IRE Major Project Scope Document

Team 19:

1. Pratyanshu Pandey (2019101025)
2. Pavan Baswani (2021701035)
3. Prince Singh Tomar (2019101021)

INDEX

Problem Statement

Collecting domain specific data given a seed set of structured and unstructured sources

Given a domain, the task is to build a structured, clean and high quality dataset for the domain. This will involve searching for structured and unstructured sources for the domain, scraping the web pages to get the raw data, extracting tabular information from the raw data, processing and formatting this tabular information to give a structured, clean and high quality json output.

Planning & Implementation

The project is divided into subtasks:

Getting the Domain and obtaining the sources

1. In this phase of the project we aim to look at several possible domains on which we can work on.
2. We will choose a domain with a good number of wikipedia pages and external structured and unstructured sources.
3. We will then get the domain and the sources finalised and approved by the mentors.

Getting the raw data

1. In this phase we scrape the web pages to extract the raw textual data from the sources. This scraping will be done using libraries like BeautifulSoup and Mediawiki.
2. If the source already has some tabular data present in it, that data is stored separately.
3. In the case of wikipedia this will involve storing infobox and other tables separately and storing the text content separately for further analysis.

Extracting Tabular Data from raw data

1. Upon completion of raw data collection, the significant step is to extract the meaningful entities and find the entity-relation for the tabular columns. In this phase, two major processes involved are Entity extraction and Entity-relation.
2. Using the infobox in Wikipedia, the tabular data will be collected. To extract the tabular data from the wiki text, we will use the dbpedia library. Also, the tables residing in the text will be crawled using the mediaWiKi/beautiful soup.
3. From the other sources (other than wikipedia), we'll use the NER to identify the entities, then the entities will be mapped to certain columns using rule based techniques.

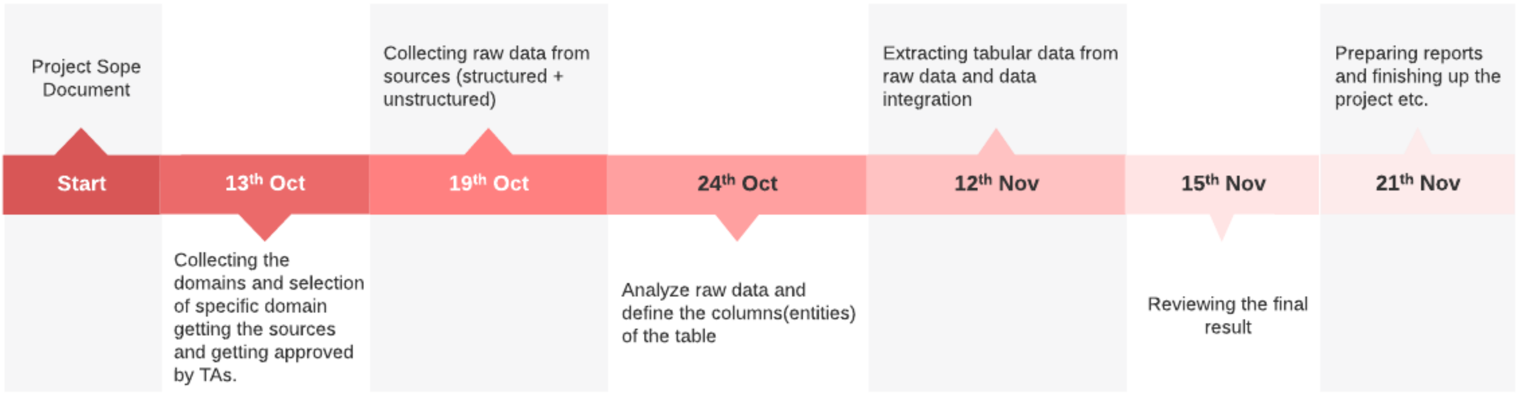
Data Integration

1. Once all the data is in tabular format, there are multiple tables with lots of data from multiple sources from the same entity. All these tables need to be merged.
2. There might be columns with different names but the same meaning. Eg: Date of Birth, Birth date. These issues need to be resolved.
3. Columns with unnecessary, irrelevant or redundant information will be pruned to give a dense high quality dataset at the end.
4. With data from multiple sources there might be conflicts. To resolve that we set a priority order of the sources based on validity of information.
5. Data needs to be formatted uniformly. An example is that date can be present in various formats MM/DD/YYYY or DD/MM/YYYY. A uniformity will be brought in within the same column.
6. Finally the tabular data will be formatted as json.

Review and Completion

1. Once data is collected we evaluate and review the dataset and make modifications if required.
2. We then write a report and complete the documentation of the project.

Timeline and Milestones



Start
Project Sope Document

13th Oct
Collecting the domains and selection of specific domain getting the sources and getting approved by TAs.

19th Oct
Collecting raw data from sources (structured + unstructured)

24th Oct
Analyze raw data and define the columns(entities) of the table

12th Nov
Extracting tabular data from raw data and data integration

15th Nov
Reviewing the final result

21th Nov
Preparing reports and finishing up the project etc.

Deliverables

Mid -

- Collected Raw Data
- Report pdf
  - Analysis/Findings from raw data
  - Methodologies tried
  - Future improvements.
  - Objectives achieved
  - Column definitions
- Crawler build
- Code link to the baseline implementations

Final -

- Project Report pdf
  - Analysis of the tabular data created
  - Methodology tried
  - Modification made to previous deliverable
  - Architecture (Contains diagram & explanation of our approach)
  - Related works
- Domain specific tabular data (in JSON Format)
- Code link to the baseline implementations

Relevant Readings and References

- Data collection & harvesting :

1. Building domain specific dataset or corpora from Wikipedia for ML
2. Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation
3. Domain-Specific Corpus Expansion with Focused Webcrawling
4. (PDF) Harvesting Domain Specific Ontologies from Text
5. https://arxiv.org/pdf/1906.11405.pdf
6. Extracting tabular data from PDFs made easy with Camelot.
7. Extraction of Biographical Information from Wikipedia Texts
8. Using Wikipedia for Hierarchical Finer Categorization of Named Entities
9. (PDF) A Journey of Tabular Information from Unstructured to Structured Data World Using a Rule Engine

- Data integration:

1. Data Preprocessing in Data Mining -A Hands On Guide

- Dbpedia:

1. (PDF) DBpedia and the live extraction of structured data from Wikipedia
2. Building domain specific dataset or corpora from Wikipedia for ML
3. Web Scraping: Introduction, Best Practices & Caveats | by Velotio Technologies | Velotio Perspectives
4. An introduction to web scraping with Python | by Jonathan Oheix
5. Domain-Specific Entity Extraction from Noisy, Unstructured Data Using Ontology-Guided Search
6. Beautiful Soup: Build a Web Scraper With Python – Real Python
7. How to build a simple web crawler | by Low Wei Hong