CS5560 Knowledge Discovery and Management

Project Team: 8

Increment-2 Report

Team Members:

Megha Nagabhushan

Rohith Kumar Nagulapati

Motivation:

Regular **exercise** is necessary for physical fitness and good health**.** Physical activity or **Exercise** helps people lose weight and lower the risk of some diseases. Exercising to maintain a healthy weight decreases a person's risk of developing diseases.

Objective:

The objective of this project is to develop a question answering system based on the **Fitness** dataset to guide people through the right kind of exercise based on their body type and need.

Dataset:

We are using our own **fitness** dataset containing information about Aerobics, Yoga, Walking and Gym workouts. Our model will be able to answer user’s query regarding the benefits of each type of exercise, how long they can be done each day and also about when a particular type of exercise is not advised.

Workflow and Implementation:

The complete system architecture is as shown in figure-2. We are going to run our answer sets through pre-processing.

I. Pre-processing steps (figure-1):

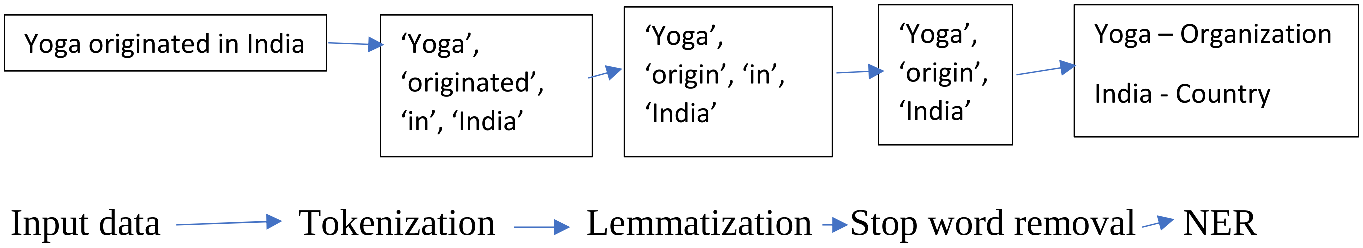

Figure-1. Preprocessing steps example.

1. **Tokenization:** This is the task of chopping up the document into pieces called tokens.
2. **Lemmatization:** This process will return the root word for every token.
3. **Stop word removal:** This process removes the frequently appearing redundant words.
4. **Name entity recognition:** This process will assign the pre-defined categories to the words.

II. Tf-idf, Word2Vec and LDA.

We will perform tf-idf on the dataset obtained after stop-word removal to determine the most important words (using term frequency and inverse document frequency) and then find their synonyms or similar words in the dataset using Word2Vec. The next step would be to apply Latent Dirichlet Allocation to classify the words into clusters.

III. OpenIE

We will use OpenIE parallelly to generate triplets from the original dataset.

IV. Mapping the open IE triplets to the cluster obtained after LDA.

V. Question Answering System

We will use the mapped data to intelligently retrieve answers.

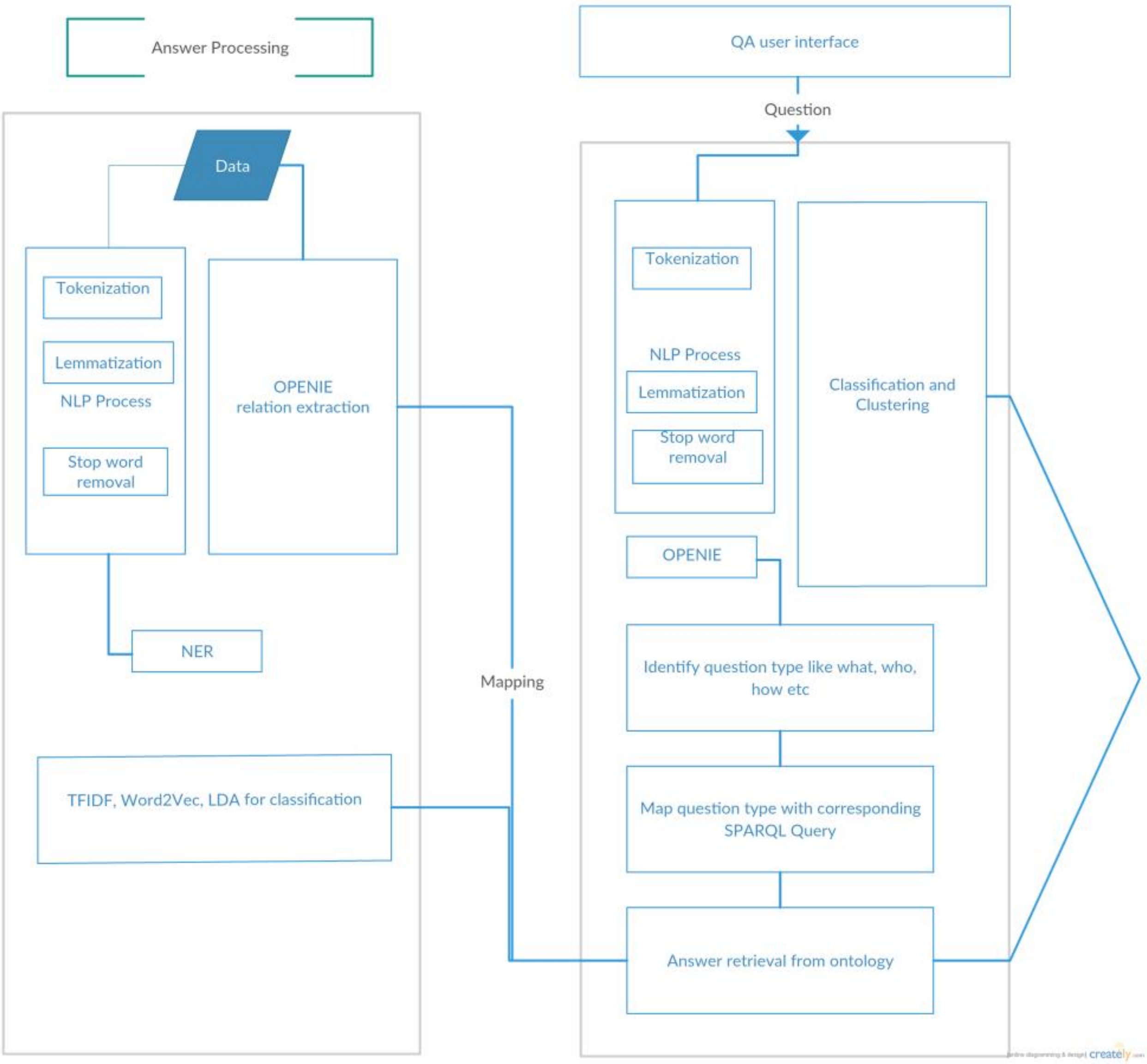

Figure -2 System Workflow

Project Management:

1. **Contribution of each member -**
   Megha Nagabhushan – 50%
   RohithKumar Nagulapati – 50%

2. **GitHub Link for the project -**
   https://github.com/ROHITHKUMARN/CS5560_KDM_Project

3. **Future work –**

   We will be implementing the Question processing and also generate ontology for our dataset. We will also be visualizing and querying our ontology. We will also be comparing our results from k-means and LDA to determine the best clustering method for our dataset.