**Dynamic Intelligent Q/A Systems**

**CS5560 Knowledge Discovery and Management**

**Project 2  
Summer 2017**

**Team 7:** TechGeeks



**Team members:**

* *Sai Jyothi Gudibandi*
* *Kalyan Kilaru*
* *Chaitanya Kumar Peravalli*
  1. **Motivation:**

Nowadays, with the rapid growth in the use of internet, the information is growing more and more rapidly. To answer the user’s questions, internet has become an important intermediary or a resource. The traditional search engines like google, yahoo search engine help the users to get the information they are looking for to some extent, but it sometimes returns irrelevant search results along with the relevant results. So, making a more Dynamic Intelligent Question answering system may help in information retrieval effectively.

Question-Answering system which supports natural language processing provides the users with a human-machine interface. This system be in similar type the way how the people ask questions to the search engine and it has more advantage compared to traditional search engines by understanding the purpose of the question and respond accordingly. Question-Answering system works more efficiently because it gives more appropriate information or answers to a question.

* 1. **Objectives:**

The main goal of this Question and Answering System is to answer the queries posed by the humans in their normal language, using a pre-structured database or a assembly of Natural language documents. Here Question Answer system deals with the wide range of question like What, Why, When, How much, How many and Is/Are-means Yes or No type. In this system we will process the human language queries by using NLP(Natural Language Processing) and other techniques, based on the results it will automatically generate the answer for those questions. Question Answering is an application for the knowledge base representations. Knowledge graph come under the knowledge representation.

Now a day we do have lots of data available on net, this will increase the difficulty level of the information handling. BY using some artificial intelligence algorithms, Machine learning will help with the data handling issues and it will make the search easy. Using different techniques results the accurate answers to the queries and it will increase the knowledge managing process. Here we are representing the text data in the form of knowledge graph, which will give the better understanding of the data to the users. For this knowledge graph representation we used the WebVowl plugins. And for fetching the information we used Protégé tool, in that Spark SQL is used to get the schema of the data.

* 1. **Significance:**

Using Question Answer system searching become easy and it will give accurate and relevant answers for the questions, we can improve the document and knowledge organization. These kind of system will be very useful in case of helping desks, knowledge management systems and E-libraries.

* 1. **Chosen Domain:**

For this project, we chosen Question and Answering which will be very helpful for searching and finding the correct answers to the user queries and to increase knowledge management. Here, we have chosen two datasets the Sports domain.

* 1. **Q/A Application:**

Q/A Application is a knowledge based application which will involve in so many steps to process the questions and to give the exact answer for the queries. In Q/A Application knowledge extraction is the first step to find the type of the question, POS tagging is used to determine the answer type. And TF-IDF is used for information retrieval. Name Entity Recognition is used to find corresponding “Place”, ”Date”, and “Person”. A WordNet which is a lexical dictionary is used to understand the context of the data. In this application, user can ask When, Where, What, Why type questions, which are related to the chosen datasets. By using different techniques, this application finally give the answers for the questions posed by the humans in their language. This will improve the search results in that documents.

1. **Related Work:**

In present real world, huge data leads to many data organization issues. There is a completeness and correctness problems with the existing algorithms. The solution algorithms for this problem should provide easy maintenance of large data and should give the accurate results. But there is no single approach or algorithm which satisfies those, For that we have to integrate different algorithms to design such an accurate search engines.

Searching for the required data from the large datasets is so hard. Knowledge graph gives somewhat better solution for this problem by representing the data into relations and entities. Now, we have so many knowledge graphs in the market, but Google knowledge graph became popular for searching data. By mixing different methods, knowledge graph gives the solution for the correctness and completeness problems.

To give the accurate search results for the user queries, we need to consider all the datasets which are existing in the internet. By using Knowledge vault, we can use the all data present in the internet. Knowledge vault consider the information in the form of RDD triplets, where triplets has subject, verb and predicate. After the collection of data, organizing the data is the next problem, where we have to find the relation between the entities. By using the Deep Dive concept we can get the solution for the above problem. Deep Dive helps for extracting and integrating the data in order to make the data training procedure easy.

After getting the entity relation, by using FchSen we can represent the data as RDF Triplets, where the semantic relationship between the words is organized and it combined with the related data in order to get the further simplified information.

Founder is an framework for constructing the knowledge graph for richly formatted data. By using this framework we can handle the structured data and we can get the solutions for the data completeness and data correctness issues.

Getting accurate answer for the question is very important, this will achieve by using LDA which will extract the presented topics in datasets and cluster the different topics in the question. For getting the RDF triplets for the question, OpenIE also helpful.

Concept Net is used for data visualization which will help the users to understand the data process. This visualization of the data is possible by getting the entity relationship between the objects in the data.

DBpedia helps us to get the data which present in different languages which will improve the accuracy of the search results. After getting the data from the different sources next step is getting the accurate results for the query in less time. This can be achieved by using the spark and DL query. By using all of these different algorithms for the question process, then we will get the accurate results in less time.

1. **Data Sets:**

For this project we have chosen two datasets from the given data sets lists. Those two data sets are:

* BBCSport
* http://mlg.ucd.ie/datasets/bbc.html
* WikiRef220
* http://mklab.iti.gr/files/WikiRef\_dataset.zip

In BBCSports, we have so many interesting topics on sports like athletics, Cricket, Football, Rugby and Tennis. Here, we mainly focused on the Cricket topic because that topic has many interesting things to know.

WikiRef220 data set have data about the Barack Obama, Financial Crisis, Elections, Airlines, Parris attacks. By this we can get to know some interesting this which are going around the world and somethings about the famous personalities

* 1. **Chosen Datasets from Stanford Question Answering Dataset:**

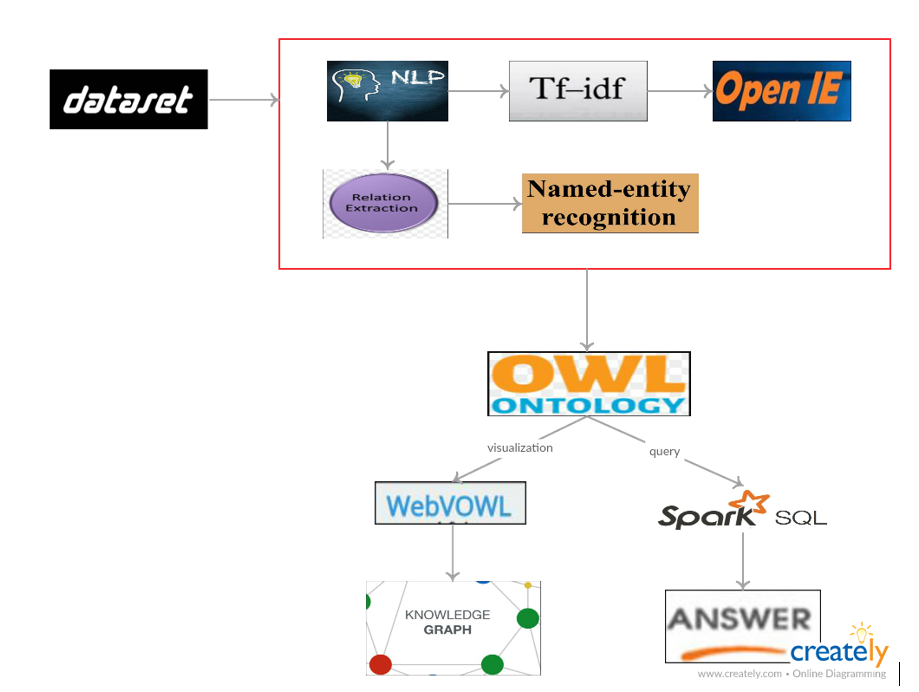
Stanford Question Answering Dataset is a collection of questions on the Wikipedia articles. Here, the answers for the single question is a collection of text or some span of text which present in the corresponding passage. This dataset is the largest data set than the other presented datasets by having one lakh plus Q/A pairs on five hundred plus articles.

* https://rajpurkar.github.io/SQuAD-explorer/
  1. **Chosen Categories from Yahoo Answers Dataset:**

Yahoo Answer dataset is a collect of more than one lakh questions and answers sets from the Yahoo Answer sites.

* https://cogcomp.cs.illinois.edu/page/resource\_view/89

1. **Design:**
   1. **Workflow:**



**1. NLP:** Nlp is the process of extracting the system understandable language from the normal user understandable language.

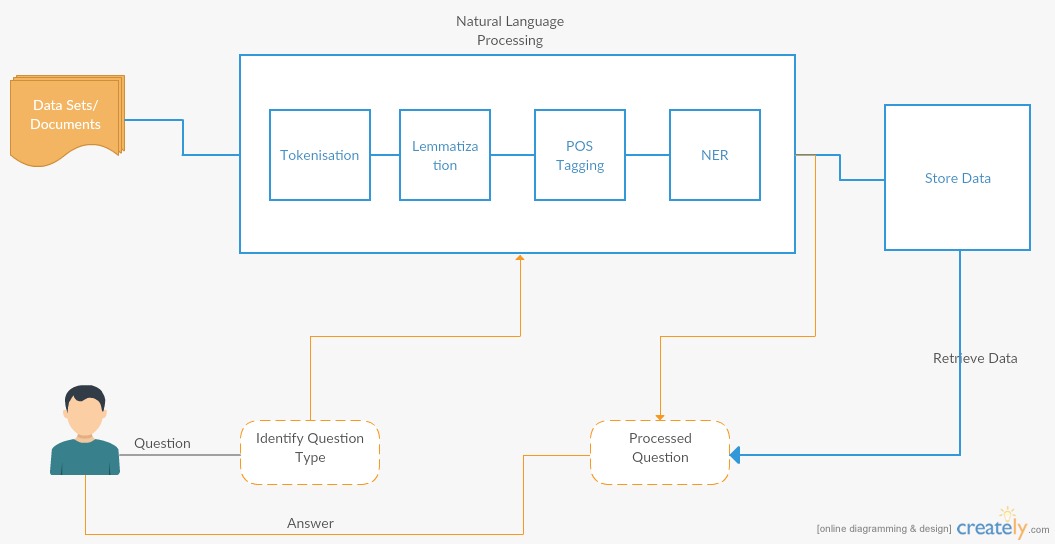
**2. Information Retrieval:** Process of Retrieving the data from the input datasets.

**3.** **Topic Discovery:** Retrieving the topics which are present in the question.

**4. Processing the Question**

**5. Getting the answer for that query**

* 1. **Processing Using NLP:**



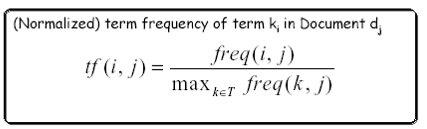
Nlp is the process of extracting the system understandable language from the normal user understandable language. Here, Nlp takes the human language documents as the input and process that data in different steps like ***Tokenization*** – is the process of cutting the data into tokens, ***Lemmatization***- reducing different forms of words to communal base form of that word, ***POS*** ***Tagging***- it is a process which allots parts of speech to individual token. ***NER***- Name entity relation, which labels order of words in the taken text.

* 1. **Feature Generation using Information Retrieval(TFIDF,Word2Vec):**

**TFIDF:**

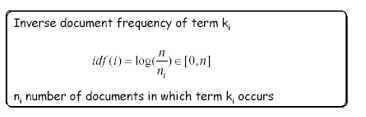
TFIDF Is nothing but the term frequency inverse document frequency, IFIDF is a mathematical measurement that is planned to reproduce how significant a word to that file in a group or corpus. The TF amount is compared to IDF count. Outcomes in matrix, by that the corpus reduced to fixed-length output.

Here is the formula for finding the Term frequency and weight of the term in that document:





Here is the formula for finding the Inverse Document frequency and weight of the term in the document:





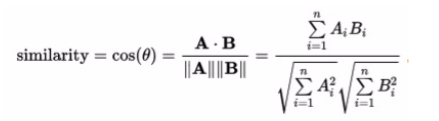
TFIF weight of term:

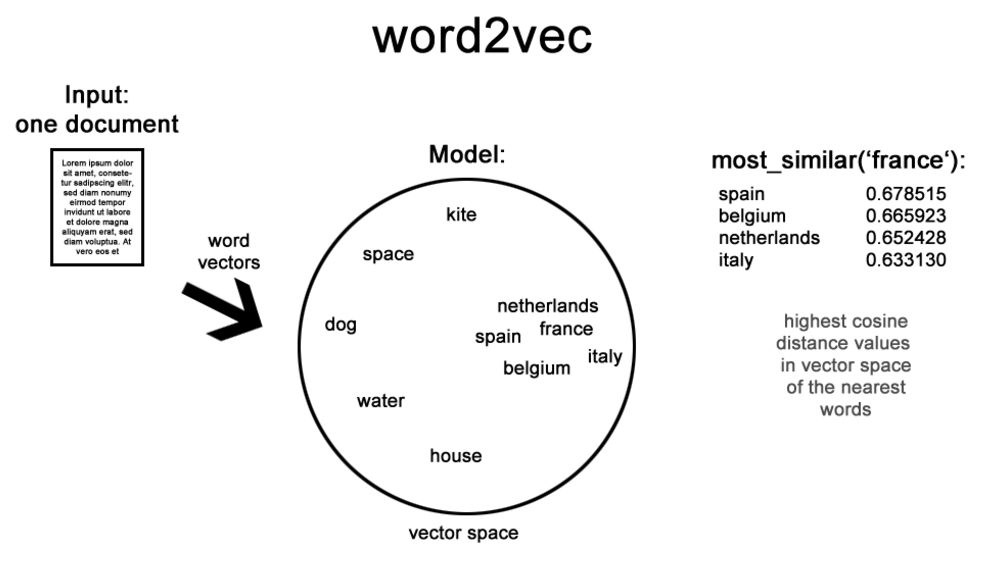


**Word2Vec:**

Word2Vec calculates scattered vector illustration of words. Main benefit of this distributed illustrations is, like words are near in vector space, which makes generalization to new patterns easier and model estimate further strong.

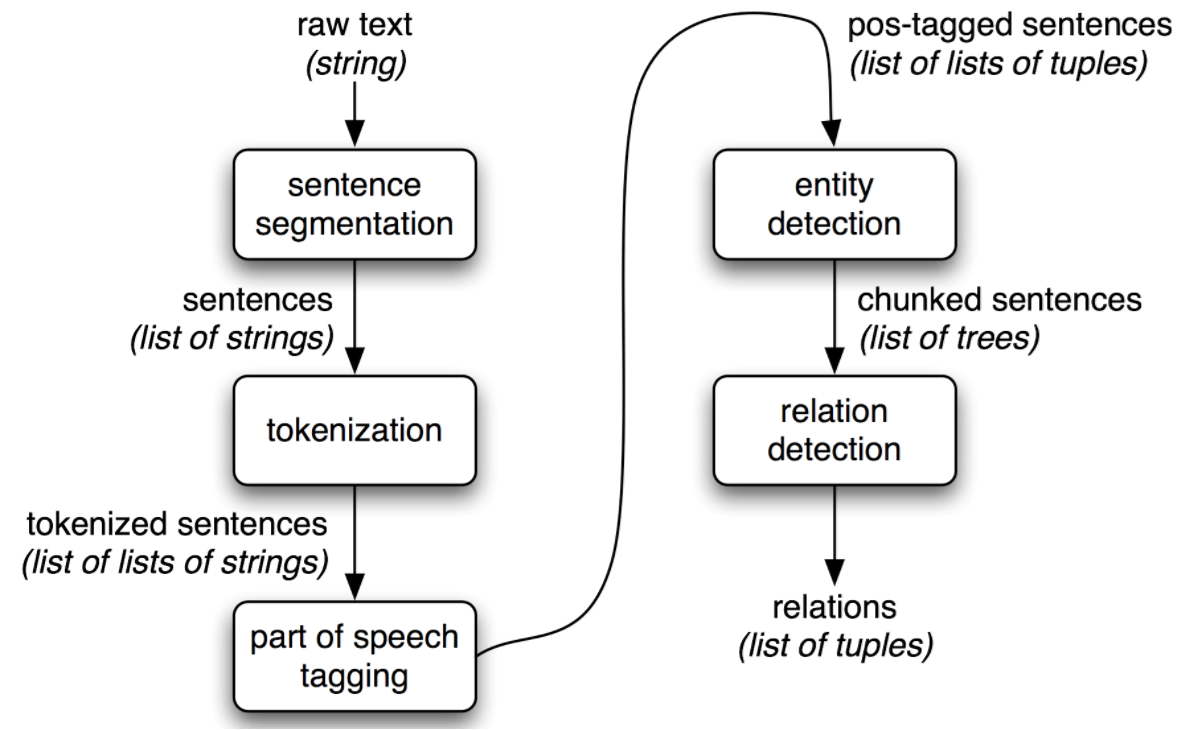
Based on the given input data, Word2Vec create a model. Word2Vector is comfortable with more input data. The input data is in sentence arrangement. Her, this program takes input is a text file and it will give output as the list of synonyms for a specified word. For example if we specified the word as “zero” depend on the document the we will get list of words like two, four, nine, eight including with corresponding cosine similarity.





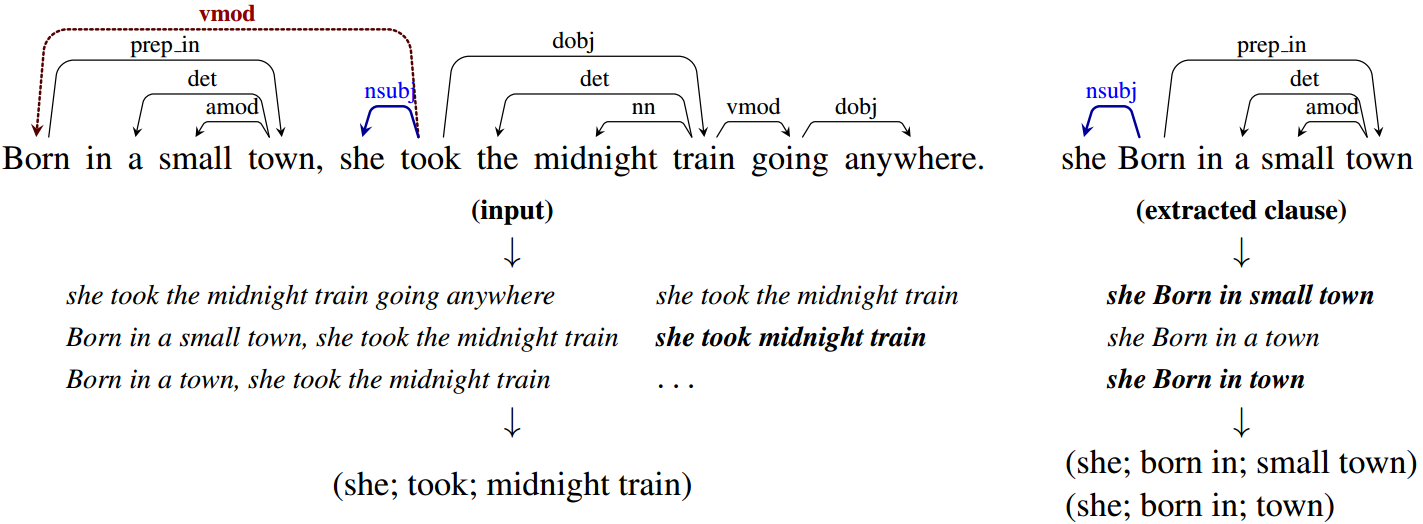
* 1. **Triple Generation: Information Extraction(OpenIE, WordNet):**

Information Extraction is the process of pull out the structured info form the formless data or semi organized data.



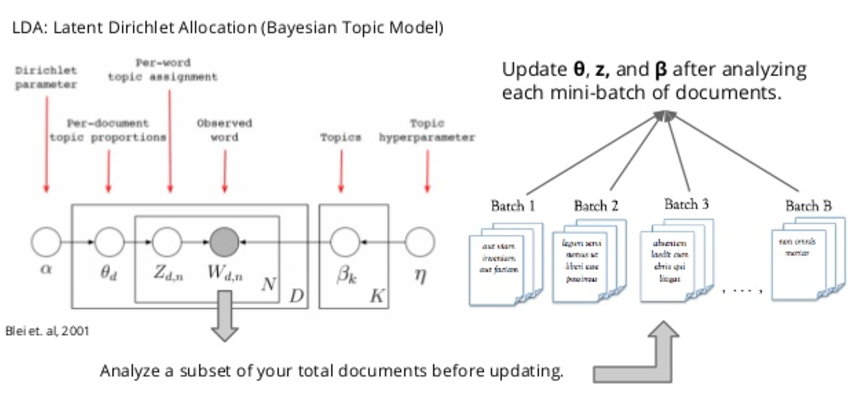
**OpenIE:**

Open IE is the process of extracting the relation between the tuples which are present in the text. This system takes the text as the input and initially splits the presented sentences into different clauses of fragments, and then these fragments are combined into triplets then gives as output.



* 1. **Answer & Question Categorization: LDA & Machine Learning(Clustering, Classification):**
     1. **LDA:**

Latent Dirichlet Allocation, is the process of extracting the topics. It will take the text as the input data and tokenize and lemmatize the text data, remove the stop words in the text then get the topics in the text by running the Spark LDA.



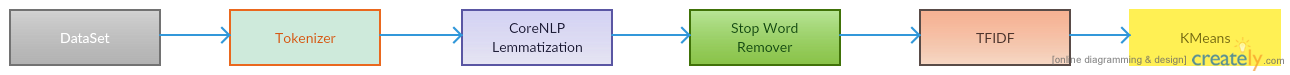
**5.5.2Machine Learning:**

**KMeans Clustering:**

It is a clustering techniques, which takes the text as input and tokenize the data and apply the lemmatization on the text and then remove the stop words in the text, give the TFIDF of the text and finally gives the KMeans.

This K-means is most frequently used algorithm which cluster input information into a given no. of clusters. Here, *k* is the no. of wanted clusters. *maxIterations* is extreme no.of repetition. *initializationMode* requires any casual initialization or by k-means. *runs* is no.of times to execute that k-means process. *initializationSteps* regulates the no.of stages in k-means process. *epsilon* controls detachment edge. *initialModel* is the non-compulsory cluster centers which are used for start the process. If that constraint is provided, only single execution is done. For creating model we used KMeans.train() method. Again we used vector concept in order to check to which cluster that goes.

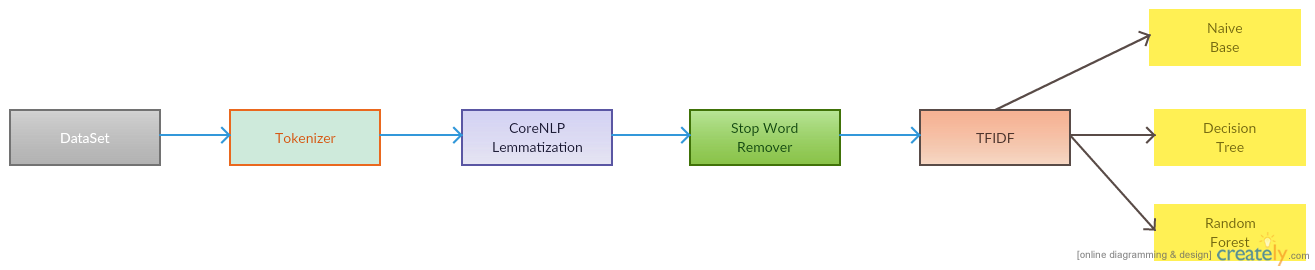
The questions from users will be categories and store based on topic/question along with processed answer.... kind etc....this helps in faster retrieval of answers in future. This can also be used to suggest recommended questions to the user based on the current query.



**Classification(Naive Base, Decision Tree, Random Forest Algorithms) :**

**Classification** is nothing but a problem of classifying the observations, to which category it goes, on the base of the training dataset.

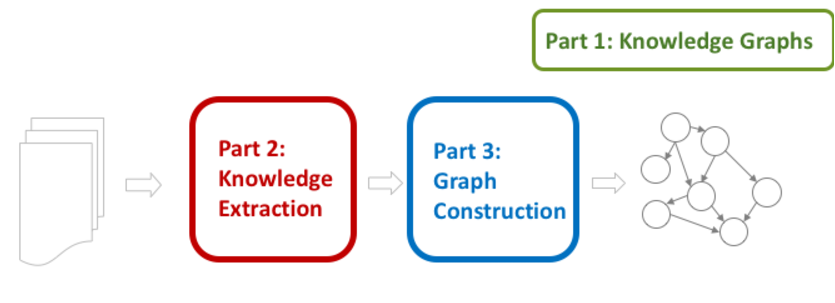
Here we used Decision Tree, Naïve Bays classifier and Random Forest Classification algorithms.



* 1. **Generation Of knowledge Graph(KG) using Answer Data:**

Knowledge graph doesn’t have specific design, the difference between different knowledge graphs is the way they handle data like the type of data.

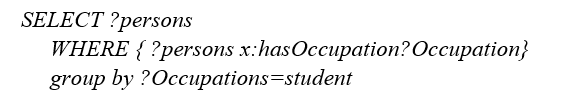
**5.6.1. Knowledge Graph construction from the text:**



For constructing knowledge the is no particular rules to follow. In knowledge graph construction, initially it will recognize the data entities in the text and find the relationship between the entities finally find the data flow between those. Here is the Knowledge Graph for chosen datasets.

* 1. **Generation of Spark or Swrl using Question Data:**
     1. **Spark SQL Query processing:**

This is a query language processing for the spark programs. It is similar to the normal SQL, like normal one it will execute the queries.



* + 1. **Swrl Query processing:**

Example for fetching the data based on some conditions:



* 1. **Question Answering using KG and Spark or Swrl:**

Knowledge Graph is useful for extracting the answer from the questions.

1. **Implementation:**
2. **Results and Evaluation:**
   1. **Data Sets used:**

* Stanford Answer dataset
* Yahoo dataset4
  1. **Results:**
* **NLP,** Convert the human understandable language to machine understandable language.
* **TFIDF** results corresponding weight of the words in the input datasets. **Word2Vec** denote the word in the form of words in vector space.
* **OpenIE,** gives the triplets<S, O, P> as output and **WordNet** gives the synonyms of the particular words.
* **KMeans/LDA,** gives the output as the topics of the input dataclustering algorithms like Naive Bayes gives the confusion matrix of the topics.
* **Using Ontology,** we can generate the knowledge graph for the datasets.
* **By using Spark SQL/DL** we can process the given query to get the answer for that.
  1. **Evaluation and Validation:**
* **Accuracy:** 65%
* **Performance:** Good
* **Run Time:** 0.89sec
  1. **Correct and Incorrect cases:**

For topic Extraction, K means is the better approach than the LDA. Spark SQL gives the accurate results than the DL in query language processing

1. **Project Management:**
   1. **Contribution:**

*SaiJyothi Gudibandi*

*Kalyan Kilaru*

*Chaitanya Kumar Peravalli*

A picture containing screenshot

Description generated with high confidence

* 1. **Statistics:**
  2. **Concerns/Issues:**

**NA**

* 1. **Future Work:**

Focus on implementation of the Question Answer System to build a best system which will give the better answers for the user queries.

References:

* <https://www.google.com/imgres?imgurl=http://www.nltk.org/images/ie-architecture.png&imgrefurl=http://www.nltk.org/book/ch07.html&h=1138&w=1805&tbnid=C2FGsvbH2JhijM:&tbnh=132&tbnw=211&usg=__bFmy6rqEF44K80kQLhpBxvgUOv0=&vet=10ahUKEwi24tmc2aDVAhXGxYMKHXvjAcUQ9QEILDAA..i&docid=jcxqlDCdaJG-VM&sa=X&ved=0ahUKEwi24tmc2aDVAhXGxYMKHXvjAcUQ9QEILDAA#h=1138&imgdii=C2FGsvbH2JhijM:&tbnh=132&tbnw=211&vet=10ahUKEwi24tmc2aDVAhXGxYMKHXvjAcUQ9QEILDAA..i&w=1805>
* https://en.wikipedia.org/wiki/Information\_extraction