

Homework 5

Spell Checking, AutoComplete and Snippets

NYPost Search Engine

By: Mukesh Kumar Dangi

ReadMe: Tool and Steps Followed in the project

We used HW4 as based for this project and added 3 more functionalities 1. Spell correction, auto suggestion and snippet extraction.

GitHub: <https://github.com/mukeshkdangi/nypost-searchengine>

Step 1: Created Big.txt:

Extracted text from NYPost crawled data sets using Tika TagSoup parser written in Java. and added the text into Big.txt, which is the vocabulary for spell check and auto suggest.

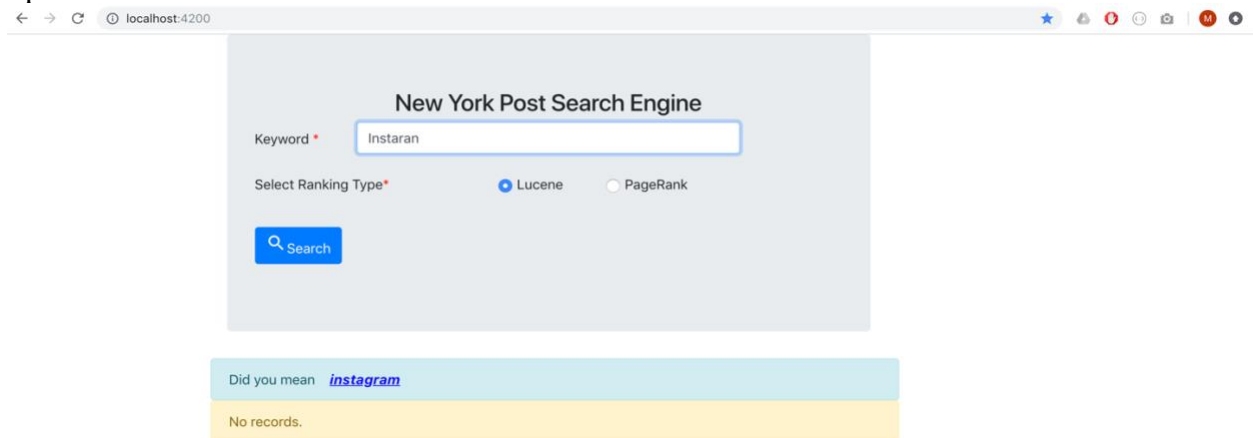
Step 2: Spell Correction:

Used spelling-corrector NodeJS module for spelling correction. We trained this module for our custom spell check. Following Steps were followed to initialize and train the module:

```
let SpellCorrector = require('spelling-corrector');  
let spellCorrector = new SpellCorrector();  
spellCorrector.loadDictionary('/Users/mukesh/Office/Tools/Solr/solr-7.5.0/big.txt');
```

When a user inputs an incorrect query and submits the page, the program checks to see if it's a valid word or not. If it's a valid word corresponding results are displayed else a phrase with “**Did you mean <correct_word>**”. When user clicks on correct word, results gets displayed accordingly.

Spell corrector:



The screenshot shows a web browser window with the address bar displaying 'localhost:4200'. The main content area is titled 'New York Post Search Engine'. It features a search input field with the text 'Instaran'. Below the input field, there are two radio buttons for 'Select Ranking Type': 'Lucene' (selected) and 'PageRank'. A blue 'Search' button is located below the ranking options. At the bottom of the form, there is a light blue bar that says 'Did you mean [instagram](#)' and a yellow bar below it that says 'No records.'

Step 3: Auto complete:

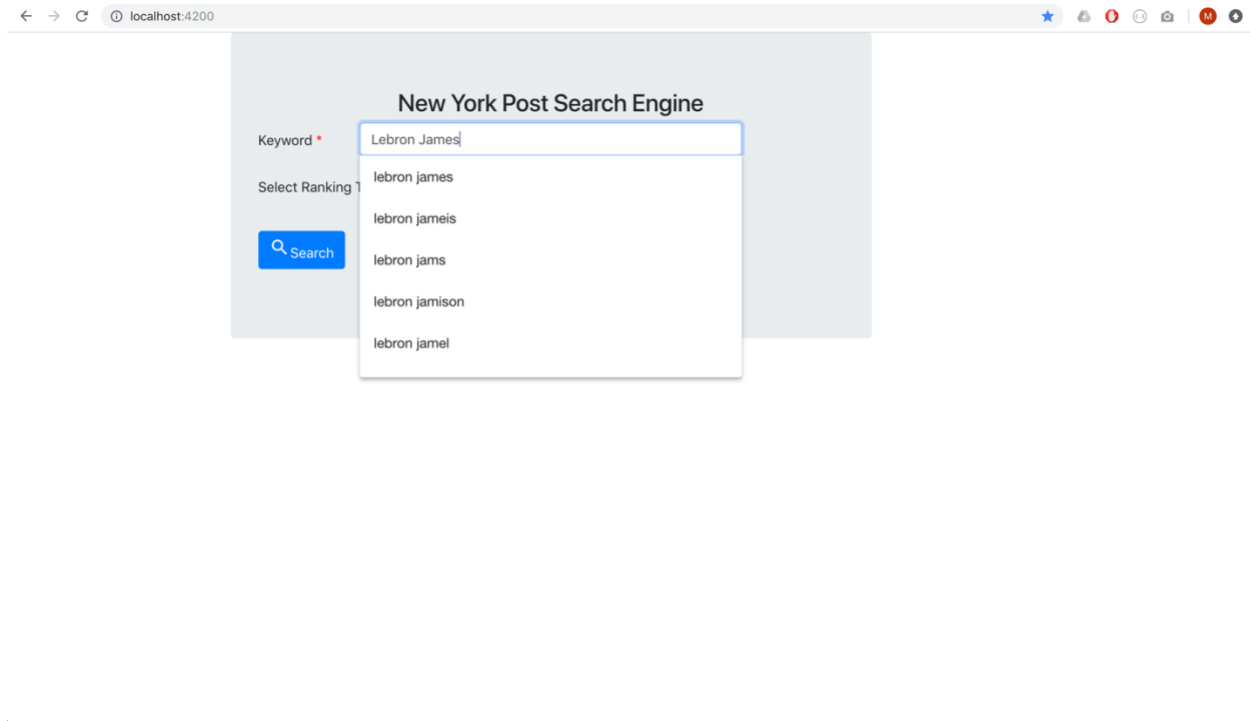
I have implemented autosuggestion using **Angular 7** following libraries *MatAutocompleteModule*, *MatOptionModule*, *BrowserAnimationsModule*, *BrowserModule*, *FormsModule*, *HttpClientModule*

Called the autocomplete API on key down or any text change (**keydown**)="searchAutoComplete(\$event)" from the main page.

Then called 'http://localhost:8983/solr/nypost/suggest?q=' + req.query.keyword API for suggestion for the entered key.

<http://localhost:8983/solr/nypost/suggest?q=> + req.query.keyword and sent it back to the home page.

The results of auto complete are displayed in a dropdown list. For Multiple keywords I appended the last space indexed word to the drop-down list.



4. Analysis of the results:

1. SPELL CORRECTION:

Incorrect Word	Correct Word
instagrm	instagram
Donal Trmp	donal trump
Snapcht	snapchat
north koea	north korea
Illegal immgraton	illegal immigration

2. AUTO-COMPLETE:

Prefix	Autocompletion
kor	kor korea kerry karol kardashian korean
cali	cali click called claims clinton
Immi	immi immediately immigration immigrant immigrants
face	face facebook faces facing fact
inst	inst instagram instagram.com istax inside

3. Snippet

For snippet, we first extracted the body text for all the files and maintained a map of doc id and corresponding text. Next, we searched 10 docs for every term user entered. And retrieved the 10 docs with URL, doc id, title and snippet. For multiword query, we tried to match the whole query term in the 10 docs returned from Solr and if found we returned otherwise we broke the query terms into pieces and tried to look for individual term in the docs and returned results.

← → ↻ localhost:4200

New York Post Search Engine

Keyword *

Select Ranking 1

- instagram acc
- instagram account
- instagram action
- instagram adchoices
- instagram archive

Results 1 - 10 of 15403

Query - **instagram** 14 ms

Title : [Instagram starts taking online bullying very seriously](#)

URL : <https://nypost.com/2018/10/10/instagram-starts-taking-online-bullying-very-seriously/>

ID : /Users/mukesh/Office/Tools/Solr/solr-7.5.0/nypost/nypost/d21f0ac7-31a0-48cf-8a0b-eff5dfa09502.html

Description : While the tech world was busy discovering Google's newest toys, Instagram on Tuesday announced the introduction of new safety features meant to reduce online bullying and encourage the spread of ki...

Snippet : ... this is just an effect. Filed under artificial intelligence , cyber bullying , instagram , social media trending now Tom Brady's postgame words to Gronk are an...

Title : [Instagram starts taking online bullying very seriously](#)

URL : <https://nypost.com/2018/10/10/instagram-starts-taking-online-bullying-very-seriously/>

ID : /Users/mukesh/Office/Tools/Solr/solr-7.5.0/nypost/nypost/23783b2d-5320-40e7-877e-f1892a63e2f1.html

Description : While the tech world was busy discovering Google's newest toys, Instagram on Tuesday announced the introduction of new safety features meant to reduce online bullying and encourage the spread of ki...

Snippet : ...nger WhatsApp Email Copy Filed under artificial intelligence , cyber bullying , instagram , social media Share this article: Share this: Facebook Twitter Google...