**CS5560 Knowledge Discovery and Management**

**Project Team: 8**

**Increment-2 Report**

**Team Members:**

**Megha Nagabhushan**

**Rohith Kumar Nagulapati**

**Motivation:**

In the present data centric world, about 80% of the total data is semi structured and is difficult to derive information from these sources in its raw form. This fact motivated us in developing a dynamic question and answering system which is the most welcoming model in all the fields as this can be one’s personal assistant in responding to user queries. Also, this model finds worthy in every use case where we require some help from computer.

**Objective:**

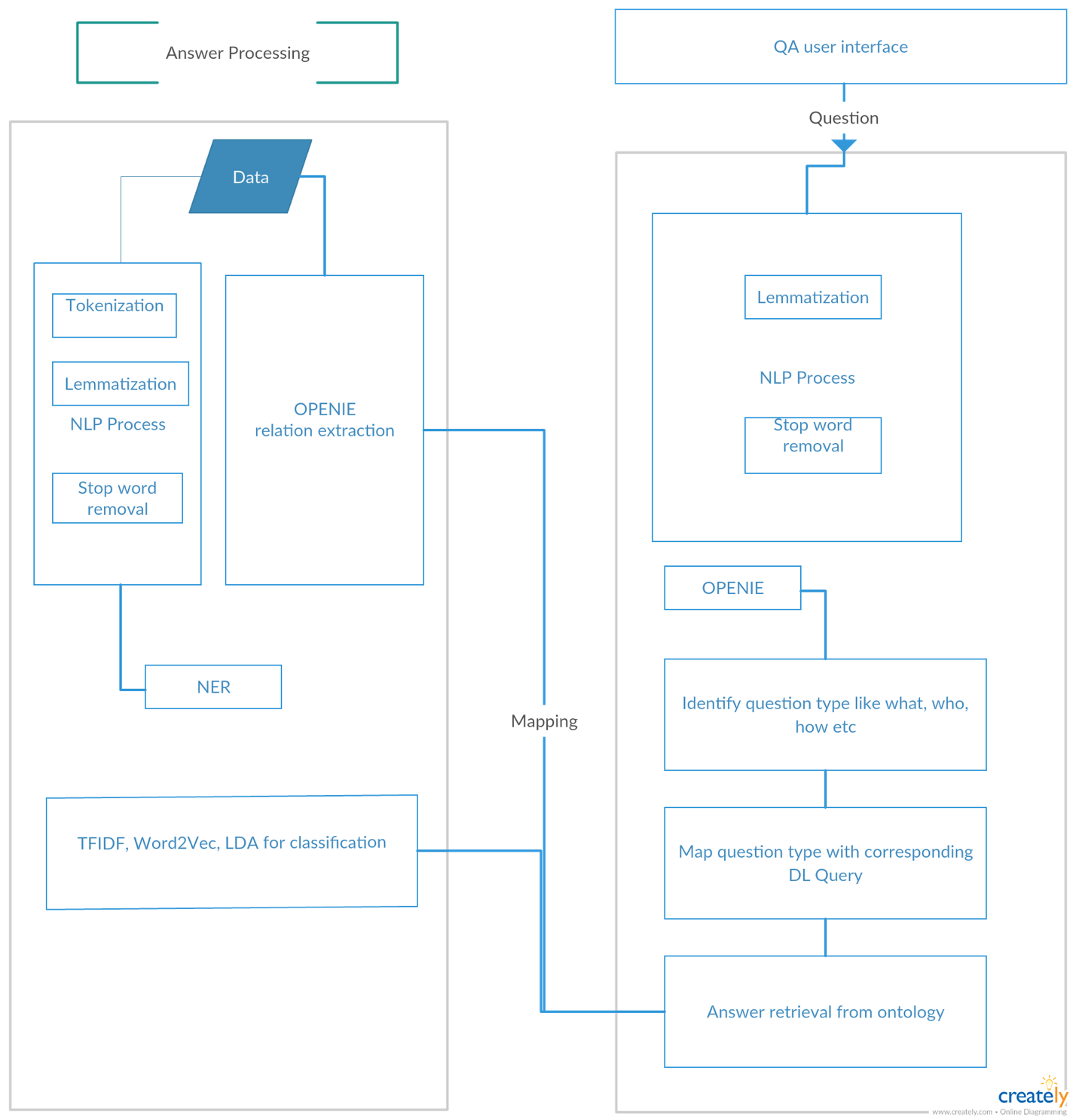
The main objective of this project is to extract information from a huge collection of medical data sources using latest techniques of natural language processing like lemmatization, named entity recognition and using OpenIE, Word2Vec, WordNet for extracting relationships, using clustering and classification techniques for topic discovery. Finally, to represent the obtained information in the form of ontology which then becomes available to question and answering system to query the results from the system.

**Dataset:**

## For this project, we have taken the NCI-PID-PubMed Genomics Knowledge Base Completion Dataset. From this Data set we try to construct knowledge graph and making system dynamic to answer all possible questions on medical domain related questions.

**Workflow and Implementation:**

The complete system architecture is as shown in the below figure.

****

1. **Natural Language Processing(NLP)**

**NLP** in the field of computer science is helpful in giving computer the ability to understand human speech as it is spoken. The most common tasks include tokenization, Lemmatization, Parts-of-Speech Tagging, Named Entity Recognition and breaking documents to sentences. It is helpful in pre-processing raw text data.

* 1. **Tokenization**

Tokenization is the process of breaking sentences in to tokens which are the smallest constructs of a huge text data.

Sentence: South Africa far too strong again

Tokens: South, Africa, far, too, strong, again

* 1. **Lemmatization**

Lemmatization is the process of separating words into individual [morphemes](https://en.wikipedia.org/wiki/Morpheme) and identify the class of the morphemes.

Original string =South Africa won the series

Lemmatized form=South Africa win the series

* 1. **Stop word Removal**

Stop word Removal is the process of removing stop words from the data.

For example, the stop words in English be:

able, about, above, according, accordingly, across, actually, after, afterwards, again

* 1. **Named Entity Recognition**

Given a collection of text data, NER determines which items in the text map to proper names, such as people or places, and what the type of each such name is e.g. person, location, organization etc.

PERSON Location Numbers

Hayden South Africa 281-9  
 Ricky Australia 2-0  
 Ponting New Zealand 60  
 Damien  
 Martyn  
 Hayden  
 Adam  
 Gilchrist  
 Daniel  
 Vettori  
 Stephen  
 Fleming

1. OpenIE relation Extraction
2. WordNet
3. TFIDF, Word2Vec
4. Feature Vector
5. K-means
6. DL query
7. Spark Play framework

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

Yoga – Organization

India - Country

‘Yoga’, ‘origin’, ‘India’

‘Yoga’, ‘origin’, ‘in’, ‘India’

‘Yoga’, ‘originated’, ‘in’, ‘India’

Yoga originated in India

Input data Tokenization Lemmatization Stop word removal NER

**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.
5. **TF-IDF, Word2Vec and LDA.**

# We will perform tied 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.

# OpenIE

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

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

# Question Answering System

# We will use the mapped data to intelligently retrieve answers.

# Figure -2 System Workflow

# Project Management:

# Contribution of each member -

# Megha Nagabhushan – 50%

# Rohithkumar Nagulapati – 50%

# GitHub Link for the project - <https://github.com/ROHITHKUMARN/CS5560_KDM_Project>

# 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.