CS 6350 Big Data Analytics and Management

Spring 2018

Project Proposal

Members

Anagha Asok : axa151631

Sashidhar Donthiri: sxd173730

Ravikiran Kolanpaka : rxk171530

Santhosh Medide: sxm174930

Sathya Pooja Rami Reddy : sxr176830

Statement

The implementation is based on the IEEE paper “Spark-based political event coding”. The project aims at retrieving new events and actors in the form of “Who-did-what-to-whom” format using a distributed framework like Apache Spark. The attempt is to develop framework that will integrate both CoreNLP and PETRARCH in a parallel manner and perform the event coding in close to real time.

Related Literature

- IEEE Paper “ Spark-based political event coding”
- https://github.com/openeventdata/petrarch2”
- https://stanfordnlp.github.io/CoreNLP/”
- https://github.com/openeventdata/scraper/blob/master/whitelist_urls.csv”

Tentative Method

- Crawling the data form the list of URLs provided.
- Convert the data into xml format compatible with StanfordCoreNLP.
- Pass the meta data to StanfordCoreNLP to extract parse trees, named entities, lemmas and sentiment using Apache Spark.
- Store the output in MongoDB.
- Parse all the BSON data from MongoDB to PETRARCH to extract political events in the form of “who-did-what-to-whom”.

Programming Environment

- Parser, StanfordCoreNLP, PETRARCH == Python Environment
- Storage Database -MongoDB

Tentative Schedule

March 22 – March 31

- Study the literature
- Crawl and preprocess the Data

April 1 – April 14

- Complete literature study
- Research about libraries and programming environment to be used
- Implement the primitive working algorithm

April 15 – April 29

- Implement the fully working algorithm
- Test it for various inputs and debug
-

By May 4

- Finish the project and make the report
- Make the presentation.