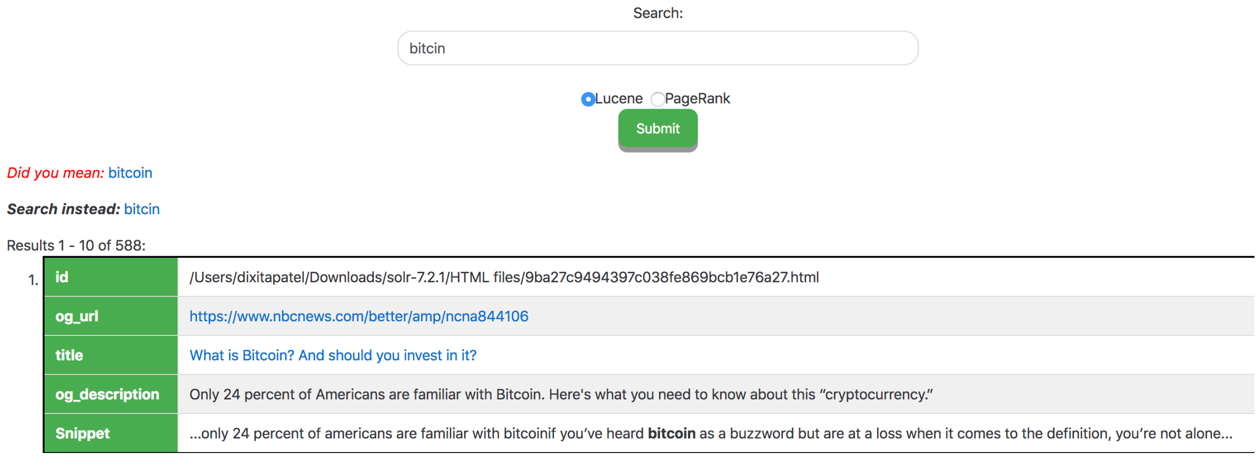
Search instead: donad

Results 1 - 10 of 3472:

1.

id	/Users/dixitapatel/Downloads/solr-7.2.1/HTML files/059a8efc7bd0d14f2284e5c5dfdb2001.html
og_url	https://www.nbcnews.com/meet-the-press/meet-press-transcript-august-16-2015-n412636
title	Meet the Press Transcript - August 16, 2015 - NBC News
og_description	MEET THE PRESS -- SUNDAY, AUGUST 16, 2015CHUCK TODD:This Sunday, face to face with Donald Trump.CHUCK TODD:Are we on a reality show now?DONALD TRUMP:T
Snippet	... but there's no denying that **donald** trump has completely shaken up the republican race" that's not a sentence that easily comes to **donald** trump's mind...

b) <u>Bitcin</u>



Search: bitcin

● Lucene ○ PageRank

Did you mean: bitcoin

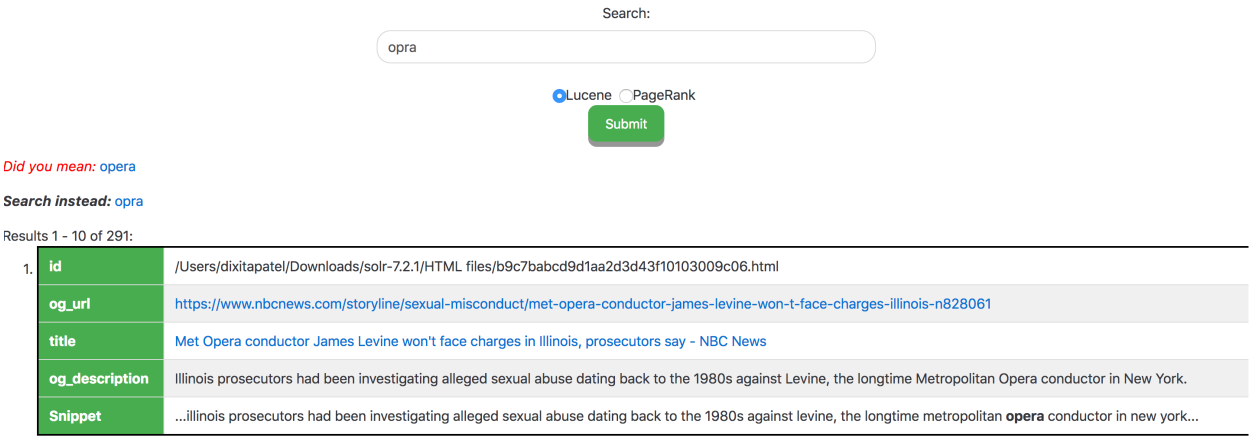
Search instead: bitcin

Results 1 - 10 of 588:

1.

id	/Users/dixitapatel/Downloads/solr-7.2.1/HTML files/9ba27c9494397c038fe869bcb1e76a27.html
og_url	https://www.nbcnews.com/better/amp/ncna844106
title	What is Bitcoin? And should you invest in it?
og_description	Only 24 percent of Americans are familiar with Bitcoin. Here's what you need to know about this “cryptocurrency.”
Snippet	...only 24 percent of americans are familiar with bitcoinif you’ve heard **bitcoin** as a buzzword but are at a loss when it comes to the definition, you’re not alone...

c) opra



Search:

opra

● Lucene ○ PageRank

Submit

Did you mean: opera

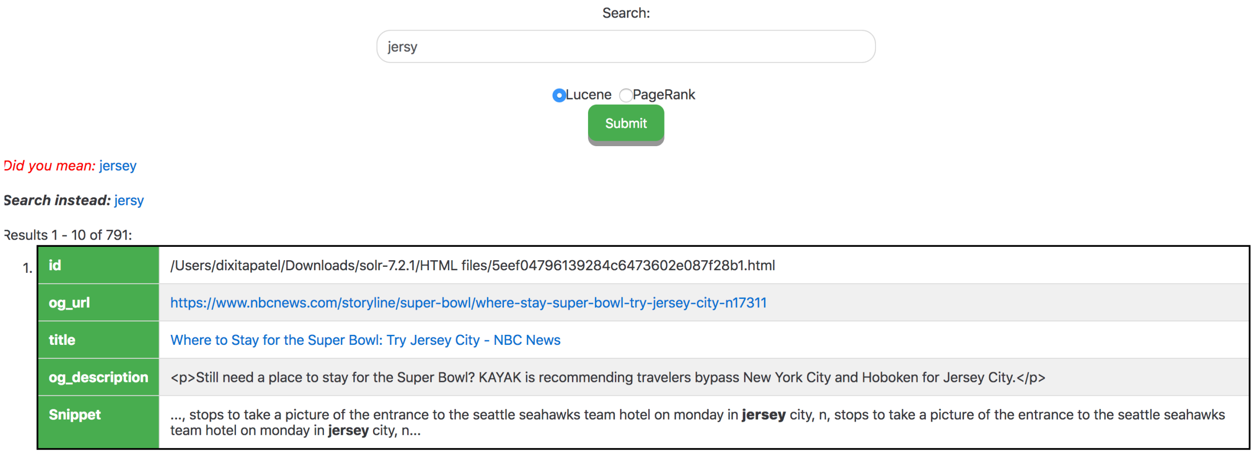
Search instead: opra

Results 1 - 10 of 291:

1.

id	/Users/dixitapatel/Downloads/solr-7.2.1/HTML files/b9c7babcd9d1aa2d3d43f10103009c06.html
og_url	https://www.nbcnews.com/storyline/sexual-misconduct/met-opera-conductor-james-levine-won-t-face-charges-illinois-n828061
title	Met Opera conductor James Levine won't face charges in Illinois, prosecutors say - NBC News
og_description	Illinois prosecutors had been investigating alleged sexual abuse dating back to the 1980s against Levine, the longtime Metropolitan Opera conductor in New York.
Snippet	...illinois prosecutors had been investigating alleged sexual abuse dating back to the 1980s against levine, the longtime metropolitan **opera** conductor in new york...

d) jersy



Search:

jersy

● Lucene ○ PageRank

Submit

Did you mean: jersey

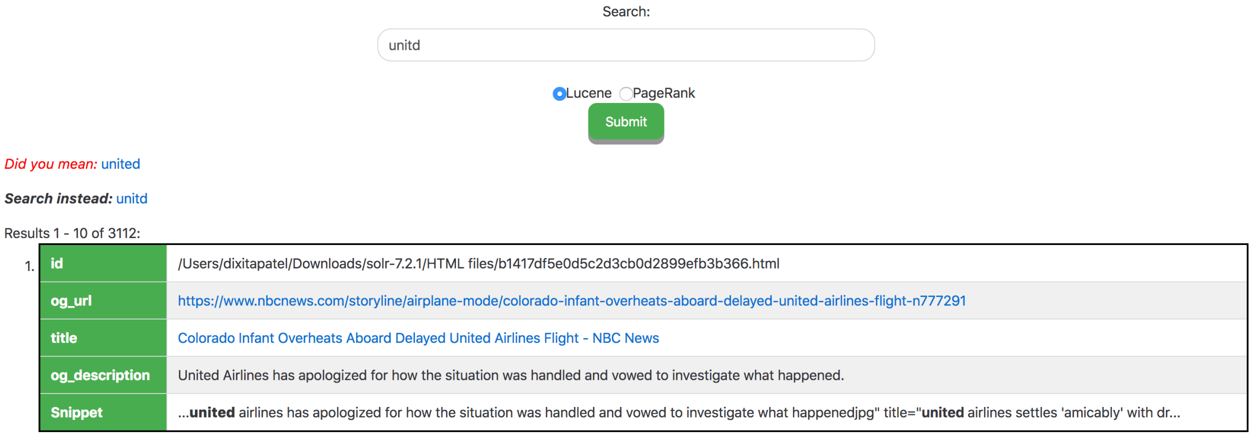
Search instead: jersy

Results 1 - 10 of 791:

1.

id	/Users/dixitapatel/Downloads/solr-7.2.1/HTML files/5eef04796139284c6473602e087f28b1.html
og_url	https://www.nbcnews.com/storyline/super-bowl/where-stay-super-bowl-try-jersey-city-n17311
title	Where to Stay for the Super Bowl: Try Jersey City - NBC News
og_description	<p>Still need a place to stay for the Super Bowl? KAYAK is recommending travelers bypass New York City and Hoboken for Jersey City.</p>
Snippet	..., stops to take a picture of the entrance to the seattle seahawks team hotel on monday in **jersey** city, n, stops to take a picture of the entrance to the seattle seahawks team hotel on monday in **jersey** city, n...

e) unitd



Search:

unitd

● Lucene ○ PageRank

Submit

Did you mean: united

Search instead: unitd

Results 1 - 10 of 3112:

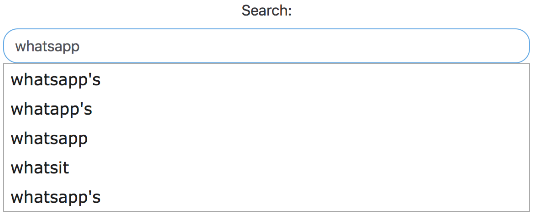
1.

id	/Users/dixitapatel/Downloads/solr-7.2.1/HTML files/b1417df5e0d5c2d3cb0d2899efb3b366.html
og_url	https://www.nbcnews.com/storyline/airplane-mode/colorado-infant-overheats-aboard-delayed-united-airlines-flight-n777291
title	Colorado Infant Overheats Aboard Delayed United Airlines Flight - NBC News
og_description	United Airlines has apologized for how the situation was handled and vowed to investigate what happened.
Snippet	...**united** airlines has apologized for how the situation was handled and vowed to investigate what happenedjpg" title="**united** airlines settles 'amicably' with dr...

2. Autocomplete

a) Whatsapp

W – width, with, when, which, whatsapp’s
Wh – who, white, when, which, whatapp’s
Whatsa – whats, whatsoever, whatsapp, whatsit, whatsapp’s



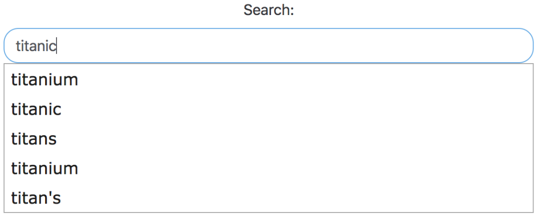
Search:

whatsapp

whatsapp's
whatapp's
whatsapp
whatsit
whatsapp's

b) titanic

tita – title, total, totally, totalsides, tirades
titan – titanic, titans, tiananmen, tetanus, tirades



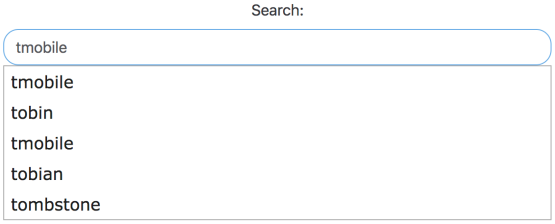
Search:

titanic

titanium
titanic
titans
titanium
titan's

c) tmobile

tmob – tomb, tobel, tobacco, tobby, tombstone
tmobil – tmobile, tobin, tobacco, tobby, tombstone



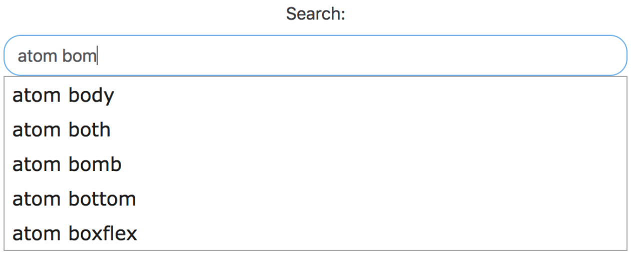
Search:

tmobile

tmobile
tobin
tmobile
tobian
tombstone

d) **atom bomb**

atom – automatic, atmosphere, atomic, automakers
atom b – atom by, atom body, atom be, atom but



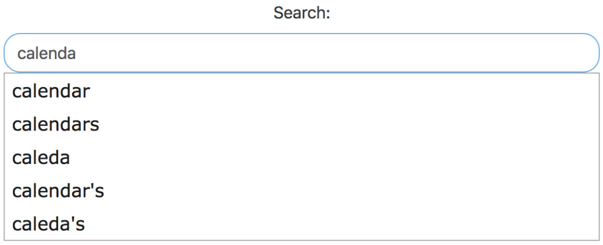
Search:

atom bom

atom body
atom both
atom bomb
atom bottom
atom boxflex

e) **calendar**

ca – canonical, career, can
cale – careers, called, callback, came



Search:

calenda

calendar
calendars
caleda
calendar's
caleda's