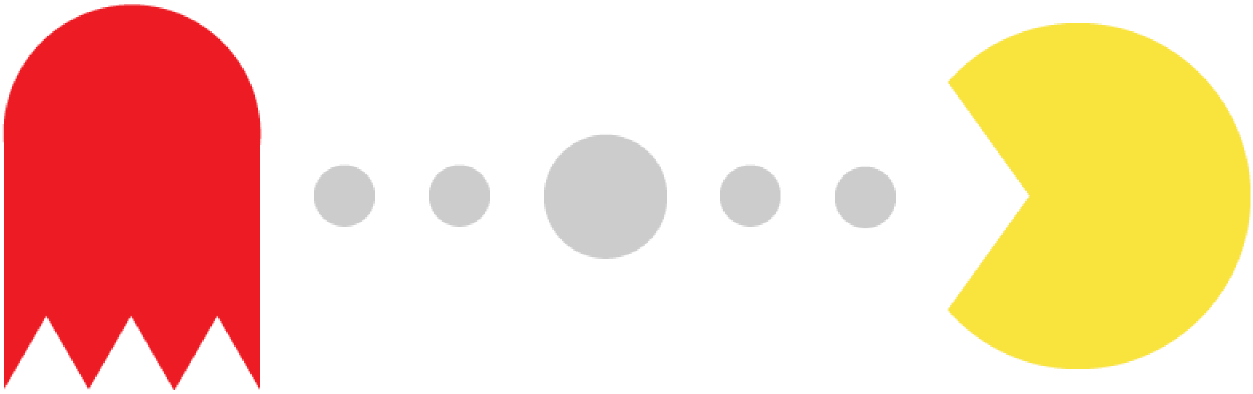
UNIVERSITY OF LOUISIANA LAFAYETTE

Entity recognition and extraction from RDF Files

Information Storage and Retrieval Project

Project proposal document for classification of noun set into named entities, concepts and Extraction of related hierarchical data i.e. properties of extracted named entities.

DBpedia



Submitted to:

Dr. Vijay Raghavan

Center for Advanced Computer Studies

University of Louisiana at Lafayette

Advisor:

Shaimaa

elshaimaa.ali@hotmail.com

Prepared by:

Ramesh Sunkara (rxs6616)

ram@louisiana.edu

Rajkiran Kommineni (rxk1565)

rxk1565@louisiana.edu

Abstract

Ontology is a study of existence of an entity and how this entity is related to the other entities in a hierarchy, and subdivided according to similarities and differences. The entities are classified into named entities and concepts. Named Entity recognition is very important for grouping of related data that enables better understanding of their semantics, and to extract the related concepts. It requires an approach to mine recourses to extract named entities for various classes. For example, the Wikipedia articles consist mostly of free text, but also include structured information embedded in the articles, such as “InfoBox" tables, that contains categorized information, images, geo-coordinates and links to external Web pages. This structured information is extracted and put in a uniform dataset which can be queried. DBpedia is a project which extracts structured content from the information created as part of the Wikipedia project. This structured information is organized into variety of categories such as ‘base ball team’ or broader class of entities such as ‘teams’. An approach has to be developed to recognize the named entity types and extract the properties of the extracted named entities. The evaluation process is conducted manually with manually collected test data.

Table of Contents:

Problem Statement:

DBpedia is a project which extracts structured content from the information created as part of the Wikipedia project. This structured information is organized into variety of categories such as ‘base ball team’ or broader class of entities such as ‘teams’. An approach has to be developed to recognize the named entity types and concepts, and an approach to define the type of each named entity, and an approach to extract the properties of the extracted named entities.

Implementation Details:

This would be stand-alone java project which can be used as a library in any other projects where classification of concepts is required.

As part of this project,

- We classify the given set of nouns into named entities and the concepts. Named entities are the subjects and concepts are the topics that relate to the subjects. We will develop a SPARQL queries and execute them using Jena API for each type of named entity. Jena API is the java interface by which the SPARQL query interacts with the RDF files. We use it to check if the noun is a "NAMED ENTITY" or a "CONCEPT".
- We need to extract all the properties of the named entities.
- One of our goal is to keep codebase very modular to achieve loose coupling and high cohesiveness and also to make the code more readable.
- Good no.of test cases will be developed to ensure the accuracy of the classification we do with the application.

This project is part of the ongoing research on "*Building Light weight ontologies for the purpose of annotations*".

Individual Contribution:

Ramesh S (RXS6616): Setting up the infrastructure for the project like Version control system and project artifact. Will be involved in mainly using Jena API, evaluation and partly on report.

Raj Kiran K (RXK1565): Will be involved mostly in writing SPARQL queries, evaluation and report.

Tools and Dataset:

Following are the different tools identified to use for the development of this project:

- Eclipse Juno – Java IDE.
- Apache Maven – Project Management Tool.
- GitHub – Version Control and Cloud Host service for source code.
- Apache Log4j – Logging different events.
- Apache Jena – API to interact with RDF files.
- SPARQL – Query language to extract data from RDF files.
- RDF – Resource Description Framework.

Data Set:

- DBPedia dumps will be used. [http://dbpedia.org/Downloads39]

Roadmap:

1. **November 1-8** – Download the required DBpedia files, Download JENA, and connect from Jena to DBpedia files with a test query.
2. **November 9-16-** Formulate a SPARQL query for each type of named entity that will be executed through Jena to check if the noun is a concept or a named entity.
3. **November 17-24** – Run the queries over the RDF files and write SPARQL query to extract properties of the extracted named entities.
4. **November 25 – December 1**– Write a report including computation analysis, and obstacles.

Evaluation:

The evaluation process is conducted manually. Following are the checks we want to perform after we run the program:

- If the given noun set is classified correctly or not.
- If the program picks all the named entities correctly or not.
- If the program makes the classification of named entity types correctly or not.
- If the program picks required data or not.
- If the program picks any junk data or unwanted data or not.
- If the program retrieves accurate properties of the named entities or not.

NOTE: We will be adding more test cases, as we progress.

References:

- http://dbpedia.org/About
- http://jena.apache.org/
- https://github.com/
- http://maven.apache.org/
- http://opentox.org/data/documents/development/RDF%20files/JavaOnly/query-reasoning-with-jena-and-sparql
- http://jena.apache.org/tutorials/sparql.html