CSE 535 – Information Retrieval

PROJECT REPORT

Project Part 3

Team Name: Team7

Members:

1. Blesson Mathew Sam - 50169279
2. Manjeet Singh - 50169105
3. Sanjay Surendranath Girija – 50170327

Links:

Code Source: https://github.com/manjeetsingh87/IR3-Maven

Project URL: http://blesson.me:8080/IR3-Maven/search

SOLR Server project: http://blesson.me:8983/solr/projectb

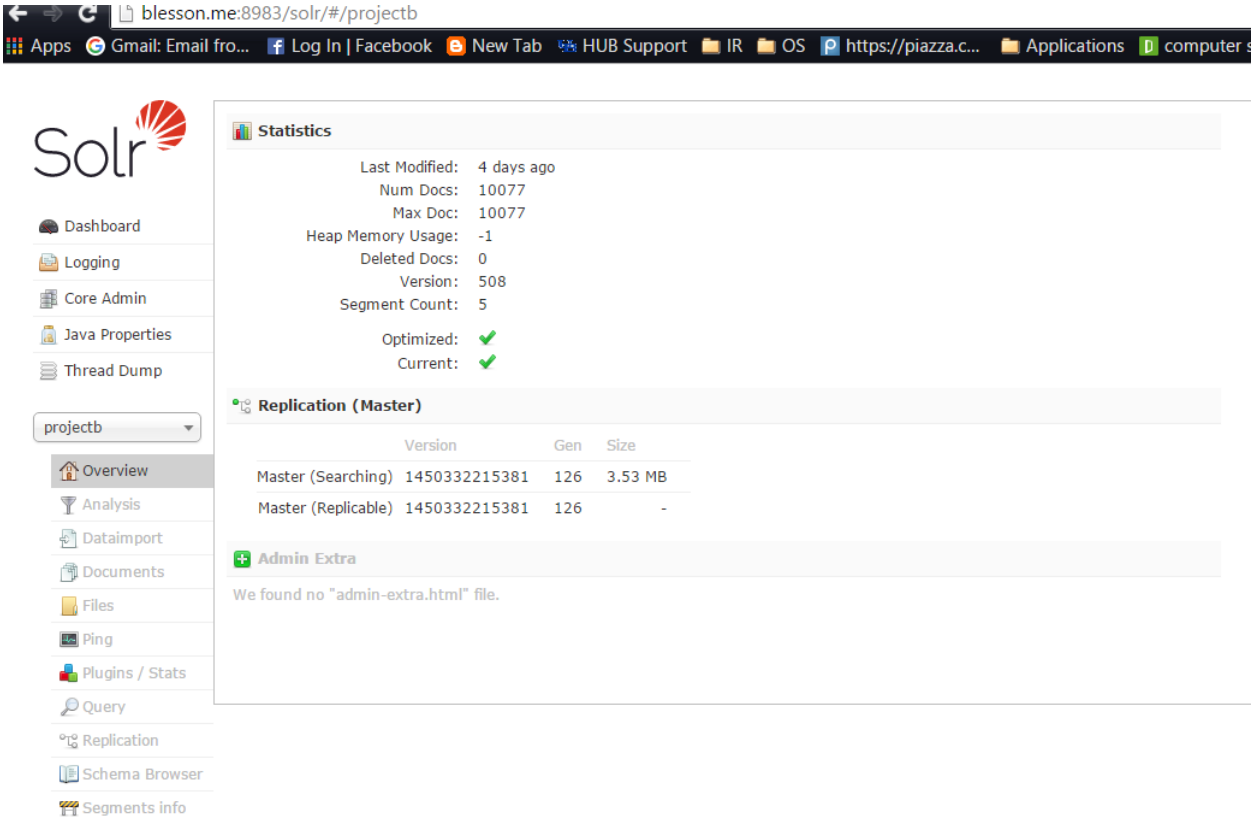
Overview

The following report provides the details of the implementation of Project Part C by Team7. We developed a multilingual search engine working on data from social media. The components incorporated into the project were – Content Tagging, Faceted Search and Cross-Document Analysis.

Data

The data we worked on was primarily twitter data over a span of two weeks. Tweet data was collected in four languages, namely:

- English - 3538
- German - 2477
- Russian - 1459
- French – 2600



Data was obtained over a broad range of topics like movies, sports, television, terrorism, Paris bombings, ISIS, refugee crisis.

The data was indexed in Solr which was deployed on a personal domain.

Components

The following components were included in the project:

1. **Content Tagging**

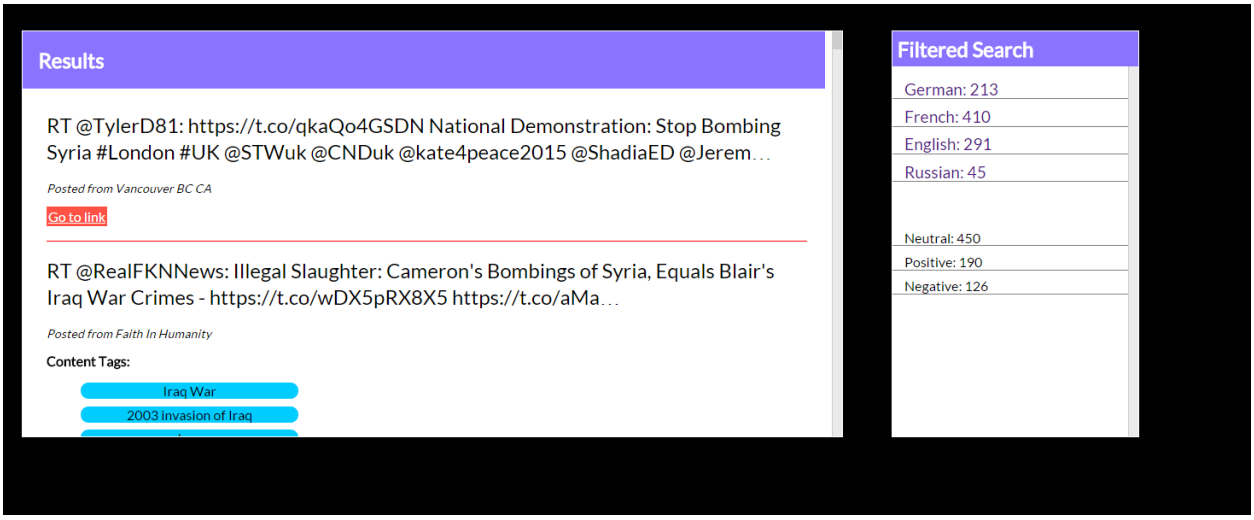
   Each tweet was translated into English and processed using the Alchemy API for obtaining content tags. The tweet data were individually sent to the Alchemy API's concept tagging module and the contents for each were retrieved. The retrieved content tags were added to the tweet data in the index.

The tagged contents are displayed in blue.

2. **Faceted Search**

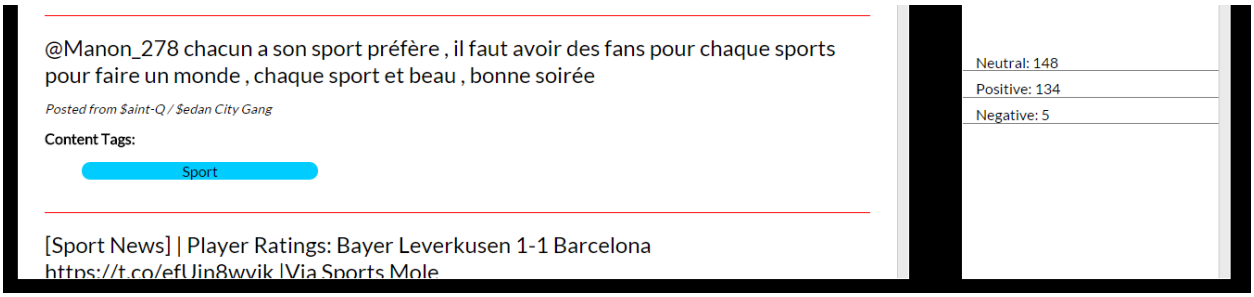
   The results of a search were retrieved and faceted based on language. This was built using the facet feature provided within Solr. The search results for a specific language could be selected based on the facet selected.

   The facets are displayed on the right



3. Cross-Lingual Analysis

Cross-lingual analysis, specifically sentiment analysis was done on each tweet using the Alchemy API's sentiment detection module. The collated sentiment results for each search is displayed on the right sidebar, representing the number of positive, negative and neutral tweets for every result.



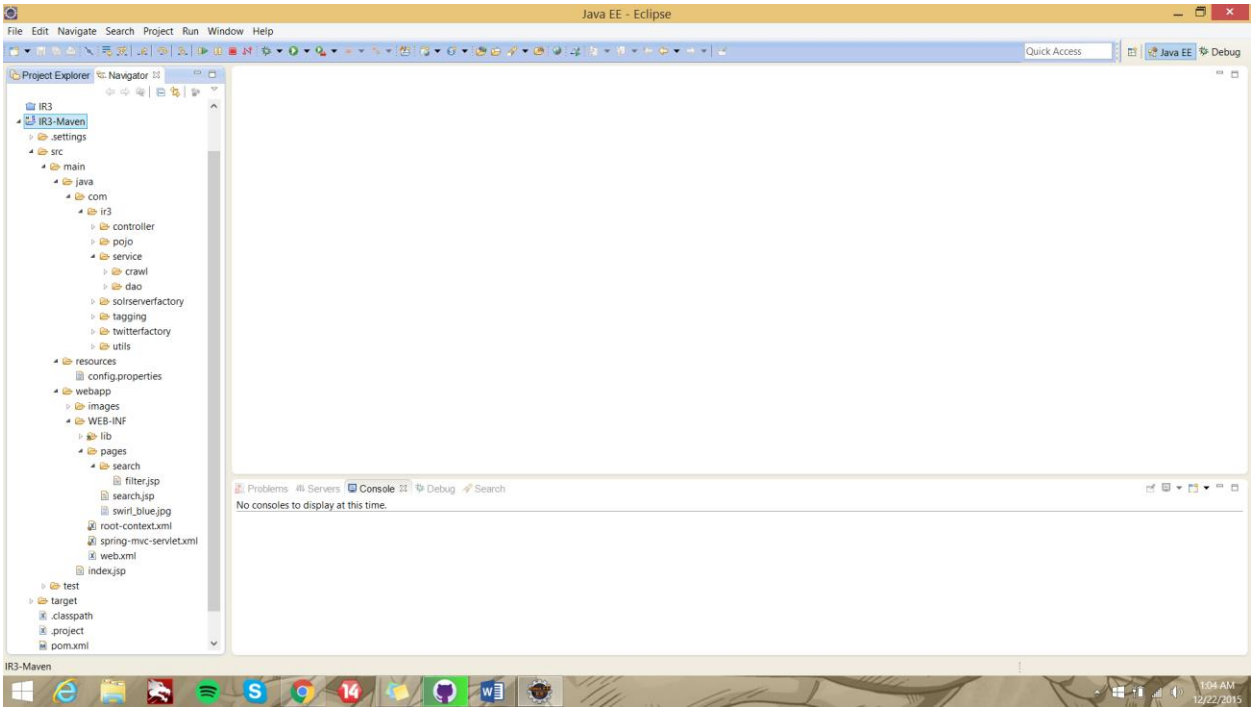
The Microsoft Bing Translator API was used for language translation. This also indirectly aids Cross-Lingual Retrieval in an indirect manner.

<u>Back End</u>

The project was developed using Maven framework, Java 1.7, Spring MVC Annotations, Solrj api, Twitter4j api, Alchemy api. This was done for the ease of deployment and synchronization when working as a team.

The servlet was written using JSP with HTML5 capability and the whole project was deployed on an Apache Tomcat Server.

Screenshot of Project Source Tree:

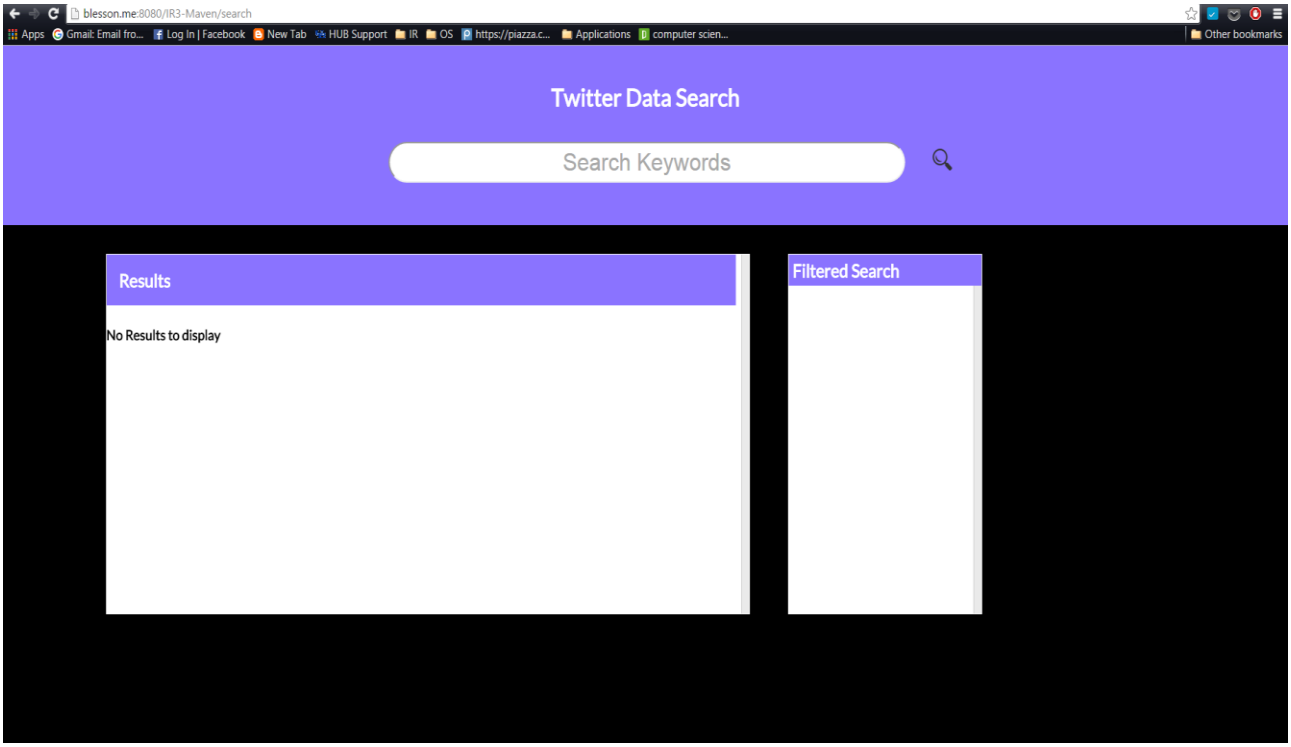


<u>Front End</u>

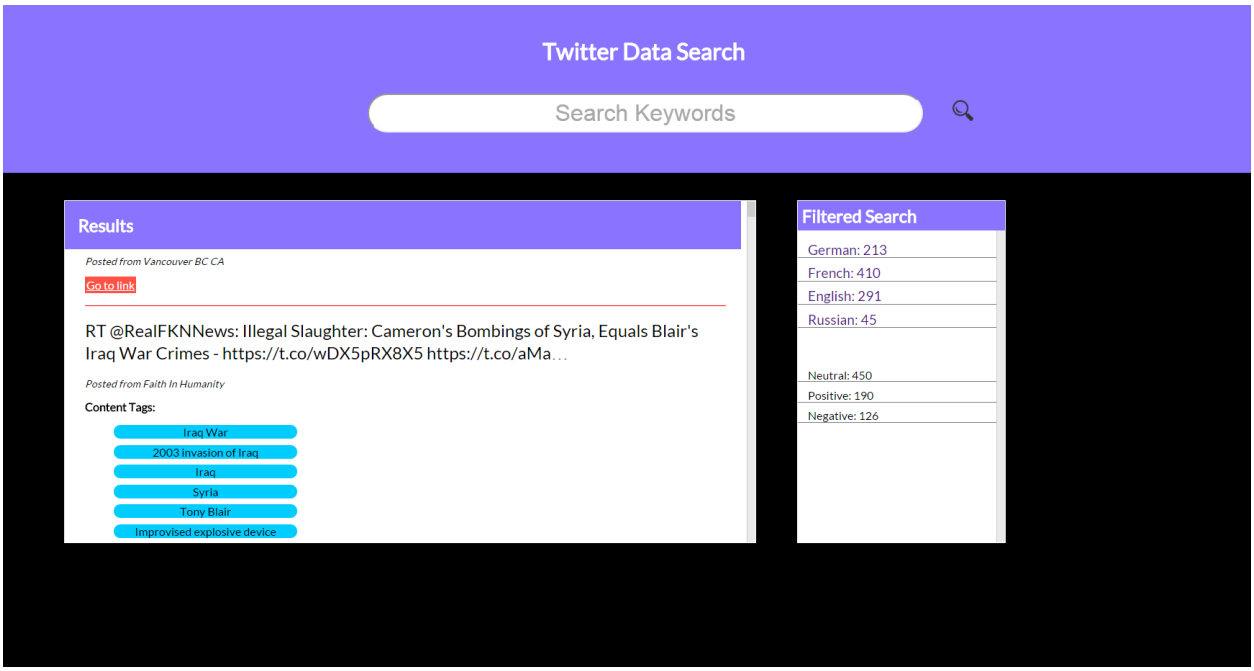
The UI was developed using HTML 5, javascript and jquery. There are two JSP pages called

a. Search

   The landing page for the website. It contains a search bar at the top.



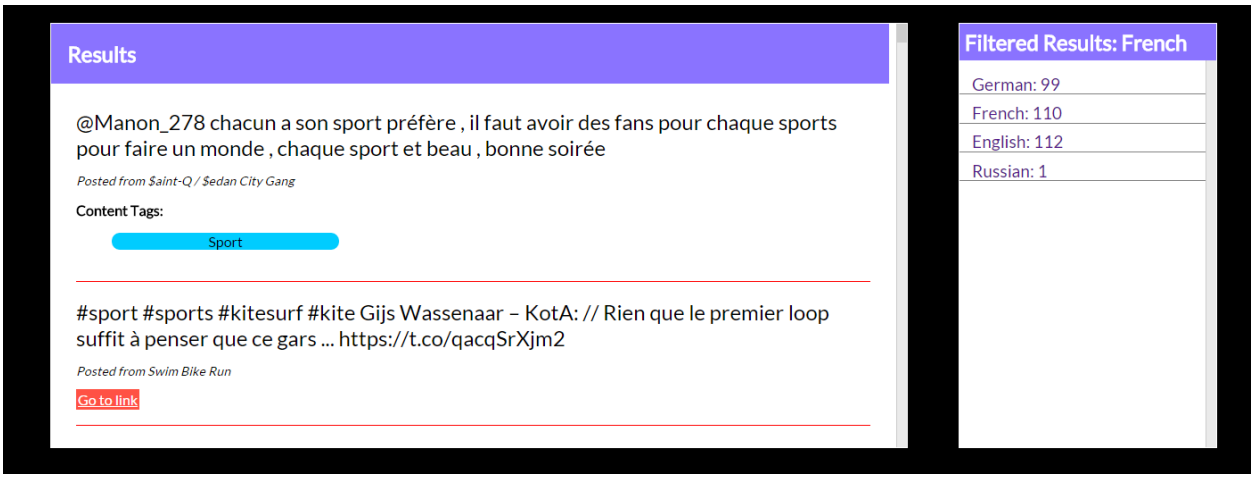
The search results are also displayed in the same page, once a search is made.



b. Filter

Used for displaying the results filtered by the facet.

Results filtered in French



Limitations and Future Work

We had attempted to implement summarization by obtaining relevant data for each tweet from DBPedia. This required the use of the Twinkle API and SPARQL query language. Due to errors in the query format we were not able to implement Summarization. This is one of the areas in which future work would be pursued.

Usage of Graphs to display the data from the Cross-Lingual analysis would also be a possible improvement.