Project 4: Complete Search and Analytics Solution based on dissecting twitter data

Abstract

The aim of this project is to build a solution that provides insight related to social conversations on important societal issues and to gain experience of building an end-to-end IR solution including data collection, search relevance, and analytics.

1 Introduction

The project aims to index tweets using “Solr”, which have tweets with the following specifications:

- 5 topics: Environment, Politics, Crime, Social Unrest and Infrastructure
- 5 cities: NYC, Delhi, Bangkok, Paris and Mexico City
- 5 languages: English, Hindi, Spanish, French and Thai

Our main task after indexing the tweets was to:

- Detect trending phrases/hashtags from each topic/city.
- Retrieve top relevant tweets for each trending phrase/hashtag.

Perform analysis such as:

- Time series – for a given city
- Comparison across the cities – sentiment, volume etc.
- Sentiment analysis – overall sentiment of general public for a phrase/hashtag

Some optional ideas:

- Faceted search on named entity
- Summarization – either on hashtags or topics

2 Implementation

2.1 Back End Implementation

2.1.1 Java Servlet

The implementation of Back end part has been achieved using Java Servlet. A Java servlet is a Java software component that extends the capabilities of a server. Although servlets can respond to any types of requests they most commonly implement web containers for hosting web applications on web servers and thus qualify as a server-side servlet web API.

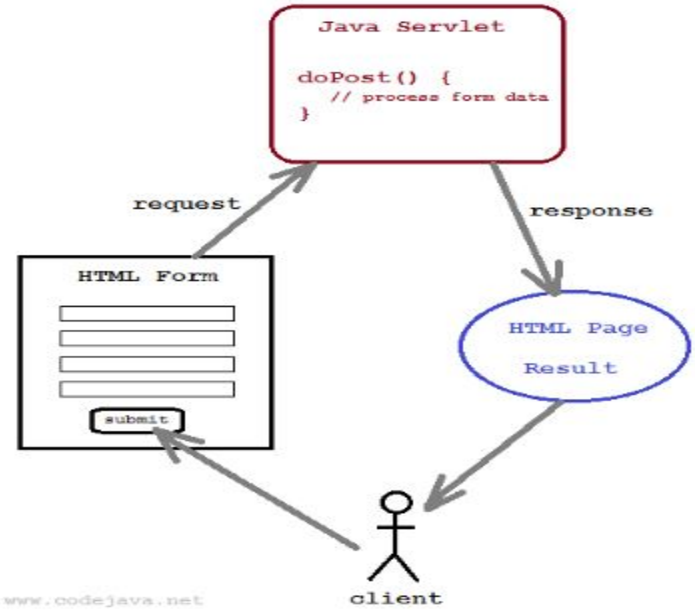


Fig 1: Java Servlet's communication with HTML

2.2 Front End Implementation

2.2.1 JavaServer Pages (JSP)

JavaServer Pages (JSP) is a technology that helps software developers create dynamically generated web pages based on HTML, XML, or other document types. Released in 1999 by Sun Microsystems, JSP is similar to PHP and ASP, but it uses the Java programming language.

In our project, we have used JSP to create webpages based on HTML and Cascading Style Sheets (CSS).

2.2.2 Hypertext Markup Language (HTML)

HyperText Markup Language (HTML) is a markup language for creating a webpage. Webpages are usually viewed in a web browser. They can include writing, links, pictures, and even sound and video. HTML is used to mark and describe each of these kinds of content so the web browser can display them correctly.

2.2.3 Cascading Style Sheets (CSS)

CSS stands for Cascading Style Sheets. CSS describes how HTML elements are to be displayed on screen, paper, or in other media. CSS saves a lot of work. It can control the layout of multiple web pages all at once. External stylesheets are stored in CSS files.

3 Member Contributions

3.1 Jui Kate- Implementation of Java Servlet

Jui has implemented the back-end part of the project. She has used Java Servlet inorder to fetch the tweet data from Solr and the tweets are processed to obtain the desired form. Then, the processed tweets are pushed to front end using Servlet.

3.2 Vakul Bhatia- Implementation of JavaServer Pages

Vakul has done the part of obtaining the processed tweets from Servlet and displaying it using HTML and CSS. He has used JavaServer pages inorder to display the webpage using both HTML and CSS.

Vakul has implemented JSP pages by using several delimiters for scripting functions. The most basic is <% ... %>, which encloses a JSP scriptlet. A scriptlet is a fragment of Java code that is run when the user requests the page. Other common delimiters include <%=... %> for expressions, where the scriptlet and delimiters are replaced with the result of evaluating the expression, and directives, denoted with <%@ ... %>. A JSP page looks similar to an HTML page, but a JSP page also has Java code in it. We can put any regular Java Code in a JSP file using a scriplet tag which start with <% and ends with %>.

3.4 Tannu Priya Singh - Creation of HTML webpage

Tannu has done the part of displaying the processed tweets, which are obtained from the Java Servlet using Hypertext Markup Language. She has written the HTML skeleton and then created the webpage with the required specifications.

3.3 Sheryl Evangelene Pulikandala - Implementation of CSS

Sheryl has done the part of displaying the processed tweets, which are obtained from the Java Servlet using Cascading Style Sheets. A CSS code is used to style an HTML document. After the creation of a website by Tannu, a CSS file is created with the required styling and a link to the .css file is added to the HTML file. This helps the browser to know that a stylesheet is used. When the HTML file is opened in the browser, the required styling is done to the webpage.

3.5 Smrati Singh- Collecting and Indexing Tweets into Solr

Smrati has collected over 150,000 tweets spread over different cities, topics and languages. Then, she has used Solr to index these tweets in the required json format. She has used Tweepy API to collect tweets spread over 4 weeks and has indexed the files to Solr, taking care of the emoji's, hashtags, urls etc..

4 Results

The front page of our application is shown below. A search box is present which allows the user to type the keywords and the tweets which contain information of these keywords are displayed to the user.

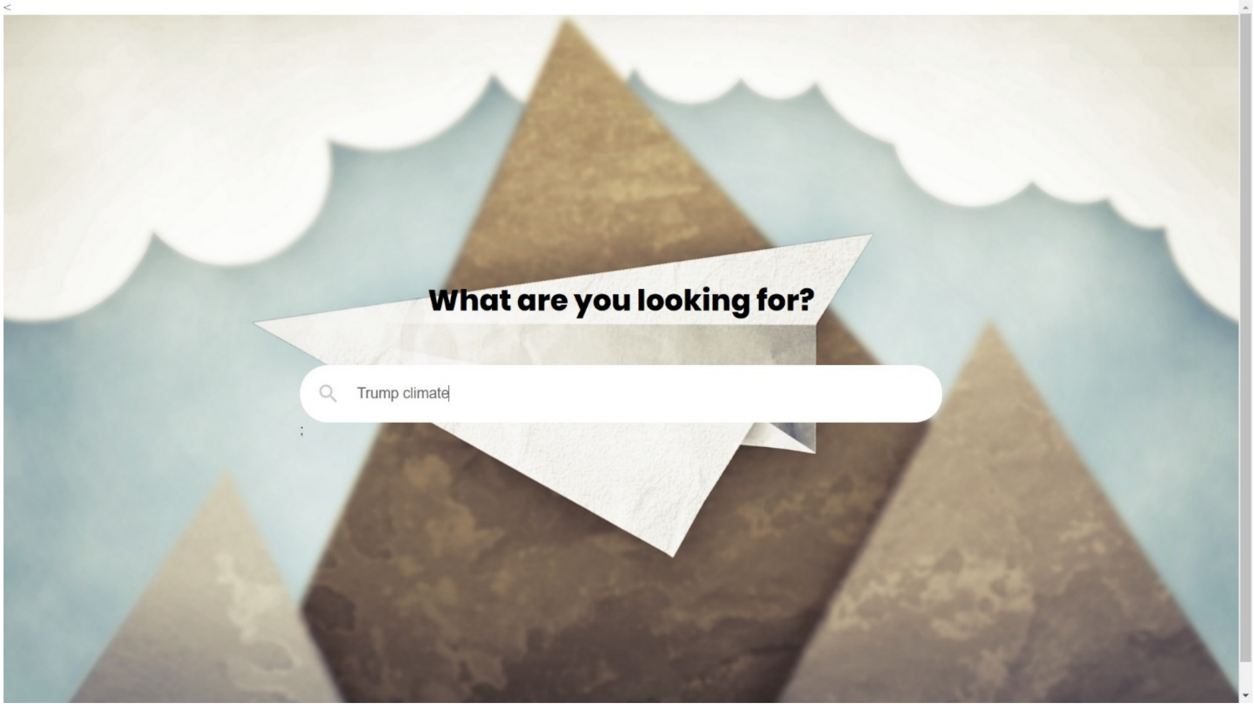
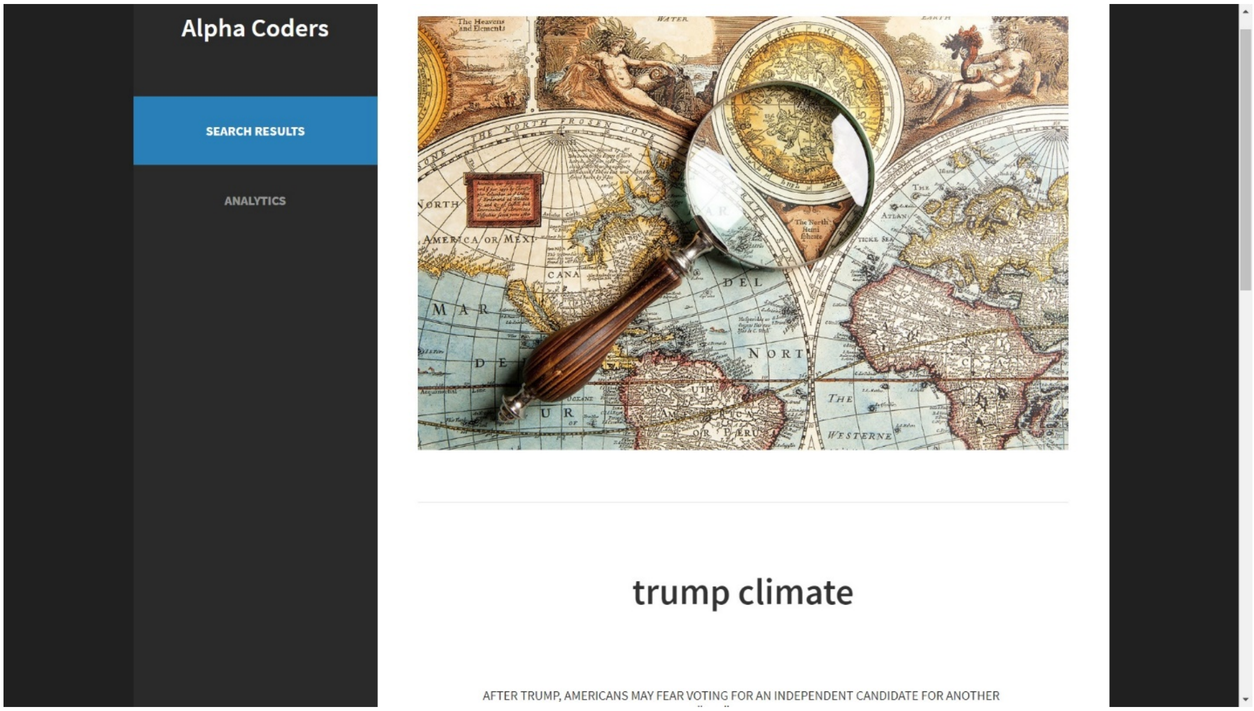


Fig 2: Front page

For example, if the user wishes to see the tweets which contain the keywords “Trump” and “Climate”, they have to type the two keywords in the search box. The page is then directed to the tweet results page.

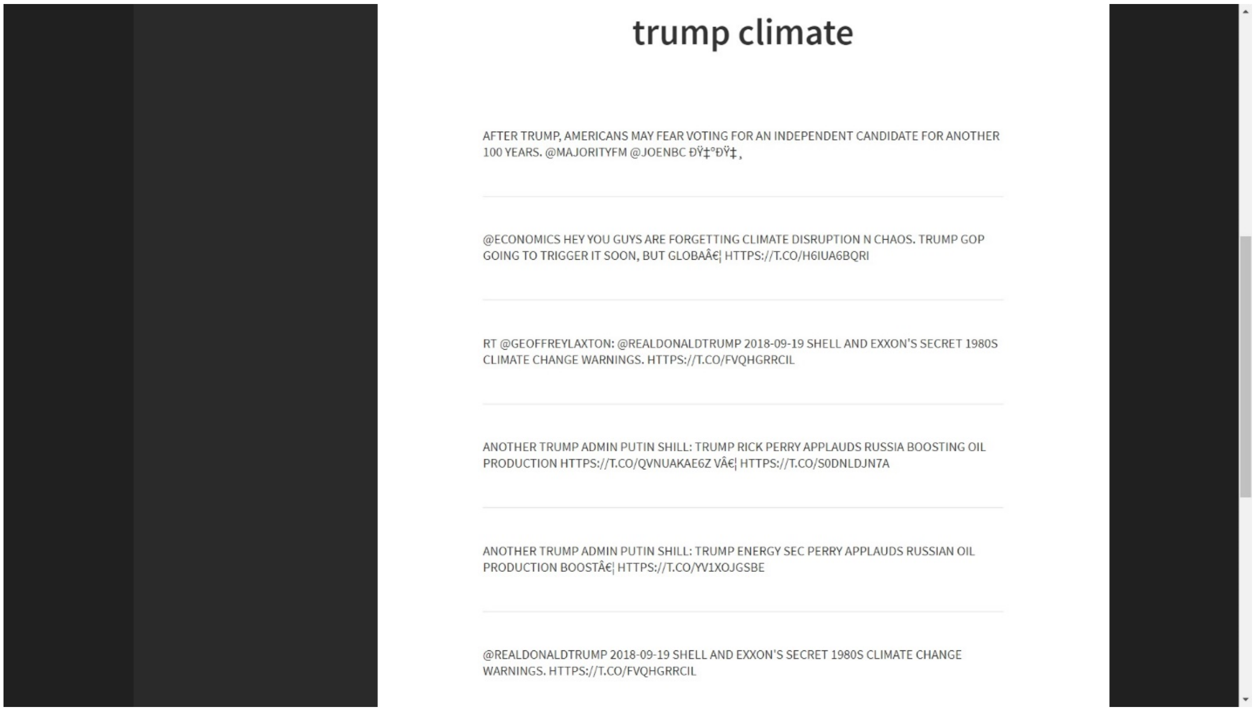


trump climate

AFTER TRUMP, AMERICANS MAY FEAR VOTING FOR AN INDEPENDENT CANDIDATE FOR ANOTHER

Fig 3: Page is directed to tweet results

Then, the results are displayed as shown below in Fig 4.



trump climate

AFTER TRUMP, AMERICANS MAY FEAR VOTING FOR AN INDEPENDENT CANDIDATE FOR ANOTHER 100 YEARS. @MAJORITYFM @JOENBC ĐŸ‡°ĐŸ‡¸

@ECONOMICS HEY YOU GUYS ARE FORGETTING CLIMATE DISRUPTION N CHAOS. TRUMP GOP GOING TO TRIGGER IT SOON, BUT GLOBAÂ€¦ HTTPS://T.CO/H6IUA6BQRI

RT @GEOFFREYLAXTON: @REALDONALDTRUMP 2018-09-19 SHELL AND EXXON'S SECRET 1980S CLIMATE CHANGE WARNINGS. HTTPS://T.CO/FVQHGRRCIL

ANOTHER TRUMP ADMIN PUTIN SHILL: TRUMP RICK PERRY APPLAUDS RUSSIA BOOSTING OIL PRODUCTION HTTPS://T.CO/QVNUAKAE6Z VÂ€¦ HTTPS://T.CO/S0DNLDJN7A

ANOTHER TRUMP ADMIN PUTIN SHILL: TRUMP ENERGY SEC PERRY APPLAUDS RUSSIAN OIL PRODUCTION BOOSTÂ€¦ HTTPS://T.CO/YV1XOJGSBE

@REALDONALDTRUMP 2018-09-19 SHELL AND EXXON'S SECRET 1980S CLIMATE CHANGE WARNINGS. HTTPS://T.CO/FVQHGRRCIL

Fig 4: Tweet results displayed

5 Conclusion

1. Tweets are indexed using Solr.
2. Java Servlet is used to fetch the tweets from Solr. This is the Back-end Implementation.
3. HTML provides the structure of the webpage.
4. CSS provides the visual and aural layout.
5. HTML and CSS are the two core technologies used for the Front End Implementation of the project.

References

[1]https://www.google.com/search?q=java+servlet&rlz=1C1JZAP_enUS810US810&source=lnms&tbm=isch&sa=X&ved=0ahUKEwj8haK_mZPfAhXJ11kKHd-bB0UQ_AUIECgD&biw=1366&bih=657#imgrc=b0HAPigtwylFeM:

[2]https://www.google.com/search?q=html+wiki&rlz=1C1JZAP_enUS810US810&source=lnms&sa=X&ved=0ahUKEwiy56qPoZPfAhUv1VkKHVpeD7YQ_AUICSgA&biw=1366&bih=657&dpr=1

[3] https://turbofuture.com/computers/How-to-Create-a-CSS-Stylesheet-with-Notepad

[4] https://www.codecademy.com/articles/local-web-page