Big Data Visualization

Project 1: Multilingual Sentiment Analysis (Natural Language Processing (NLP))

The multilingual sentiment analysis revealed the actual results based on the region and the language spoken in that region. As part of this research the system was implemented in R-language,

Source code https://github.com/mimm1/Sentiment-Analysis

They accessed twitter data using TwitterAPI using the keywords “women driving in Saudi” in both Arabic and English languages for a period of 10 days. A negation vector was generated that maintains the score of each tweet by comparing it against a provided list of positive and negative words in both the languages. Positive and negative words were translated using online Translate API that were used to score each of the tweets. The analysis showed average “Negative” sentiments for “women driving in Saudi” were 4% with 6.3% and 1.2% for tweets in Arabic and English respectively. The “Positive” sentiments were 32.1% with 36.1% for tweets in English and 27.1% in Arabic. However, 63.8% sentiments in both the languages were classified as neutral**.**

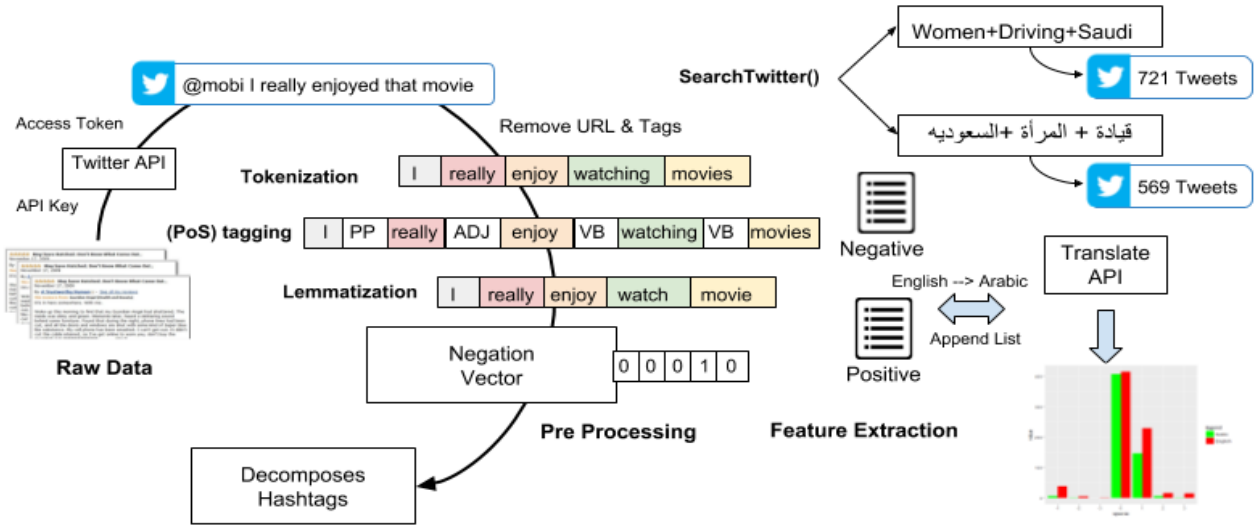


Figure . Lexical approach for Sentiment Analysis

Project 2: Crowd Management (Online Clustering)

Crowd control is another Big Data application useful for police, security and emergency service. OpenCelliD is Big Data collection of Cell Towers and WiFi Aps consists of around 39 million rows, this database is built by the community and is free available for download . Draggable pushpins were dynamically created on the map by retrieving the Cell Towers database from the SQL Server containing location code, range, network code as well as the latitude and longitude of the towers.

Source code: https://github.com/mimm1/OpenCell

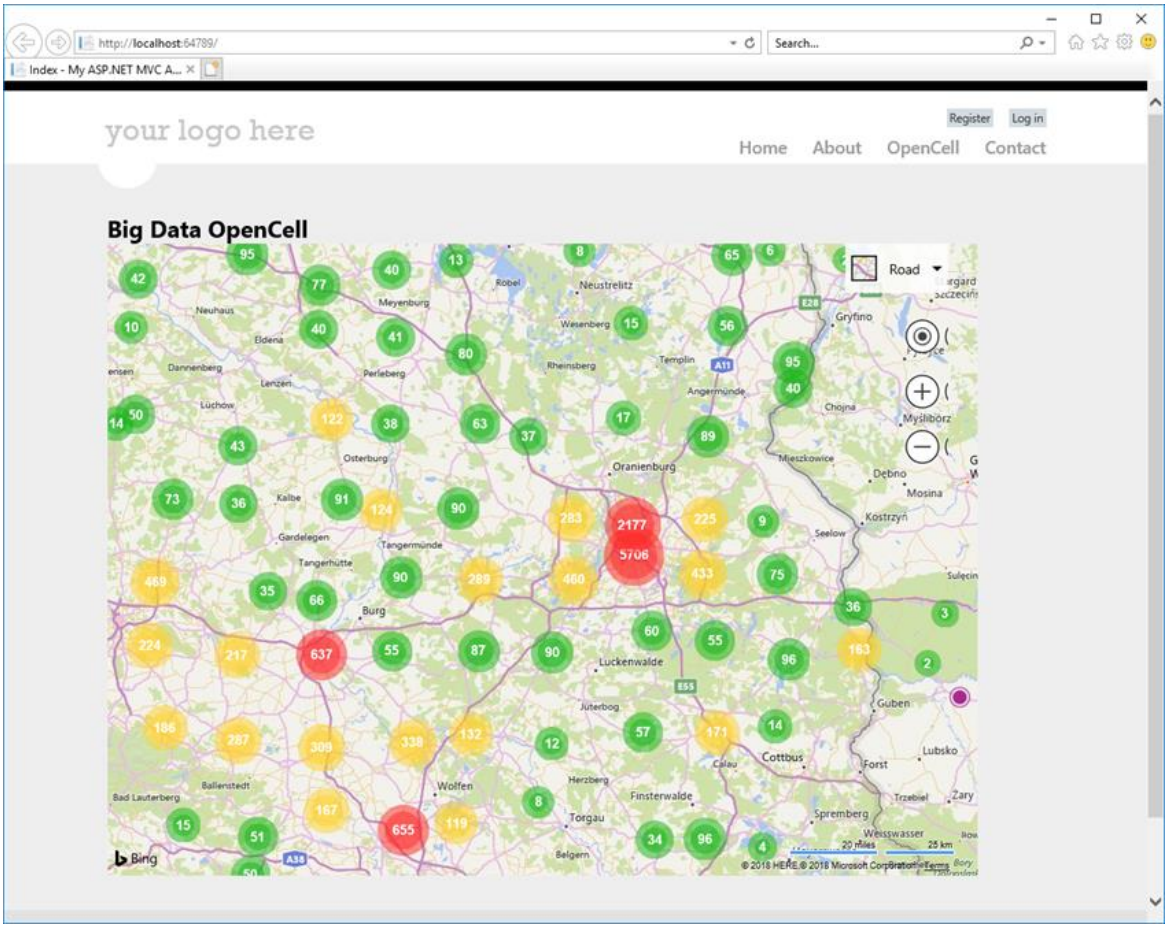


Figure . Clustering of OpenCellid Towers

Partial views were designed to create, edit, and view the geographic locations of the Cell towers. The interface is developed in CSharp using Model-View-Controller to create interfaces. The Bing Map is integrated in the application using Asynchronous JavaScript and XML (AJAX) and the JavaScript Object Notation (JSON) format has been used to serialize and transmit the structured data. The clustering was achieved by creating an instance of the ClusterLayer class and passing the pushpins data from the SQL Server for clustering and updating into the map.

Project 3: Climate Science (Online Machine Learning)

Climate science is the study of the planet's environment. In the prediction system the input data may be received, changes in the sea level, atmospheric composition, hydrological cycle and changes in the land surface. The training data usually covers decades or centuries, although a shorter time scale such as weather prediction may include days or weeks. As shown in Figure , the learning system must be capable of detecting concept drifts and respond to these drifts by automating its configurations based on the seasonal drifts.

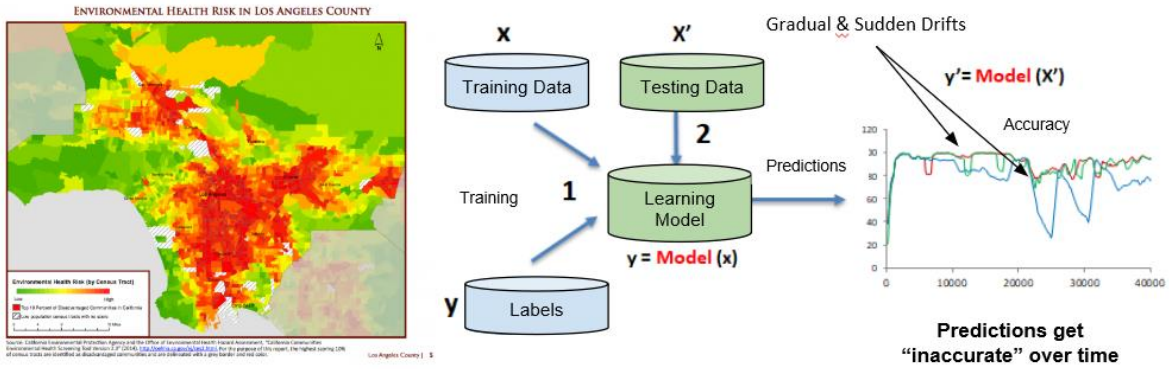


Figure . Environmental Monitoring and weather prediction

Project 4: Gene Ontology Mining

Gene Ontology (GO) is a framework for the modeling of biology. The GO defines concepts or classes used to describe gene functions and relationships between these concepts. Figure shows the class hierarchy of gene ontology in Protégé Software.

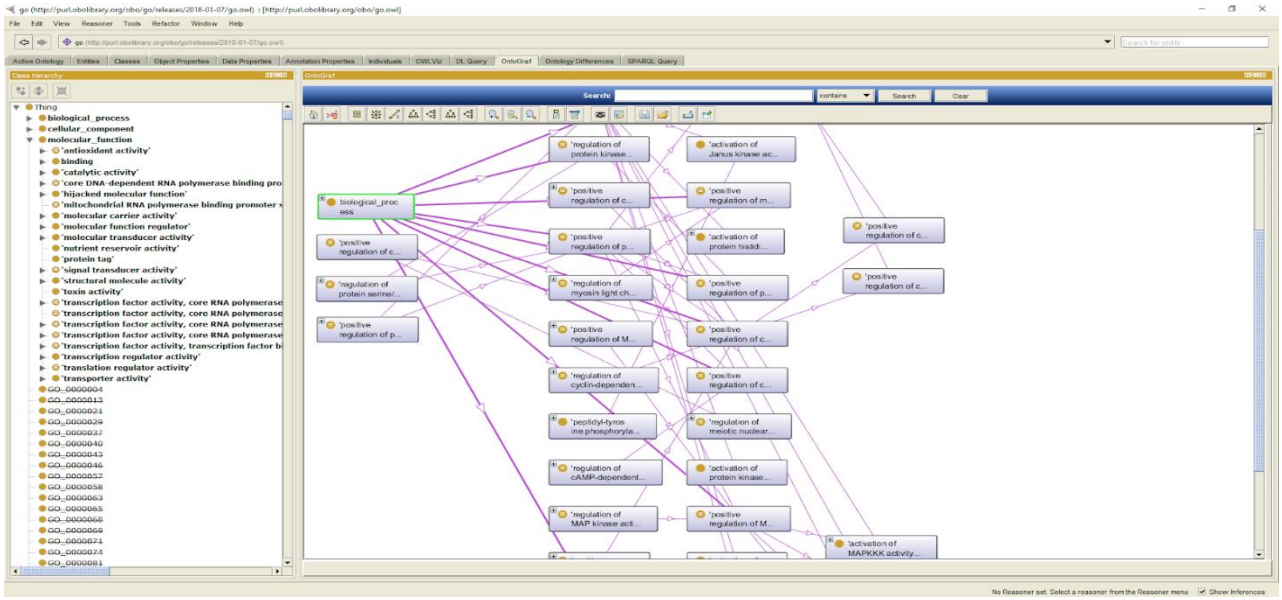


Figure . Visualization of Gene Ontology in Protégé

Ontologies assists in understanding the relations of massive amount of data it generates, such as Semantic Web to provide uniform access to resources and structuring the information in machine readable format. By annotating experiments and literature with Gene Ontology terms, researchers are able to integrate results and gain insight about relations that were previously difficult to discover. Many software applications can relate differential gene expression to functions using GO enrichment analysis. The encouraging results obtained that provide a first evidence of the potential of deep learning techniques towards long term ontology learning.

Project 5: Genomic Medicine

Big Data analytics assists the genomic medicine in diagnostic and prevention of genetic diseases and foreincis. Deoxyribonucleic acid (DNA) fingerprinting was invented at the University of Leicester by Professor Sir Alec Jeffreys in 1984.

Used 'Windows_FinchTV' program, the data in (Step 2) sequence Figure represents the detection of fragments differing by one base pair in size and the gray bars behind the lines reflects the quality of the base call, the low bars represents poor sequence quality and is completely unusable. The letters on the top of there window present the base called by the software at each nucleotide location.

Figure. DNA Sequencing

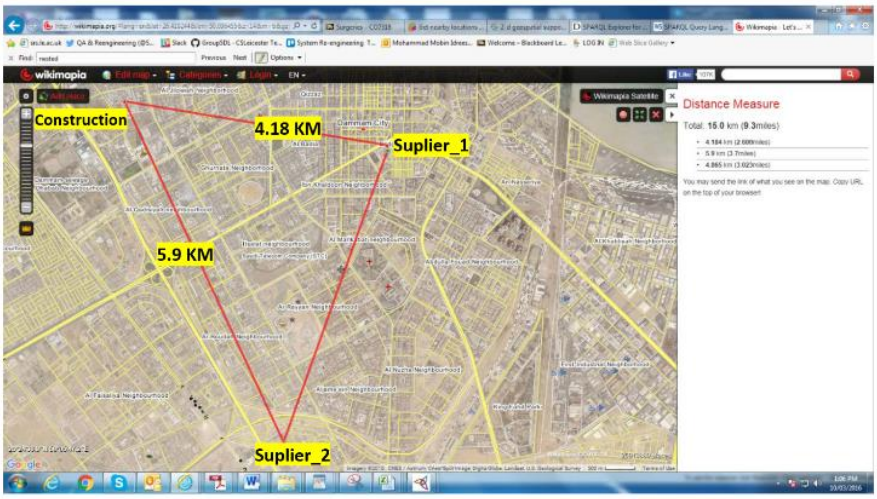
A good quality sequence lines should not overlap and have distinct line peaks representing well separated fragments. The sequences needs to be blasted (Step 3) to determine if the sequences are representative of the gene you intended to analyze. Step 4 shows the results of Nucleotide blast in which the input sequence matches 98% of the base calls with score 34,455 indicating the gene as human.

Project 6: Ontology on GIS Data

This projects is about Ontology Engineering and SPARQL using GIS information.

Source Code: https://github.com/mimm1/ontology

Supplier	Material	Latitude	Longitude	distance
University_of_Leicester		"26.434687"^^	"50.059977"^^	"0.000000000000"
Supplier_1	Nano_Concrete	"26.430152"^^	"50.100145"^^	"0.001634034449"
Supplier_2	Nano_Concrete	"26.387333"^^	"50.090790"^^	"0.003191842285"



Query 2: What material is best for a particular kind of construction?

```
SELECT  DISTINCT ?Construction_Type ?Material ?toxic_level
WHERE  {
            ?Construction_Type  a scm:Residential ;
                                    scm:usingMaterial ?Material .
            ?Material scm:hasToxicLevel ?toxic_level .
             FILTER (?toxic_level <=3) .
             OPTIONAL { ?Material  scm:ManufacturedBy ?m .
             }
}
```

Figure. Ontology on SPARQL

Project 7: Visualizing Airline Routes

Visualization of worldwide flight routes network as connected graph of airport dataset created in Gephi. The directed graph is using 5,623 nodes(airports) and 37,596 edges(routes) presented as a mixture overlay between network graph and geographic data using the Geo Layout plugin.

Source code: https://github.com/mimm1/BigData-Visualisation/blob/master/airlines.png

Figure . Visualizing Airline Routes Network using Gephi

Project 8: Social Network Analytics

In social network analytics explored the information within the Panama Papers. Figure shows the connections of Saudi Aramco Energy Ventures using the CrunchBase database. It is a graph of companies linked to market sectors, cities, and investors.

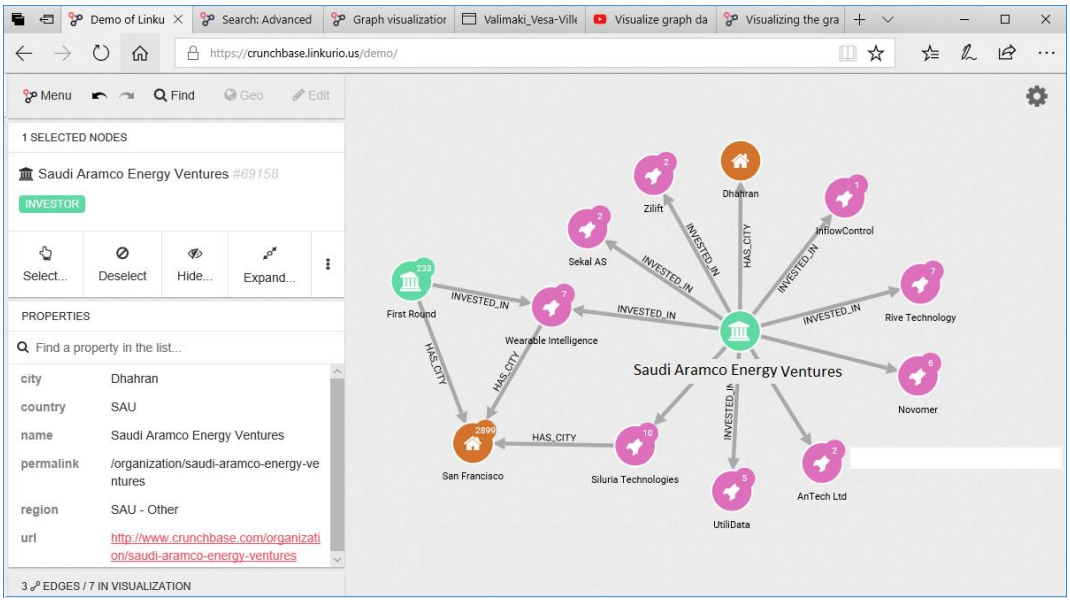


Figure . Exploring Panama Papers connection using Linkurious