CS 6350 Project Spring 2017 – Project Proposal

Topic

Create a Structured database for multilingual named entity resource to identify different political actors in multiple languages (primarily in English, Arabic and Spanish)

The project mainly comprises of mining the worldwide news press articles to identify the geographies that are sensitive to social unrest, agitations and campaigns. This is accomplished by working on multiple datasets to gather information on the various political actors regarding whom the society is alarmed and alerted. The motivation behind the project is to apply data resources and methods to help make data-driven decisions about foreign policy, civil war prevention, human rights policies, and the effects of other factors such as environmental or economic policies on these phenomena. The core of the project is to use data for the purposes of decision making which is not constrained by the language.

Team Member Details

#	Name	E-mail
1	Subhasis Dutta	sxd150830@utdallas.edu
2	Chandrika Cherukuri	cxc161530@utdallas.edu
3	Sneha Ramesh Neranki	snr150130@utdallas.edu
4	Jayprakash Rout	jxr152730@utdallas.edu
5	Priyanka Chadalavada	pxc162330@utdallas.edu

Proposal

The goal of our project is to ingest unstructured data from multiple data source and convert it into a coherent structured data stored and indexed in a database so that it can be searched by REST API calls. As the data sources are unstructured and incomplete we will be using map reduce to do different join operations. Also to identify unidentified labels we will use different machine learning classification techniques to identify the language given a person's name.

Some of the objectives are:

1. Using Map Reduce Join a many-to-many relation between different data source

When a user searches for a particular person in a web interface we will retrieve the data and visualize the data and expose APIs for other systems to consume. Based on user search for a person or organization we need to retrieve his/her name in different languages especially Spanish and Arabic. Along with this data, we are trying to combine data from CAMEO to get more details about that person. To combine data from CAMEO and JRC datasets, if we write the batch code to generate the data, it consumes a lot of time. So, we develop the Map Reduce relation between Cameo and JRC relations.

2. Using BableNet Knowledge Base identify political actors and their name variations

Using BableNet's data corpus try to identify the political actors and their name variants in different languages.

3. Using Machine Learning Identify the language of unidentified entity in the JRC Named Entity dataset

One of the main milestones of the implementation is to extract the different variants of the names of the political actors in multiple languages. We collate this information with the other available data sources to extract credentials and their political significance I am studying about how to utilize the Naive Bayes approach to perform text classification for multi class problems. My project would mainly encompass achieving accurate language classification using Naive Bayes N-gram approach.

One strategy for dealing with continuous data in naive Bayes classification would be to discretize the features and form distinct categories or to use a Gaussian kernel to calculate the class-conditional probabilities. Under the assumption that the probability distributions of the features follow a normal (Gaussian) distribution. Text classification is a typical case of categorical data, however, naive Bayes can also be used on continuous data.

Being an eager learner, naive Bayes classifiers are known to be relatively fast in classifying new instances. Eager learners are learning algorithms that learn a model from a training dataset as soon as the data becomes available. Once the model is learned, the training data does not have to be re-evaluated in order to make a new prediction. In case of eager learners, the computationally most expensive step is the model building step whereas the classification of new instances is relatively fast.

A lot of work is done on Text Classification using N-gram approach in Python using the nltk library.

4. User Interface to search the Named Entity dataset and display the name variation

An easy to use search interface which will make call to a REST API get the results and display it and cache the results.

Data Source

1. JRC-Named Entity Dataset - is a highly multilingual named entity resource for person and organization names. It consists of large lists of names and their different spelling variants including across scripts. JRC-Names is a technical resource that can be used to find names even if they are spelled differently, but it is also a useful ingredient for IT systems that process text, e.g. for text mining.
2. CAMEO Dataset - a framework designed to categorize all types of political interactions, rather than a limited repertoire of actions.
3. BableNet Data and API – An open source knowledge base with data of entity and their name in different languages.
4. dbPedia API – API interface to Wikipedia.

Technology Stack

1. Python – Primary coding language
2. Java – For map reduce coding
3. MongoDB – for storing the final structured data
4. Hadoop – For Map reduce and storing the flat file data
5. Spark – Use for machine learning implementation
6. MEAN stack – For building the User Interface.

Code Base

https://github.com/SubhasisDutta/CAMEO-JRC-Database