

CSE 535: Information Retrieval

PROJECT PART C

Team Pentatonics

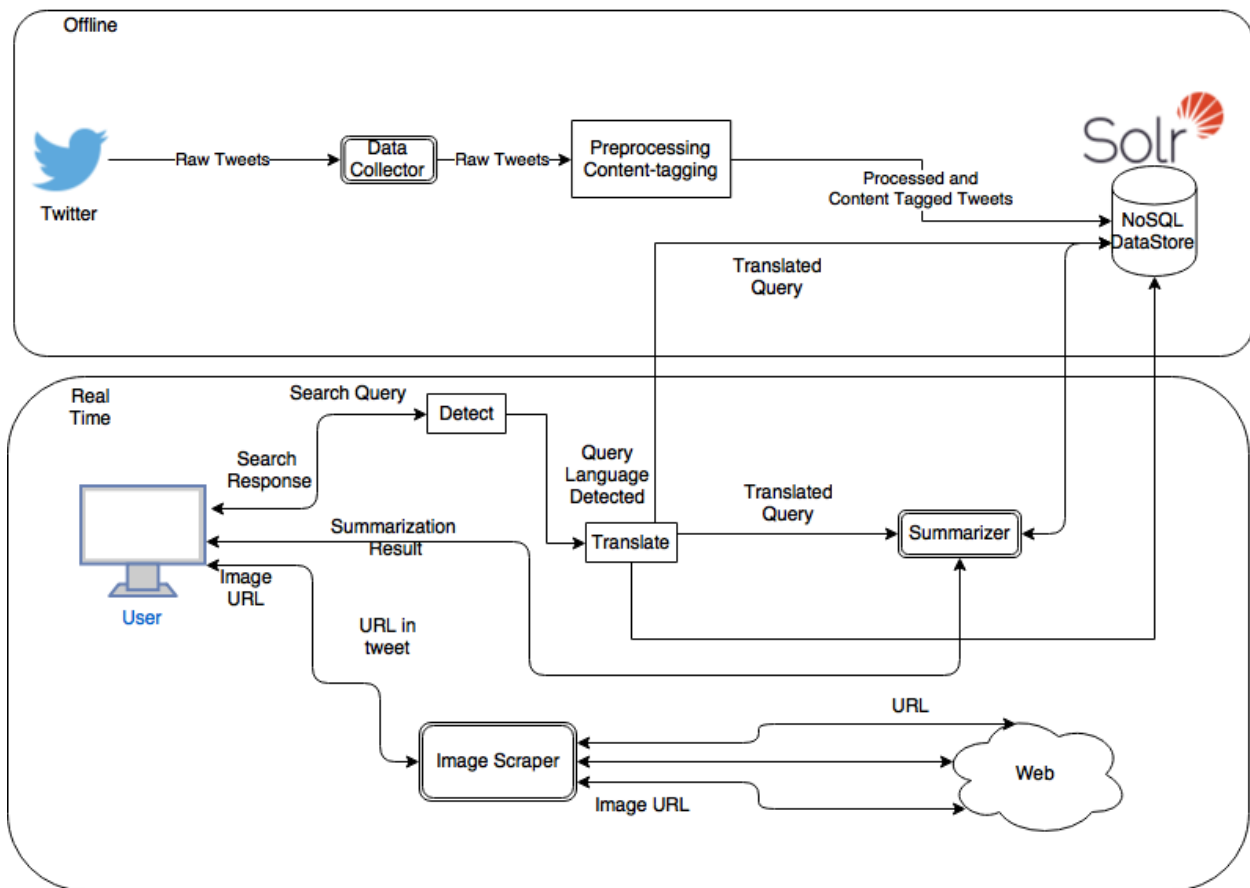
Mithun Atri	5017 0054
Abhay Dhundiraju Sastry	5016 9277
Sunny Tulsi Sreedhar Murthy	5017 0328
Fayaz Ahmed Vellore Farooque	5017 0331
Vasavi Manasa Chindalur Lakshminarayana Gupta	5017 0613

Overview

The goal of this project was to develop a multi-lingual faceted search system for social media data, in particular, Twitter. It also involved developing a front end that allowed users to search and browse the multi-lingual data based on various criteria such as language, location, hashtags etc.

We have built a custom Twitter search engine with a rich user interface capable of cross-lingual retrieval, providing a summary of the tweets and performing analytics to gain meaningful information. The UI also utilizes Open Graph protocol to fetch additional metadata, such as images/videos, from the websites mentioned in the tweets or from The Guardian, which is used to get the latest news based on the user query and the corresponding results.

The below diagram shows our system architecture:



Brief description of the architecture: The first step in our system is to pull tweets from Twitter using the Twitter API. The tweets that are pulled are pre-processed and content tagged using Alchemy API. The pre-processed tweets were indexed in Solr.

In real-time, when the user enters the query, the language of the query is detected and appropriate boosts are applied for the corresponding language fields. The same query is passed to the summarizer which queries Solr to fetch the search result and summarize the tweets. The

summarization result is then sent back to the UI. The URL from the tweet are sent to a web service which extracts the image from the URL using og-protocol. The results of all these calls are collected in an asynchronous fashion and displayed in the UI.

Data

We indexed Twitter data over a period of approximately 1 week and collected around 15,000 tweets across 5 different languages, namely English, German, Russian, French and Arabic. The topics included the Syrian refugee crisis, Paris attacks, San Bernardino shootings and the Paris Climate talks. This provided us a rich data set to work with.

Implementation

- Content Tagging

Service used: Alchemy

Input: "Text" field of tweets

*"RT @KimKaosDK: Anonymous declares December 11 'Isis Trolling Day'\n#Daeshbags
#Daesh #ISISTrollingDay <https://t.co/OSXqqxtUTC>"*

Output: JSON object with tagged content

```
{
  "concepts": [],
  "entities": [
    {
      "count": "1",
      "relevance": "0.01",
      "text": "#ISISTrollingDay",
      "type": "Hashtag"
    },
    {
      "count": "1",
      "relevance": "0.01",
      "text": "#Daeshbags",
      "type": "Hashtag"
    },
    {
      "count": "1",
      "relevance": "0.01",
      "text": "@KimKaosDK",
      "type": "TwitterHandle"
    },
    {
      "count": "1",
      "relevance": "0.01",
      "text": "#Daesh",
      "type": "Hashtag"
    }
  ],
  "keywords": [
    {
      "relevance": "0.954151",
      "text": "Anonymous declares"
    },
    {
      "relevance": "0.718812",
      "text": "Trolling"
    },
    {
      "relevance": "0.699331",
      "text": "Isis"
    },
    {
      "relevance": "0.463185",
      "text": "https://t.co/OSXqqxtUTC"
    }
  ],
  "language": "english",
  "status": "OK",
  "taxonomy": [
    {
      "label": "/business and industrial/manufacturing",
      "score": "0.6498"
    },
    {
      "label": "/society",
      "score": "0.41411"
    },
    {
      "confident": "no",
      "label": "/society/dating",
      "score": "0.297524"
    }
  ]
}
```

We implemented Content Tagging using Alchemy API. The text field of each tweet is given as parameter to the API which returns a JSON response with multiple fields such as entities such as Person, Country, City, and Organization. The content tagging results have been used for faceted search and keyword extraction based News results.

Note: Content tagging has been applied only for tweets in English and German since Alchemy did not support content tagging for the other languages that we indexed.

- Cross-lingual Retrieval

Service Used: Bing Translation API

Input: “Text” field of tweets

Output: Text translated in 4 other languages

We have implemented methods to fetch results in 5 languages using translation and detection with the Bing Translation API. During query, an API call is made to detect the language of the query entered by the user. Another API invocation is made to translate this query into the 4 other languages (if user entered in German, the query is translated into English, Arabic, French and Russian). An additional goal here is to display tweets in the original query language at the top.

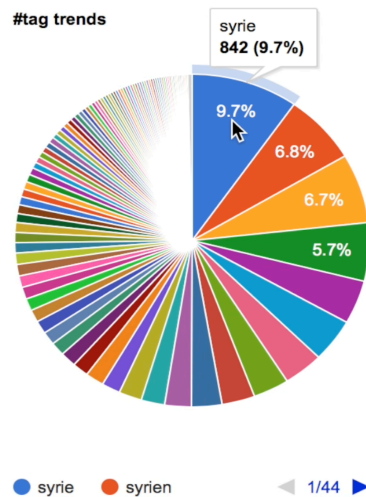
- Summarization

As mentioned earlier, Alchemy API generates a field called “keywords” for each tweet text. This generates the key terms in the tweet which when utilized effectively can be used to generate more information about the tweet topic. These keywords are used to query The Guardian news website to get the latest news on the related topic. If at all alchemy does not provide keywords, we take the user query as it is and invoke The Guardian API to fetch the news.

The *Representational Tweet section* which provides another type of summary. We used the LSA Summarizer provided in PyTLDR python module to summarize the tweets. The LSA Summarizer reduces the dimensionality of the article into several “topic” clusters using singular value decomposition, and selects the sentences that are most relevant to these topics. In the representational tweets for a given search result, top tweets are displayed. However in few cases the summarization achieved through this module does not really match the user’s expectation of a summary of the tweets. This is an area that can be explored further to provide better summarization.

- Analytics

We have performed some basic analytics on the tweets retrieved. The pie chart provides an indication of the top hashtags associated with the query. The pie chart is generated by leveraging Google charts API.



As you can see, we queried the Arabic translation of Syria and the corresponding pie chart is generated by counting the number of occurrences of a particular hashtag across all the documents. One can make some inference out of it. For this particular query we can conclude that “Syria” as a hashtag has appeared more in French language tweets than from Arabic. This is a result of the Cross-lingual retrieval which we showcased earlier.

- UI

The user interface of our search engine features a wide array of traits like facets, hashtag analytics and pagination. The UI was based on a MVC pattern in which the model and the controller were designed using AngularJS and the view using HTML and CSS. Bootstrap framework was used to assist us in building a quick and minimal UI. The default CSS file provided by Bootstrap was modified to suit our needs. Various AngularJS modules like ‘angular google charts’, ‘angular checklist model’, etc. were used to add facets, hashtag visualization and make the UI visually appealing. The ‘ng-repeat’ directive was widely used in this project to display data from lists. Our UI made many requests to various backend services and the response was made available to the user in a neat format.

- Solr

To obtain better search results, we used the ExtendedDisMax query parser with query boosting features. Below is an example query generated by our system which will be sent to Solr.

```
q=(obama)
&start=0
&rows=10
&fl=*,score
&defType=edismax
&qf= text_ar text_de text_en^5 text_fr text_ru
&bq=lang:en^30
&bq=(*:~ -RT)^10
&bf=recip(ms(NOW,created_at),3.16e-11,1,1)
&stopwords=true
&lowercaseOperators=true
```

```
&wt=json
&indent=true
&facet=true
&facet.field=lang
&facet.field=tweet_hashtags
&facet.field=Country
&facet.field=Person
&facet.field=date
&spellcheck=true
&spellcheck.collate=true
&spellcheck.q=obama
&spellcheck.maxCollations=1
&spellcheck.maxCollationTries=1
```

Explanations for the important fields in the request:

- ***“qf=text_ar text_de text_en^5 text_fr text_ru”***
The query term will be searched in fields text_ar, text_de, text_en, text_fr and text_ru with a boost given to the text_en field because the language of the query was identified as English. Suppose the query language was Arabic, then the field text_ar would get the boosting.
- ***“bq=lang:en^30&bq=(*: *-RT)^10”***
The ***“bq=lang:en”*** field boosts the documents whose language is English by a factor of 30. If the query language was Arabic, then documents whose language is “ar” would be boosted. We found that a factor of 30 seems to work well across all languages. The ***“bq=(*: *-RT)^10”*** field is used to rank the documents which are not retweets higher than the retweets. Since EDisMax does not provide a negation field, we boost tweets which do not contain “RT”.
- ***“bf=recip(ms(NOW,created_at),3.16e-11,1,1)”***
The boost function field is used to rank the recent tweets higher than the ones which were tweeted on later date.

Use Case

- *Simple query*
Our query here is “Syria”. In the center panel the search results are displayed. As you can see along with the search results, we are also displaying the metadata which is present in the external website mentioned in the tweet.

Pentatonics

Assad

Facets

Lang

- ☐ de 71
- ☐ en 30
- ☐ ru 25
- ☐ ar 21

Tweet Hashtags

- ☐ Assad 28
- ☐ ISIS 20
- ☐ syria 20
- ☐ syrieneinsatz 17
- ☐ cupra 16
- ☐ سوريا 13

Country

- ☐ Syria 29
- ☐ Turkey 9
- ☐ Russia 5
- ☐ Iran 4
- ☐ France 1
- ☐ russia 1

Person

- ☐ Assad 24
- ☐ Erdogan 9
- ☐ Bashar Al-Assad 8

Result

Total number of documents retrieved: 147

eu sollte nur mit usa die is+assad -aera beenden , putin ist kriegstreiber + verbrecher wie assad, genauso erdogan.
Date: 2015-12-03 @weppolivo3

Sister @TulsiGabbard you are so ethical in saying Assad should not be overthrow but ISIS is created by colonial https://t.co/ro04A00Ops

Tulsi Gabbard: Keep Assad in place, fight ISIS - CNNPolitics.com

News

- Europe must unite to deliver Syria from Isis and Assad | Guy Verhofstadt

Boris Johnson: allies should join Assad and Russia against Isis

Faceted Search

To let the user drill down into the search results, the UI has a provision for facets, which is a list of checkboxes, with values under categories *Languages*, *Tweet Hashtags*, *Country* and *Person*. The values under the facets *Country* and *Person* are obtained from the content tagging using Alchemy whereas the *Language* and *Tweet Hashtags* is part of the JSON obtained from Twitter. Checking any of the checkboxes will filter the search results by restricting them to the tweets containing the value for the chosen facet. The facet could be applied multiple times, wherein each filter step narrows down the search results and the facet values displayed under each category is refreshed based on the new search result obtained.

Pentatonics

التحريات سوريا

Facets

Lang

- ☐ fr 208

Tweet Hashtags

- ☒ daech 208
- ☒ syrie 208
- ☐ el 142
- ☐ daech 134
- ☐ sassoufi 117
- ☐ etatislamique 108

Country

Person

Date

- ☐ 2015-12-02 11
- ☐ 2015-12-01 44
- ☐ 2015-12-03 41
- ☐ 2015-11-27 11
- ☐ 2015-11-25 10
- ☐ 2015-11-28 10

Result

tweet_hashtags:syrie tweet_hashtags:daech

Total number of documents retrieved: 208

#Sassoufi RT Breaking3zero: En 12 uns, l'entrée en guerre de la Grande-Bretagne contre #Daech en #Syrie. (jutto...
https://t.co/CcfeTDqenY
Date: 2015-12-03 @the_sanghe

#Sassoufi Vidéo: Erdogan humilié, Obama confirme les propos de Poutine #Turquie #Daech #Syrie https://t.co/57njPUPjYM
Date: 2015-12-02 @the_sanghe

#Sassoufi RT spuznik_fr: La #Russie porte des frappes aériennes contre la raffinerie de pétrole de #Daech #Syrie ...
https://t.co/ZuuTdm72pH
Date: 2015-12-02 @the_sanghe

Summary

News

- سوريا بين مليشيات: انكزى حصة الجند | وسانج خاهر

the guardian

- ملقا تحارب داعش

- *Cross lingual retrieval*

Here, we have queried for the term Syria in Arabic (سوريا). The figure (1a) displays the initial set of results in Arabic. By making use of faceted search, we then selected the language as “en” which retrieved the results as shows in figure (1b).

Pentatonics

الضربات سوريا

Facets

Lang

- ☐ de 1176
- ☐ fr 1065
- ☐ en 1057
- ☐ ru 565
- ☐ ar 137

Tweet Hashtags

- ☐ syria 841
- ☐ syrien 585
- ☐ sassoufit 577
- ☐ isis 491
- ☐ syria 366
- ☐ daesh 363

Country

- ☐ Syria 423
- ☐ RUSSIA 102
- ☐ Russia 99
- ☐ U.S. 96
- ☐ Iraq 52
- ☐ Iran 28

Person

- ☐ HOLLANDE 102
- ☐ PUTIN 102

Result

Total number of documents retrieved: 4000

RT @AhmadTalk: بسعد صباحين ليلتي الجيش السوري الحر. أمنا ونصرنا #سوريا #الجهنم_الحر_أنت_الأمل #Syria https://t.co/FSQIVtC...
Date: 2015-12-13 @etisalat1

RT @HamadMo7amad: ادعوهم بما توجد به أنفكم شاء #سوريا اللهم إنيهم مظلومون #Syria http://...
Date: 2015-12-06 @vfh20135

جديد مواقع الرياض: بريطانيا تثنى أولى ضرباتها على مواقع داعش في سوريا
https://t.co/PcgHixPxc #saudi_enews
Date: 2015-12-09 @saudi_enews

جديد سبق: بريطانيا تثنى أولى ضرباتها الجوية ضد داعش في سوريا
https://t.co/7nc780oPl6 #saudi_enews
Date: 2015-12-09 @saudi_enews

1a

Pentatonics

الضربات سوريا

Facets

Lang

- ☒ en 1057

Tweet Hashtags

- ☐ syria 291
- ☐ isis 235
- ☐ daesh 155
- ☐ messedup 102
- ☐ isil 101
- ☐ turkey 67

Country

- ☐ Syria 399
- ☐ RUSSIA 102
- ☐ Russia 98
- ☐ U.S. 96
- ☐ Iraq 52
- ☐ Iran 28

Person

- ☐ HOLLANDE 102
- ☐ PUTIN 102
- ☐ Putin 86
- ☐ Obama 21
- ☐ Assad 18
- ☐ Erdogan 16

Date

Result


langen

Total number of documents retrieved: 1057

RT @THE_47th: After over 2mnths & 4K "targeted" strikes, Russia's Min of Def says ISIS is "expanding" & now controls 70% of Syria https://t...
Date: 2015-12-12 @mememo

RT @THE_47th: After over 2mnths & 4K "targeted" strikes, Russia's Min of Def says ISIS is "expanding" & now controls 70% of Syria https://t...

Islamic State militants seize 70% of Syria's territory & Russian defense minister



1b

- *Spellcheck*

Solr provides an excellent feature of spellchecking which we utilized to aid the user when the query is misspelt. Here we misspelt a multi-word query. As you can see the spellchecker displays the correct query which the user can click on to retrieve the results. It worked very well with queries across different languages. Here, we are trying to query air strikes in French with an incorrect spelling.

frapp|arienne

Facets

Lang

- ☐ en 141
- ☐ de 33
- ☐ fr 14
- ☐ ru 7

Tweet Hashtags

- ☐ erbil 16
- ☐ germany 16
- ☐ kurd 16
- ☐ peshmerga 16
- ☐ syria 16

Result

Did you mean **frappe aérienne** ?

Total number of documents retrieved: 195

@ja_gabon cérémonie d'ouverture Company of the year 2015
Date: 2015-12-03 @The_Blessed1996

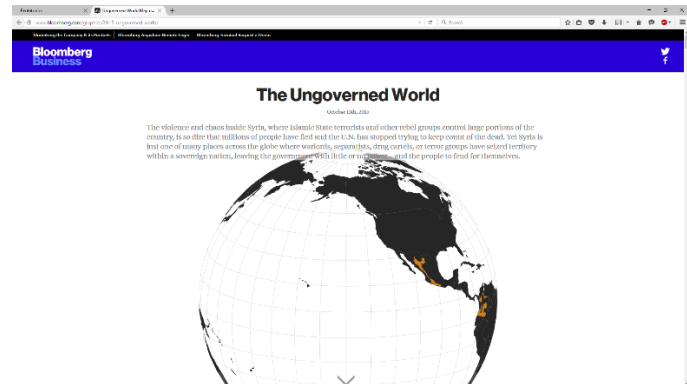
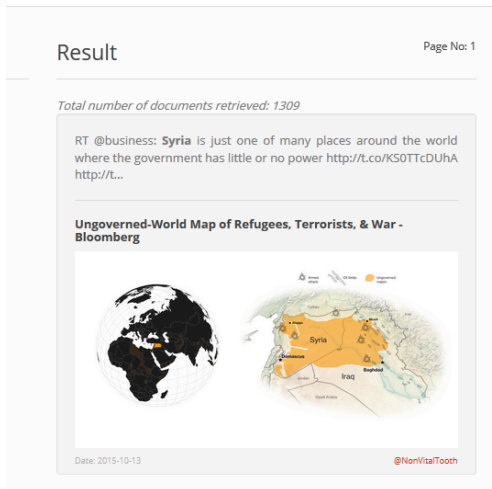
f he'd cursed the Prophet, they would have killed him." 2/2
https://t.co/lAq74ybcv2 Quel est ce prophète plus haut que son dieu...#Blasphème

Summary

News

Representational Tweets

- *Accessing external website by clicking on images/title displayed*
Upon clicking the image displayed, we are redirected to the news article/website.



Limitations

Content Tagging is available only for English and German languages. Hence, faceted search could not be implemented for tweets in other languages.

At this point of time, we rely on Bing Translation API for translation. Hence, the accuracy of the search result depends on the accuracy of the translation API.

Because the tweets are of only 140 characters, the summarization does not always give very good results which match the user expectations.

Scope for Improvement

- We can add richer Summarization Snippets from Wikipedia or any other news site.
- We noticed the BING Translation API takes a long time to translate the query terms which can be reduced if we use an internal translator and also enhance the accuracy of the translation.
- Identifying retweets using better methods during preprocessing.
- At present, the content tagging is available only for English and German languages, we can investigate more to achieve content tagging in other languages.
- Query auto completion when the user is entering the search query is a nice feature to have.
- Semantic search can be implemented to provide better search results.
- We can add more analytics with better visual representations and identify relationship between topics and hashtags etc.