

# 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>

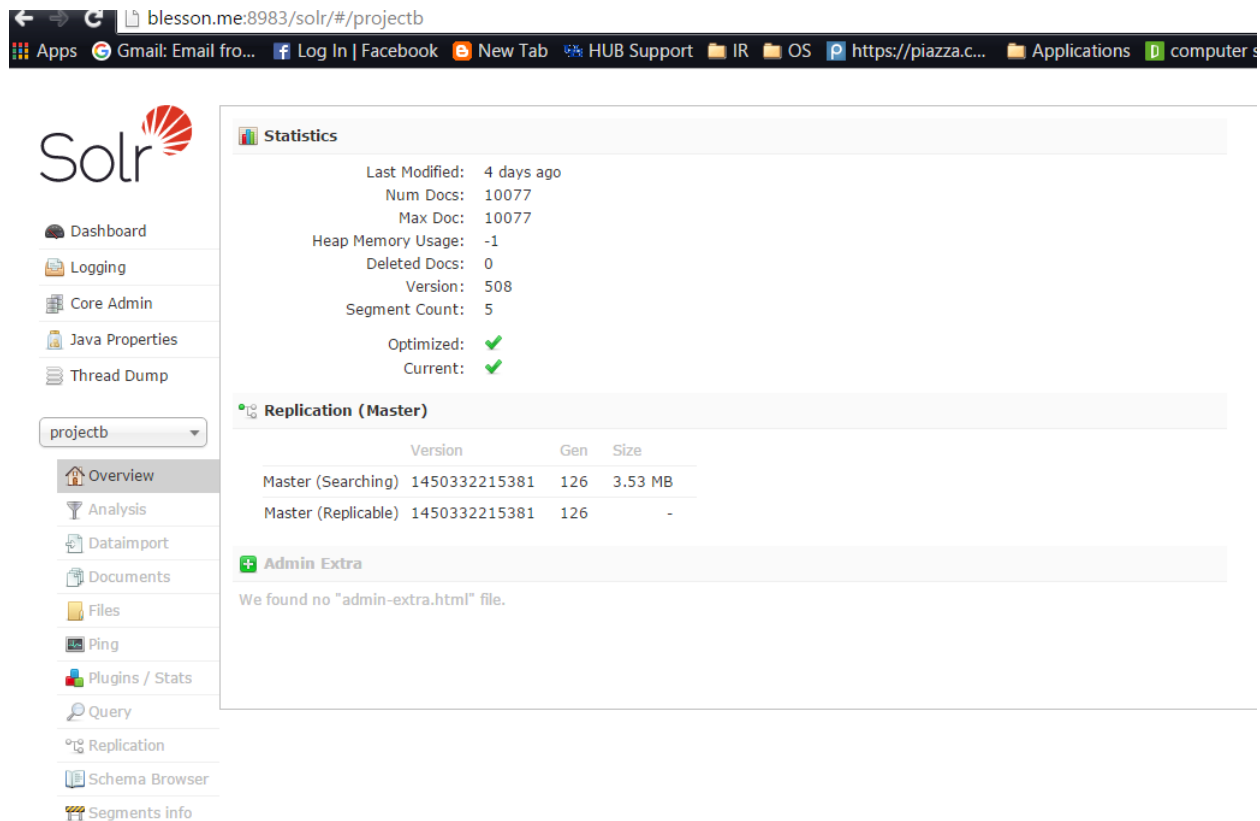
## Overview

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



The screenshot shows the Apache Solr Admin UI in a web browser. The browser's address bar displays the URL `blesson.me:8983/solr/#/projectb`. The left sidebar contains a navigation menu with the following items: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, a dropdown menu currently showing 'projectb', and a list of links: Overview (selected), Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query, Replication, Schema Browser, and Segments info.

The main content area is divided into two sections:

- Statistics**: This section displays various metrics for the 'projectb' core:
  - Last Modified: 4 days ago
  - Num Docs: 10077
  - Max Doc: 10077
  - Heap Memory Usage: -1
  - Deleted Docs: 0
  - Version: 508
  - Segment Count: 5
  - Optimized: ✓
  - Current: ✓
- Replication (Master)**: This section shows a table of replication status for the master node:

	Version	Gen	Size
Master (Searching)	1450332215381	126	3.53 MB
Master (Replicable)	1450332215381	126	-
- Admin Extra**: This section displays a message: "We found no 'admin-extra.html' file."

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.

J'aime une vidéo @YouTube de @watchmojo - Top 10 Star Wars Lightsaber Battles in Movies and TV (Quickie) <https://t.co/0M4xFY8KA0>

Content Tags:

Lightsaber

War

Luke Skywalker

Obi-Wan Kenobi

Star Wars

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

The screenshot displays a search results page with a purple header labeled 'Results'. It shows two tweet results. The first tweet is from @TylerD81, mentioning a national demonstration in London. The second tweet is from @RealFKNNews, discussing the Iraq War. Below the tweets, there are blue content tags: 'Iraq War' and '2003 Invasion of Iraq'. To the right, a sidebar titled 'Filtered Search' shows language facets: German (213), French (410), English (291), Russian (45), Neutral (450), Positive (190), and Negative (126).

Filtered Search	
German:	213
French:	410
English:	291
Russian:	45
Neutral:	450
Positive:	190
Negative:	126

### 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 screenshot displays a web interface with a tweet on the left and a sentiment analysis sidebar on the right. The tweet text is in French: '@Manon\_278 chacun a son sport préféré , il faut avoir des fans pour chaque sports pour faire un monde , chaque sport et beau , bonne soirée'. Below the text, it says 'Posted from Saint-Q / Sedan City Gang' and 'Content Tags: Sport' with a blue pill-shaped tag. At the bottom of the tweet area, there is a link: '[Sport News] | Player Ratings: Bayer Leverkusen 1-1 Barcelona https://t.co/efUln8wvik | Via Sports Mole'. The right sidebar shows sentiment counts: 'Neutral: 148', 'Positive: 134', and 'Negative: 5'.

Sentiment	Count
Neutral	148
Positive	134
Negative	5

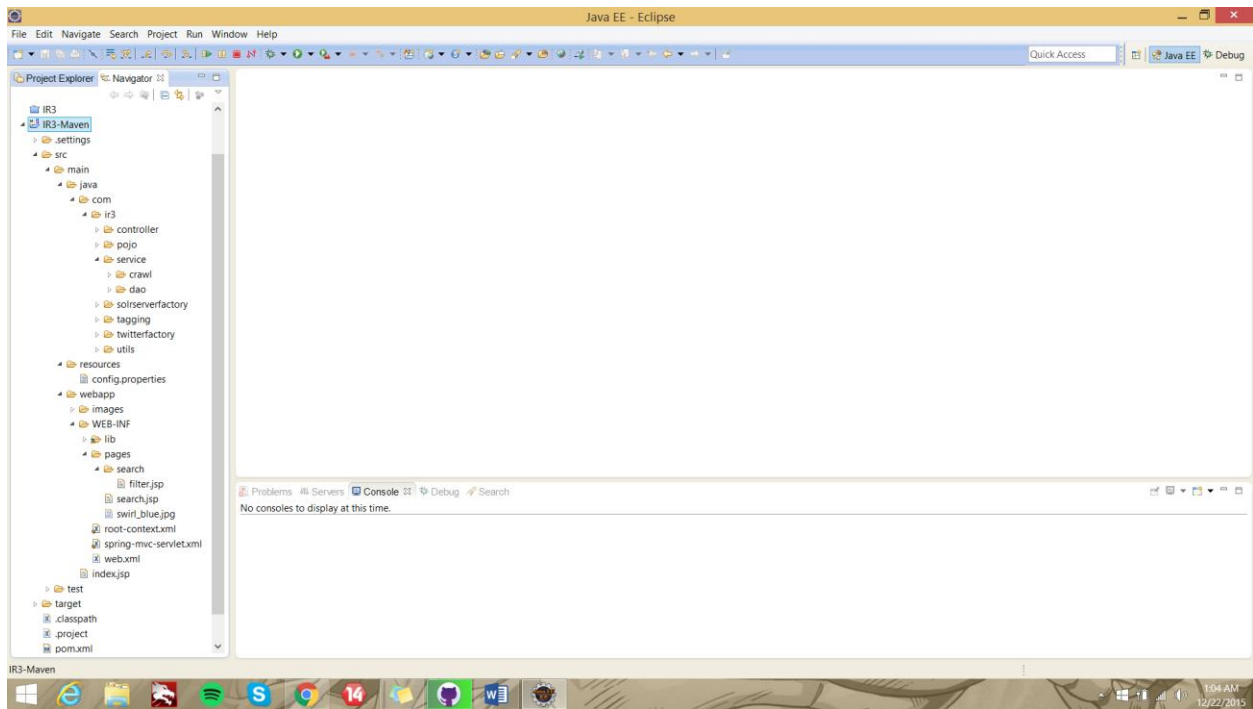
The Microsoft Bing Translator API was used for language translation. This also indirectly aids Cross-Lingual Retrieval in an indirect manner.

#### Back End

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:

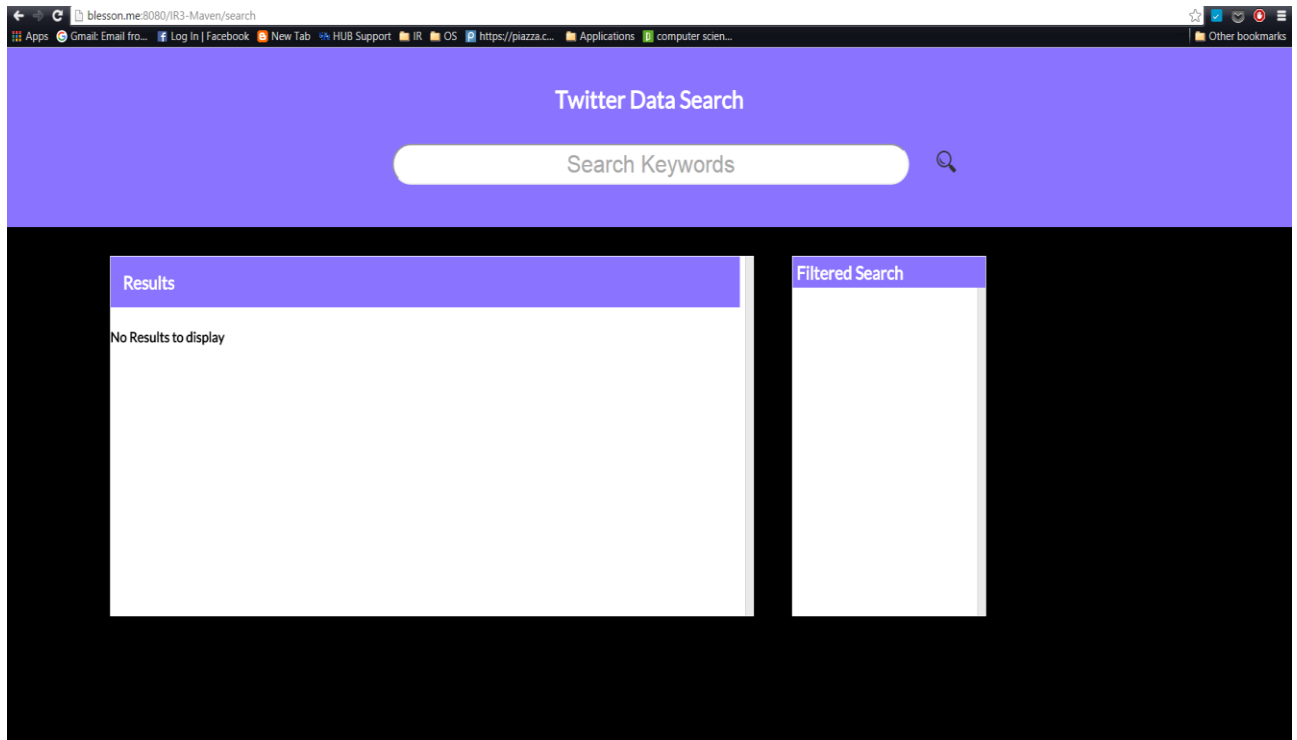


## Front End

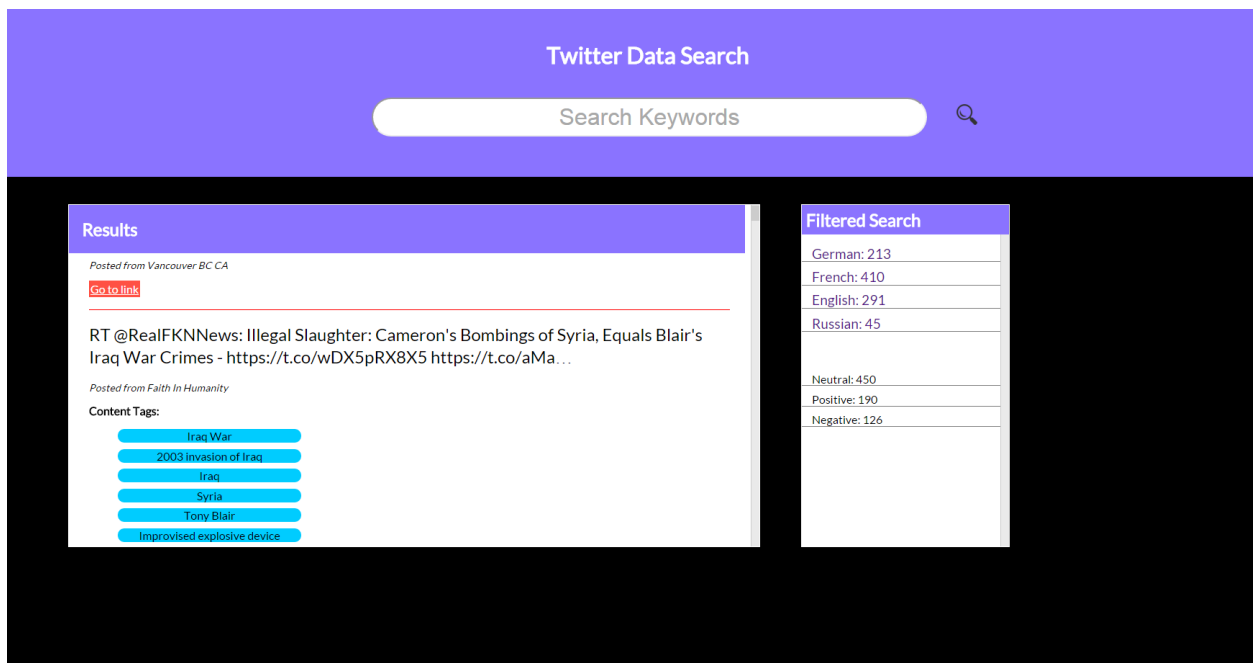
The UI was developed using HTML 5, javascript and jquery. There are two JSP pages called

- 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

The screenshot displays a search results interface with two main panels. The left panel, titled 'Results', shows two search results. The first result is from user @Manon\_278, with the text '@Manon\_278 chacun a son sport préfère , il faut avoir des fans pour chaque sports pour faire un monde , chaque sport et beau , bonne soirée'. It is tagged as 'Sport' and posted from 'Saint-Q / Sedan City Gang'. The second result is a tweet with hashtags #sport #sports #kitesurf #kite, mentioning 'Gijs Wassenaar – KotA: // Rien que le premier loop suffit à penser que ce gars ...' and a URL. It is posted from 'Swim Bike Run'. The right panel, titled 'Filtered Results: French', shows a table with language counts: German: 99, French: 110, English: 112, and Russian: 1.

Results	
@Manon_278	chacun a son sport préfère , il faut avoir des fans pour chaque sports pour faire un monde , chaque sport et beau , bonne soirée
Posted from	Saint-Q / Sedan City Gang
Content Tags:	Sport
#sport #sports #kitesurf #kite	Gijs Wassenaar – KotA: // Rien que le premier loop suffit à penser que ce gars ... https://t.co/qacqSrXjm2
Posted from	Swim Bike Run
Go to link	

Filtered Results: French	
German:	99
French:	110
English:	112
Russian:	1

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.