CSCI572 HW5 Report

Adding Spell Checking, AutoComplete and Snippets to Your Search Engine

Wei Fei

11/26/2018

❖ Introduction

- This exercise is based on previous homework. We will keep the original functionalities that support query search by Lucene and page rank, but add more powerful functionalities, such as spell checking, auto complete and snippets.

❖ Steps

- **Auto completion**
  - Configuration in solrconfig.xml
    - Add a search component and tell it to use the SuggestComponent.
    - Add a request handler into the file
  - Use "nypost" folder as my input, write a Java program to generate an output file named "big.txt" that includes books from Project Gutenberg, Wiktionary, British National Corpus. We use Apache Tika package to parse the input string to our autocompletion library.
  - Go to port 8983, use "/suggest" with some search queries to test my autocomplete functionality.
- **Spelling correction**
  - In the php file, use third party PHP library of Norvig spelling corrector to implement spelling correction functionality for search query.
  - Download the PHP version of Norvig spelling corrector
  - Enhance spelling correction program with a set of terms to "nypost"
  - Use Norvig's correction program to calculate edit distance, in order to use minimum number of operations to convert the wrong input query to correct suggestions.
- **Snippet**
  - The last functionality we need to implement is snippet, so that the page will return search results with bold query words.
  - I implemented php code to look for a string match of the query terms with the web page. Return the first sentence that provides a match. If no match is found, then no snippet is returned.
  - I use an external library called "simple_html_dom.php"

- Index.php
  - Use MMAP server to start a local host, and then put index.php into htdocs file.
  - Include two external library files in the index.php, and then implement frontend view.

❖ Analysis of the Result

- Five examples of misspelled terms that are correctly handled by my spelling correction program

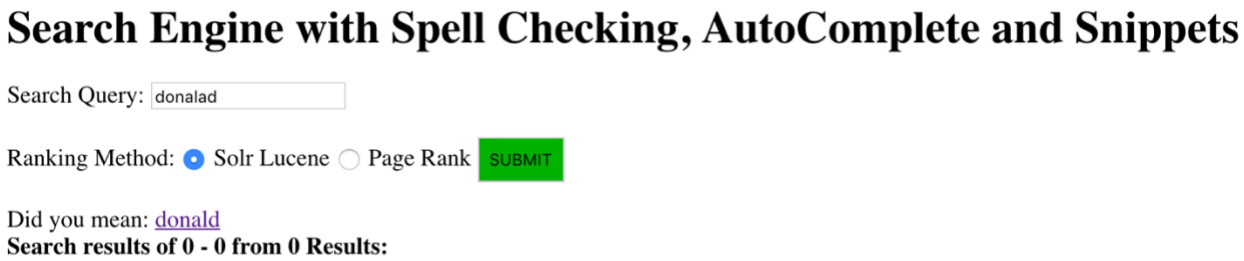
Misspelled terms	Corrected spelling
donalad	donald
lakars	lakers
las vagas	las vegas
lebron jamas	lebron james
university of califania	university of california

- Five examples of autocompletion

Incomplete terms	Autocompletion term
app	app, appid, apple, application, apps
link	link, linkedin, links, likes, linkid
fork	fork, for, form, former, forever
kobe	kobe, korean, koreans, korea's, koreas
power	power, powered, powerful, powers, powerhouse

❖ Screenshots

- **Five misspelled terms and corrected spelling**



Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: lakars

Ranking Method: Solr Lucene Page Rank SUBMIT

Did you mean: lakers
Search results of 0 - 0 from 0 Results:

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: las vagas

Ranking Method: Solr Lucene Page Rank SUBMIT

Did you mean: las vegas
Search results of 1 - 10 from 2132 Results:

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: lebron jamas

Ranking Method: Solr Lucene Page Rank SUBMIT

Did you mean: lebron james
Search results of 1 - 10 from 1362 Results:

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: university of califania

Ranking Method: Solr Lucene Page Rank SUBMIT

Did you mean: university of california
Search results of 1 - 10 from 15403 Results:

- **Five examples of auto completion**

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: app

Ranking Metho Page Rank SUBMIT

app
appid
apple
application
apps

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: link

Ranking Metho Page Rank SUBMIT

- link
- linkedin
- links
- likes
- linkid

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: fork

Ranking Metho Page Rank SUBMIT

- fork
- for
- form
- former
- forever

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: kobe

Ranking Metho Page Rank SUBMIT

- kobe
- korea
- korean
- korea's
- koreas

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: power

Ranking Metho Page Rank SUBMIT

- power
- powered
- powerful
- powers
- powerhouse

➢ Sample output with snippets

- **Default**

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: donald trump

Ranking Method: (•) Solr Lucene ( ) Page Rank SUBMIT

Search results of 1 - 10 from 8041 Results:

- **TItle:** Donald and Melania Trump tour towns destroyed by Hurricane Michael
  Url: https://nypost.com/2018/10/15/donald-and-melania-trump-begin-tour-of-towns-destroyed-by-hurricane-michael/
  ID: 3ebce63e-fba7-403d-b0b8-50f1db9e67b4.html
  Snippet: ... **donald** and Melania **Trump** tour towns destroyed by Hurricane Michael ...

- **TItle:** Donald and Melania Trump tour towns destroyed by Hurricane Michael
  Url: https://nypost.com/2018/10/15/donald-and-melania-trump-begin-tour-of-towns-destroyed-by-hurricane-michael/
  ID: b02d11fc-24b7-4eab-917d-83905c542d1e.html
  Snippet: ... **donald** and Melania **Trump** tour towns destroyed by Hurricane Michael ...

- **TItle:** Donald Trump has invented a new way to win
  Url: https://nypost.com/2016/02/15/donald-trump-has-invented-a-new-way-to-win/
  ID: 51f46b36-0e7b-4d6a-837d-938d9587b16e.html
  Snippet: ... **donald Trump** has invented a new way to win <...

- **TItle:** November 9, 2016 | New York Post
  Url: https://nypost.com/2016/11/09/
  ID: 1e666eac-ab65-4885-869e-3d057deeb87c.html
  Snippet: ... November 9, 2016 | 11:59pm **donald Trump** will quickly begin to define his presidential legacy by nominating ...

- **TItle:** How Donald Trump manipulated Bon Jovi out of Bills purchase
  Url: https://nypost.com/2017/11/06/how-donald-trump-manipulated-bon-jovi-out-of-bills-purchase/
  ID: 4882c3e5-f749-4528-8082-8df47c4c48e5.html
  Snippet: ... How **donald Trump** manipulated Bon Jovi out of Bills purchase ...

- **PageRank**

Search Engine with Spell Checking, AutoComplete and Snippets

Search Query: star wars

Ranking Method: ( ) Solr Lucene (•) Page Rank SUBMIT

Search results of 1 - 10 from 10756 Results:

- **TItle:** Photos | New York Post
  Url: https://nypost.com/photos/
  ID: 88f9f9d1-491d-4df3-981d-863a8b1b485a.html
  Snippet: ... Zendaya does her own stunts with Spider-Man and more **star** snaps Zendaya fli...

- **TItle:** Sports | New York Post
  Url: https://nypost.com/sports/
  ID: 8790e35b-13e7-4277-abec-631c5ca93fd3.html
  Snippet: ... Astros **star** borrows from Aaron Judge with risky Red Sox trolling ...

- **TItle:** Fashion News, Photos, and Video | New York Post
  Url: https://nypost.com/fashion/
  ID: 6bc5ab6f-4ca4-4588-a3a8-f9a6bb45f6ab.html
  Snippet: ...M defamation suit against Gizmodo Media Group Inside Usher's **star** -studded 40th birthday bash Tekashi 6ix9ine
  ...

- **TItle:** Technology News & Reviews | New York Post
  Url: https://nypost.com/tech/
  ID: 40b8760d-53b6-4edb-a9b7-f60c3c13258d.html
  Snippet: ... Venture capital investors pour $3B into NYC tech **star** tups October 13, 2018 | 9:33pm ...

- **TItle:** Columnists | New York Post
  Url: https://nypost.com/columnists/
  ID: fc17eac6-ad31-4c0c-b079-3bebbc17fcb9.html
  Snippet: ... Michael **star** r ...