Project Plan- Knowledge Graphs(KG01)

Entity Search

Authors:

Barun Kumar, Aman Sharma, Saud Afaq,Taslima Akter, Paaras Baru

SUPERVISOR: DIEGO MOUSSALLEM

DATE: 25-NOV-2019

Contents

0.1 OVERVIEW

The project plan comprises of several phrases, where we have been asked to create an NLP pipeline to parse the entity description generation and implement ranking functions for displaying the results as well as UI implementation which needs to be completed within a span of 1 year. We have divided the total available time in this project group in two phases, the first phase would be the time until end of the semester and the second phase would be the time till we start the integration phase. This plan describes in detail the tasks of only first phase and an abstract view of second phase.

We have categorized the tasks of phase one and described the timeline for each task.

0.2 PRIMARY TASKS

0.2.0.1 Document Retrieval

0.2.0.1.1 Fetching the documents using LD2NL framework.

0.2.0.2 Implementing indexing on documents

0.2.0.2.1 Calculation of all the parameters required in the ranking model.

0.2.0.2.2 Creating Index Table.

0.2.0.3 Scoring the documents/ranking

0.2.0.3.1 Implementation of BM25F Ranking function for each Query-Document pair.

0.2.0.4 Making it work for more than one KG

0.2.0.5 Implementation of UI

- Improvement of the existing UI to incorporate aforementioned features

0.3 FIRST PHASE

In the first phase, we take first two primary tasks of indexing and scoring the documents as our goal. We start this by getting hands-on on GENESIS (node JS), LD2NL Framework and SPARQL.

0.3.1 DOCUMENT FETCHING AND CREATE HIGH LEVEL DESIGN DOCUMENT

(2 WEEKS)

The very first task is to retrieve the text/documents (on which the further task of ranking will be implemented) that contains entity description, from the Avatar and Triple2NL packages within the LD2NL framework. Our initial approach is to build an interface which consumes Entity description text via APIs. Furthermore, each document is sent to next stage in the pipeline for Indexing.

3.2 INDEXING (4 WEEKS)

Information Retrieval engines rely on indices for efficient access to the information required for computing scores at query time. We are planning to use indexing table for example: ***R-vertical indexing*** [1] as the index structure for improved performance of the search. In this stage, each document is indexed based on the term(s) appearing in the content.

The output of this stage is the index table for each document. The table contains all the relevant parameters required for the Scoring/ ranking the document.

MG4J [4], an open-source engine for text indexing, is our choice for implementing Indexing.

3.3 SCORING USING THE RANKING MODEL (4 WEEKS)

For ranking model, BM25F ranking model [1] is our choice. Using BM25F, document D is scored against a query Q using a summation over individual scores of query terms $q \in Q$:

$$score^{BM25F}(Q, D) = \sum_{q \in Q} w_i^{BM25F}$$

0.4 SECOND PHASE

In the second phase, we have planned to execute the remaining of the assigned tasks as follows:

1) Making it work for more than one KG
2) Implementation of UI

Here the tasks mentioned above are implemented in the second phase and has described in an abstract manner. The order of tasks and concrete details will be provided at the time of implementation of second phase.

0.5 ORGANIZATION

- Responsibilities and Scrum Master: Aman Sharma.
- External Team Meeting: Every Monday, 11:00.
- Internal Team Meeting: Every Monday, 18:30 to 20:00
- Team members will fill the weekly work report of their tasks.
- Total weekly working hours: 20 hours.

0.6 Technologies and Tools used

We use the following environment in the project.

- Tools: Java 8
- Eclipse
- Node JS

- LD2NL
- GENESIS

Communication: Slack and Trello.

Versioning: GitHub

MG4J

0.7 DEADLINES AND MILESTONES FOR 1ST PHASE AND 2ND PHASE

Serial. No.	Milestone	Deadline	Phase	Duration for tasks	Number of Person working
1.	Creation of the Project Plan	26-11-2019	FIRST	2 weeks	5
2.	Document Fetching	06-12-2019	FIRST	2 weeks	5
3.	Indexing	10-01-2020	FIRST	3 weeks	5
4.	Scoring using the ranking model	30-01-2020	FIRST	3 weeks	5
5.	First Prototype submission	TBD	FIRST		5
6.	First phase Presentation	TBD	FIRST		5

References:

[1]https://drive.google.com/file/d/1C-aNUNcJtGlvg5mXDz2AterM7INnbIw-/view

[2] https://drive.google.com/open?id=1Tq4F5oJqXtH75pFqKvwgby3z5ieef5AZ

[3] https://dl.acm.org/citation.cfm?id=3106514

[4] https://dl.acm.org/citation.cfm?id=1863879.1863882