Clean Coders

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

Team Project Proposal

Semantic web



Table of Contents

[Group Information 2](#_Toc432868409)

[Details of team members 2](#_Toc432868410)

[Summary 3](#_Toc432868411)

[Data 4](#_Toc432868412)

[Tools and technologies used 5](#_Toc432868413)

[Schedule 6](#_Toc432868414)

[Expected Results 7](#_Toc432868415)

# Group Information

Group Name: Clean Coders

Project Name: Relation extraction using semantic web technologies

Team Members: 4

## Details of team members

1. Ekal Golas – exg140230
2. Navaneeth Rao – nbv140130 (In a different section - 502)
3. Tejaswini Sirlapu – txs142130
4. Prashanth Govindaraj – pxg142030

# Summary

This project proposal describes the high-level overview of the implementation of the semantic web team project. This project is about relation extraction from unstructured business data using semantic web technologies.

It explains the various technologies used in this project and various features that will be implemented to design this application.

This proposal also contains the description of the data to be used and the basic schedule for the project work with milestones.

Lastly, the proposal also discusses about expected results of the project.

# Data

The data is described as:-

1. A collection of web pages, indexed and stored in SolR
2. Data will be about people, organizations, locations and related financial information
3. Initially we take 100MB of test data, and then scale out using [technologies](#_Tools_and_technologies) discussed
4. Data will be available publicly on the internet (Restricted to preferred domains). Example: News article about a company in “businessinsider.com”, a company`s Wikipedia page etc.

# Tools and technologies used

The tools are technologies we plan to use as discussed as:-

1. Apache SolR – For indexing and storing data related to web pages
2. Apache Spark (Maybe) – For streaming large amount of data
3. Hadoop/MapReduce – For batch processing of data, if spark does not work out
4. Stanford NLP – For relation extraction in the unstructured data
5. Apache Nutch (Maybe) – For web crawling
6. Java – As the programming language
7. Apache Jena – For defining ontologies and creating graphs

This project will be hosted on GitHub for any references.

# Decision Criteria

Following were the alternatives considered and rejected due to reasons mentioned below:-

1. Using social network as data source – This includes gathering data from Twitter API etc. which is not much informative and grammatically correct
2. Visualization related to relations between location and people – Less useful than visualizing and presenting results on companies and people which can help make better sense to make business decisions
3. Using a single system architecture – This is clearly insufficient to handle large amount of data and gather meaningful results on it.

# Schedule

Following are the milestones in the project and their estimates:-

1. Oct 25, 2015 – Coding the extractors and testing on some limited sample dataset
2. Oct 30, 2015 – Project checkpoint report complete
3. Nov 1, 2015 – Describing the ontologies and structure of data by using Jena.
4. Nov 8, 2015 – Use of SPARQL to derive results from data and present them in a user-friendly manner
5. Nov 13, 2015 – Exploring the additional technologies (Ones discussed in Maybe) to extend the project to a better real world scenario and compare the results to sample data
6. Nov 20, 2015 – Project complete, final report and bug fixes if any.

# Expected Results

Following will be the expected results of this project:-

1. Getting an insight as to how people, organization and locations are related to each other.
2. Possible financial information related to these results
3. Visualization of this data as a graph

These results will be interesting because scope of this project can help marketing teams make better sales/business decision based on extracted relationships and structures data visualizations. Even casual users might benefit by gather general knowledge from this data about companies and famous personalities.