Dynamic Intelligent QA System Related Business Domain

Sujitha Puthana
Dept. of Computer Science
University of Missouri Kansas City
Kansas City, USA
Spb4b@mail.umkc.edu

Yalamanchili Sowmya
Dept. of Computer Science
University of Missouri Kansas City
Kansas City, USA
syb7c@mail.umkc.edu

Jakkepalli Rama Charan Pavan
Dept. of Computer Science
University of Missouri Kansas City
Kansas City, USA
rjhg5@mail.umkc.edu

Nandanamudi Sreelakshmi
Dept. of Computer Science
University of Missouri Kansas City
Kansas City, USA
snhnc@mail.umkc.edu

***Abstract*— Data Science is a system to extract knowledge of data in various forms, either structured or unstructured from various domains, like Knowledge Discovery in Databases(KDD). Natural language processing is used for processing the text which is machine understandable and which will help for fast retrieval of data. Ontology plays an important role with respect to entity classification to answer questions. Visualization of the data by classification into classes, subclasses, data properties, object properties using the concept of ontology tool protégé. Protégé tool has its unique features for fetching the related information by using either spark SQL or DL query, which are the simplified query to fetch instances. This application helps in fetching the answer to questions by using NLP Process, word2vec, TF-IDF, N-gram. NLP, kmean, Classification of data, NLP algorithm is useful step for text processing and then we are extracting the relevant data. Visualization of the knowledge graph is also of great use. However, all the algorithm we are using in the project have its own significance. Comparing all these processes to find the best process with respect to time, accuracy, cost to select the best process. Query using the spark sql or DL sql can fetch the information from the entity classified. These queries are very fast to extract the information to answer the relevant questions.**

1. INTRODUCTION

In present days, the amount of data is increasing and this is leading to the difficulties in handling the data. So, we need the machine learning algorithms to handle these huge data. We are making use of artificial intelligence algorithm for machine learning to handle data and search the data.

As we are using these AI algorithms for handling data, this helps in getting through different algorithms available including TFIDF, NLP algorithms, word2vec algorithm, kmean algorithm, classification of data using all these processes and analyzing the accuracy. This tremendous process leads to precise answer of the question.

In the process of going through different AI algorithm to classify data and handle them. We could understand the importance of each algorithm with the specified uniqueness. By making using use of all these algorithms simplify the management of data.

Human are more prone to understand the visualized data than the text data. Visualization includes the presentation of the data in the form of knowledge graph. The text data is classified into classes, subclasses, properties are extracted. We generate an owl class and give it as input to protégé and visualize it using either plugin VOWL or webvowl.

WebVowl: http://visualdataweb.de/webvowl/

Spark sql is like normal sql commands that can be used to fetch information in the form of schema. In our application, we are using the spark sql commands to answer some questions. Protégé tool has its own query language DL query which is more simplified version of querying. DL query fetched the instances of the classes which can answer few questions.

2. RELATED WORK

In the present days, where the data is huge leading to data management issues. There are many algorithms already existing but the main problem in the existing algorithms are completeness and correctness. To solve this problem, we need to consider all these algorithm and judge wisely which all are the algorithms that we can use to easily maintain data and give us the high accuracy. But a single algorithms or approach cannot solve this problem. Hence, we should integrate multiple algorithm for high accuracy in designing the search engine.

Searching the huge amount of data is very difficult. Knowledge Graph represents the graphical representation of the entities and interrelated relationship. There is different knowledge graph available in the market but googles knowledge graph is the popular search engine algorithm. Best knowledge graph can be designed solving the completeness and correctness issue by integrating different approaches of knowledge graph available in the market.

Data sources that are available to us are limited. We can increase the accuracy to provide the best answer to any question is by considering all the data sources that are available on the web. The solution for this approach is the knowledge vault that was made available to us by google that takes the data in RDD triplets i.e., subject, object, and predicate. After collecting the data and finding the entities our next problem would be organizing the data. We Deep Dive approach helps in resolving the problem of extraction of data and its integration to fetch accurate prediction making the training process easy.

After the data is represented in RDF triplets, the semantic relationship can be organized using the FehSen to merge the related information leading to more simplified data. It is known fact that structured data is easy to handle than unstructured data. Fonduer is the approach in focusing the construction of the structured data from the plain text. By using all these approach helps in improving the handling the data and solve the “completeness and correctness” problem.

Optimization of the questions is important to get high accuracy. Latent dirichet allocation is used to extract the topics. Applying the LDA algorithm on the question is used to cluster the question topic, measuring the similarity based on semantic between multiple questions. OpenIE algorithm is also applied on the questions to generate the RDF triplets to understand the question.

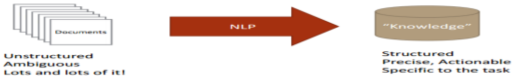
Visualization of the data plays a keys role in understanding and process huge data. Visualization is done by extracting the key entities and relationship between them. Object properties defines the property relationship between two instances. Data properties defines the relationship between two entities. Modern algorithm “Concept Net” which is an improved version to visualize the data using the labels and edges.

In our world where there exists data in multiple languages. To achieve the high accuracy information, we need to consider data from all the available sources in all the languages. DBpedia algorithm is the best approach for this process. After completing the data extraction, data retrieval our main task is to improve the processing time and accuracy to fetch the most relevant answer to the question. One good approach is the query to fetch the relevant answer. Spark query and DL query are the highly used fast processes query languages. Thus, we are processing question and data through all the available algorithm to fetch the answer very fast.

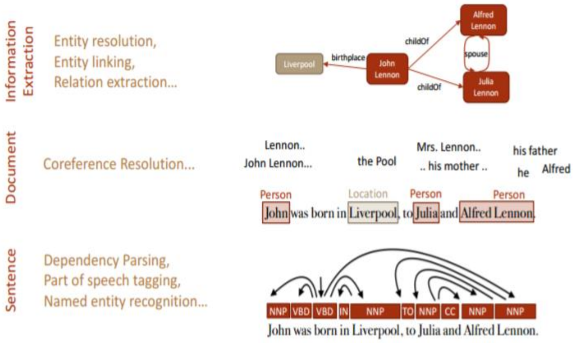
3. PROPOSED WORK

Workflow and Algorithms of Proposed System:

Step 1: Natural language processing – This process includes the identification of token, lemmatization, named entity reference(NER), co-reference resolution.



Step 2: Feature Generation using Information Retrieval – Retrieving the information from the text. We are including the identification of the NER i.e., PERSON, LOCATION, ORGANIZATION.



Step 3: Topic Discovery – Topic discovery helps identification of the topics from the context question.

Step 4: Knowledge Graph construction – Construction of the knowledge graph from generated NER.

Step 5: Preparing query for the question.

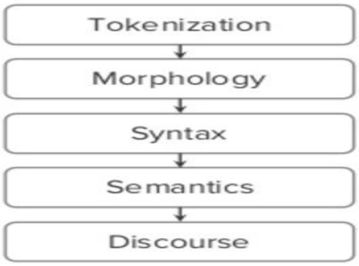
Step 6: Execute query to generate the answer.

3.1.1 Preprocessing using Natural Language Processing:

Natural Language Processing is the process that’s makes the computer to understand, analyze and extract meaning from human understandable language in a useful and smart way. NLP algorithms helps the organizing and to structure data to perform automatic summarization, named entity recognition, translation, relationship extraction, speech recognition, sentiment analysis, topic segmentation.

Algorithm for NLP designing:

- Tokenization – Break the text data into sentence, words.
- Lemmatization – Recognizing the base form of word.
- Morphology – Includes Part of Speech recognition, stemming i.e., excluding the postfix words to get the base root word, Named entity recognition.
- Syntax – Parsing Constituency or dependency
- Semantic – Coreference resolution i.e., finding the context that belongs to same entity.



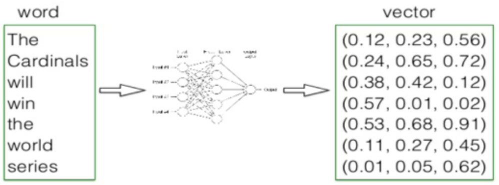
3.1.2 Feature Generation using Information Retrieval: Information retrieval is the process of tracing through the stored data and recovering specific information from huge amount of stored data. It is very difficult to find the specific data from such a huge amount of data. So, we are using the below approaches to simplify the information retrieval process.

Term Frequency Inverse Document Frequency(TFIDF): TFIDF is the numerical weight of the tokenized word that demonstrate the importance of the word in the huge document. The weight of the word increases with the repetition of word in the document. TFIDF is can be represented as TF*IDF i.e., product of term frequency i.e., occurrence of word in a document and Inverse document frequency i.e., log value number of document the word exists divided by the total number of documents.

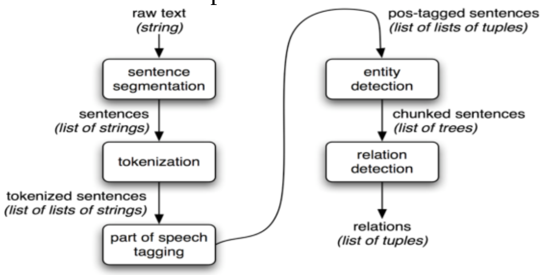
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

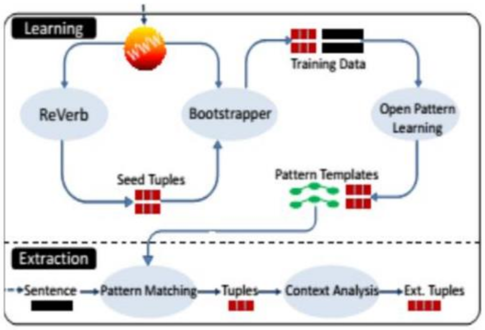
Word2Vector: Word2Vec is the process of construction of the vector from the huge text document. All the word vectors are marked in the vector space where the closely meaning words are very close to each other. Thus, mean that they are the same grouped words. This model leads to the other distributed representation model i.e., Continuous bag of words, Skip gram. Bag of words mean predicting the words from context and the skip gram is predicting the context from words.



3.1.3 Triple Generation <S,P,O> Information Extraction: It involves the process of extracting the information from the unstructured or the semi structured data i.e., normal text document. Information extraction utilizes the NLP process to extract the relationship between the entities.



OpenIE: Open information extraction is the process of extracting the RDF triplets. RDF triplets are subject, object, and predicate.



Algorithm to retrieve OpenIE Triplets:

- Input the data to the system.
- Matching the pattern from already predefined algorithm.
- Extracting the tuples.
- Analyze the context.
- Extracting RDF triplets.

WordNet: It involves the generation of the synonym for a token of word. WordNet algorithm in analyze the data to extract the correct information though we use the synonym of the word. WordNet generate the synsets, which is the group of words with similar meaning.

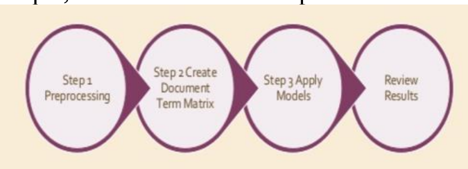
3.1.4 Answer & Question Categorization:

Machine Learning:

Machine learning involves the process of automatic analyzation of data using the advances artificial intelligence algorithm. This process simplifies the prediction from the existing huge data. Machine learning algorithm are very efficient.

Clustering: Cluster represent the group of similar kind. In data analyzation, we use clustering process to group together similar words using vector.

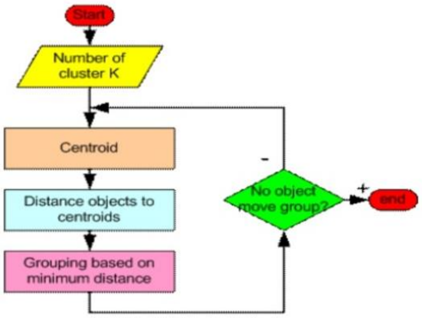
i. Latent Dirichiet Allocation: LDA is a clustering technique, used to extract the topics.



Algorithm for LDA clustering:

- Input the data i.e., either text data or question.
- Tokenize the input data.
- Implement the lemmatization i.e., generating the dictionary word.
- Remove stop words including punctuation.
- Run the spark LDA, to generate topics.

ii. K-Mean: K-mean is a clustering technique,


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

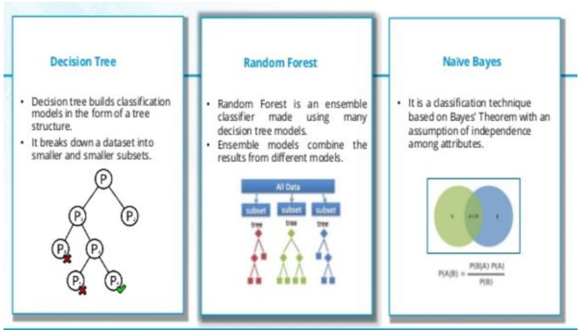
Algorithm for k-mean clustering:

- Input the dataset.
- Tokenize the input data.
- Implement the lemmatization i.e., generating the dictionary word.
- Remove the stop words.
- Generate the TFIDF.
- Determine the KMeans.

LDA vs KMean Clustering:

S NO	Latent Dirichiet Allocation	Kmean Clustering
1	Output is the collection of topics from the words in the datasets.	Generate the distinct topic collections
2	More realistic approach than Kmean.	Output is k disjoint clusters.

Classification: Classification is the extension of kmean clustering. There exists decision tree, naïve Bayes, random forest approach for classification. Below are the different classification approaches available.



- Decision tree builds classification models in the form of a tree structure.
- It breaks down a dataset into smaller and smaller subsets.

- Random Forest is an ensemble classifier made using many decision tree models.
- Ensemble models combine the results from different models.

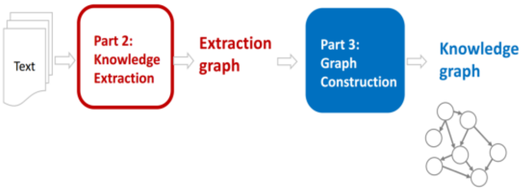
- It is a classification technique based on Bayes' Theorem with an assumption of independence among attributes.

Algorithm for classification:

- Input the dataset.
- Tokenize the input data.
- Implement the lemmatization i.e., generating the dictionary word.
- Remove the stop words.
- Generate the TFIDF.
- Process one of the above classification approach.

3.1.5 Generation of Knowledge Graph (KG): Construction Knowledge Graph is used to simplify the search results. This graph represents the graphical representation of the flow of the text data. The main advantage of using this knowledge graph is simplified diagrammatical representation of the huge data, helps in easy knowledge transfer and documentation easy.

Workflow of Knowledge Graph:

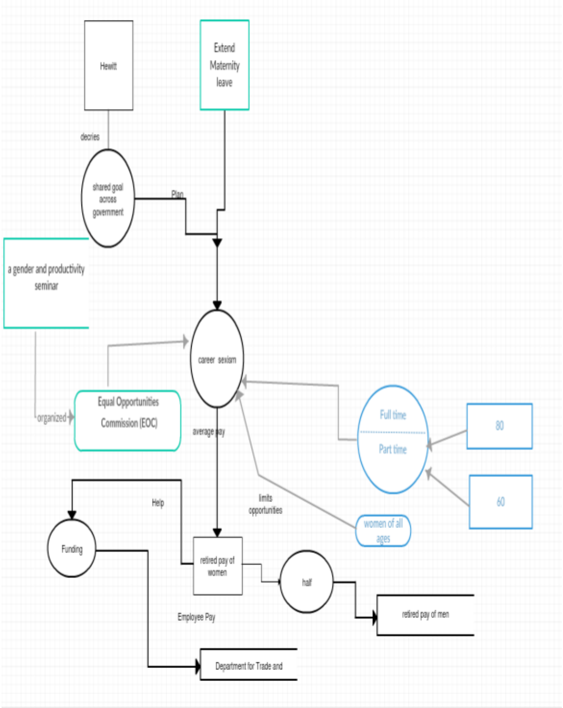


Algorithm for knowledge graph:

- Recognizing the named entity reference including the people, organization, location, date etc.
- Extracting the Classes, Subclasses, Triplets.
- Designing the data schema i.e., finding the relationship between these entities including data properties, object properties.
- Constructing the owl file for data set.
- Representing them in diagrammatical graph using protégé tool or webVowl.

Knowledge Graph for our dataset:

We do not have any specified rules for designing this knowledge graph. Different companies have their own knowledge graph construction and follows their own rules. We first recognized the entities in our dataset and designed the data schema to generate the relationships between the entities. Finalized the flow of data. Below is the diagrammatical representation of the knowledge graph that is designed for our datasets.



3.1.6 Querying for Data:

Spark query or DL query are querying types we are using our application. Constructing a query for a question to extract the answer is very fast and gives us the high accuracy answer.

i. Spark Query: Spark Sql is the structured query language which is used to query in the spark language program. This is like the general query language. To fetch the answer for a question who are the people in the community whose occupation is student can be written as below.

```
SELECT ?persons
WHERE
{ ?persons x:hasOccupation ?Occupation}
group by ?Occupations=student
```

ii. DL Query: DL query is the simplified version implemented in protégé tool to fetch the instances for the question. It is more simple and fast. To fetch the instances of people whose occupation is student can be written as:

```
hasOccupation value “student”
```

3.1.7 Question-Answering

i. A Question-Answer Set for our Dataset.

We are designing the questions from datasets considering mainly the PERSON, LOCATION, ORGANIZATION, NUMBER entity.

- When was Obama born?

Born on Aug. 4, 1961.

- Where did Obama did his schooling?

Punahou School.

- Who is father of Obama?

Barack Hussein Obama.

- Whom did Obama compete in primary race? Hillary Rodham Clinton.
- What is the minimum duration for maternity leave?

6 months.

- What is the topic about?

career sexism.

- Who is the speaker?

Ms. Hewitt.

- What is the average pay for full-time women.

80p

- What is the average pay for part-time women.

60p.

- What is the average pay for retired women compared to men?

Half.

Knowledge Graph to extract answer:

Knowledge Graph is the graphic representation with instances of the properties between the entities.

Algorithm:

- Input the dataset for which we want to construct the knowledge graph.
- Generate the entities i.e., classes, subclasses, data properties, object properties, Triplets.
- Construct the owl.
- Visualize using the protégé vowl plugin or webvowl online.

Querying for answering:

We generate the query for the question using either DL Query or SPARQL and execute. This will fetch us the answer for the question either in the table form for SPARQL or the instances for DL query.

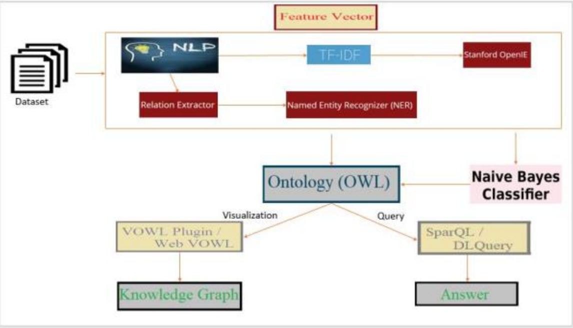
Algorithm Querying:

- Construct the. Owl for the dataset and generate the knowledge graph.
- Define the question for which we are expecting the answer.
- Construct the query.
- Execute in protégé for the answer.

4. IMPLEMENTATION AND EVALUATION

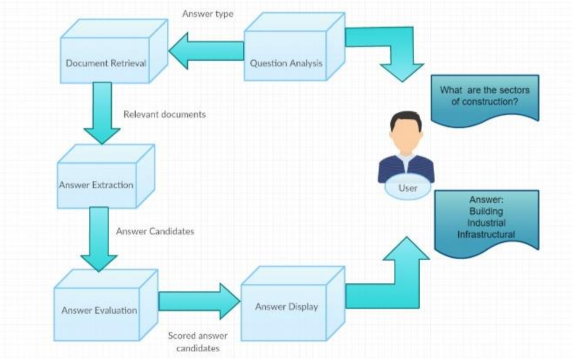
4.1 System Design and Implementation

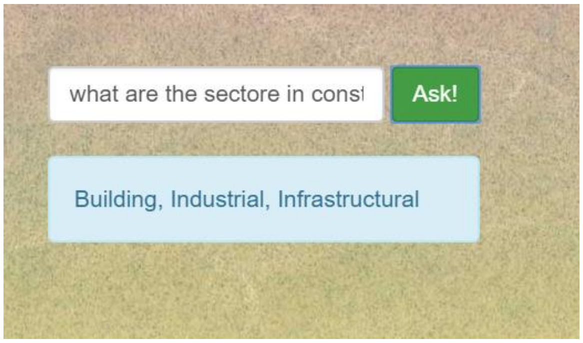
4.1.1 Software Architecture: The below diagram describes the system architecture of the Q/A system.



We need to perform the NLP on data that we choose and then generating the feature vector and then extraction of triplets using the OpenIE and classification using the Algorithms and finally generation of the owl file.

4.1.2 Implementation details: We have considered data from 3 datasets Yahoo, Stanford and WikiRef220 and applied the preprocessing on those. We have applied the TF-IDF on top of it to retrieve the important words and used the word2vec to find the similarity of the words. We have applied the OpenIE techniques for the dataset and extracted the RDF triplets and applied wordnet on the subject and predicate to find the synonyms of those. Also Implemented the OpenIE on the questions and compared the subject of the questions and then displayed the matched answers from the triplet extracted from the dataset. Also used the k-means clustering for classifying the questions into the clusters of related ones. Applied Naivebayes for the dataset to compare the accuracy. We have constructed the knowledge graph for the data dynamically and visualized it through Webowl plugin and performed the Sparql and Dl queries on it.

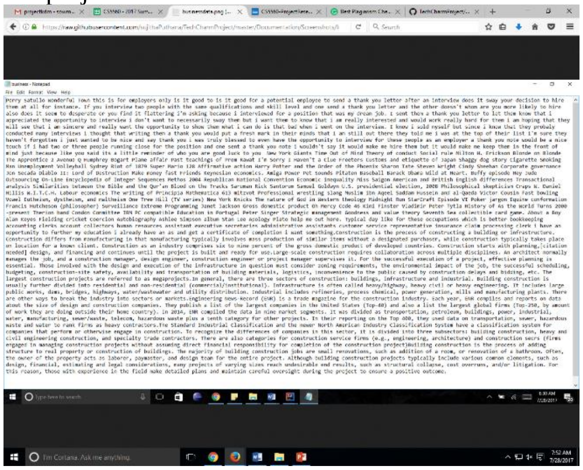



what are the sectore in cons
Ask!
Building, Industrial, Infrastructural

GitHub Link:
https://github.com/sujithaPuthana/TechCharmProject

4.1.3 Applications:

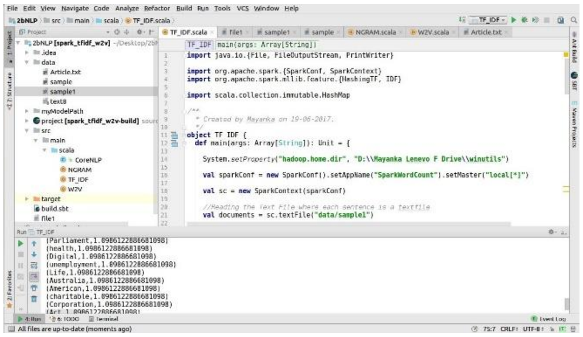
The diagram shown below will show the dataset of our project.



```
Perry satullo wonderful town this is for employers only is it good to is it good for a potential employee to send a thank you letter after an interview does it sway your decision to hire
them at all for instance. if you interview two people with the same qualifications and skill level and one send a thank you letter and the other doesn't whom are you more likely to hire
also does it seem to desperate or you find it flattering i'm asking because i interviewed for a position that was my dream job. i sent them a thank you letter to let them know that i
appreciated the opportunity to interview i don't want to necessarily sway them but i want them to know that i am really interested and would work really hard for them i am hoping that they
will see that i am sincere and really want the opportunity to show them what i can do is that bad when i went on the interview. I know i sold myself but since i know that they probaly
conducted many interviews i thought that writing then a thank you would put a fresh mark in their minds that i am still out there they told me i was at the top of their list i'm sure they
haven't forgotten i just wanted to be nice and say thank you i was truly blessed to even have the opportunity to interview for these people as an employer a thank you note would be a nice
touch if i had two or three people running close for the position and one sent a thank you note i wouldn't say it would make me hire them but it would make me keep them in the front of
mind just because like you said its a little reminder of who you are good luck to you  New York Giants Time Out of Mind Theory of conduct Social rule Milton H. Erickson Blonde on Blonde
The Apprentice 2 Avenue Q Humphrey Bogart Plame affair Past teachings of Prem Rawat I'm Sorry I Haven't a Clue Freeters Customs and etiquette of Japan Shaggy dog story Cigarette Smoking
Man Unemployment Volleyball Sydney Riot of 1879 Super Mario 128 Affirmative action Harry Potter and the Order of the Phoenix Sharon Tate Steven Wright Cindy Sheehan Corporate governance
Jon Secada Diablo II: Lord of Destruction Make money fast Friends Keynesian economics. Amiga Power Pet Sounds Pitaten Baseball Barack Obama Wild at Heart. Buffy episode Hey Jude
Outsourcing On-Line Encyclopedia of Integer Sequences Methos 2004 Republican National Convention Economic inequality Miss Saigon American and British English differences Transactional
analysis Similarities between the Bible and the Qur'an Blood on the Tracks Saruman Rick Santorum Samuel Goldwyn U.S. presidential election, 2008 Philosophical skepticism Craps W. Daniel
Hillis W.I.T.C.H. Labour economics The writing of Principia Mathematica 613 mitzvot Professional wrestling slang Muslim ibn Aqeel Saddam Hussein and al-Qaeda Victor Cousin Fast bowling
Vowel Eutheism, dystheism, and maltheism One Tree Hill (TV series) New York Knicks The nature of God in Western theology Midnight Run StarCraft Episode VI Poker jargon Equine conformation
Francis Hutcheson (philosopher) Surveillance Extreme Programming Janet Jackson Gross domestic product Oh Mercy Code 46 Kind Finster Vladimir Peter Tytla History of As the World Turns 2000
-present Therion band Condon Committee IBM PC compatible Education in Portugal Peter Singer Strategic management Goodness and value theory Seventh Sea collectible card game. About a Boy
Alan Keyes Fielding cricket Coercion Autobiography Ashlee Simpson album Stan Lee Apology Plato help me out here. Typical day like for these occupations which is better bookkeeping
accounting clerks account collectors human resources assistant executive secretaries administrative assistants customer service representative insurance claim processing clerk i have an
opportunity to further my education i already have an as and get a certificate of completion i want something.Construction is the process of constructing a building or infrastructure.
Construction differs from manufacturing in that manufacturing typically involves mass production of similar items without a designated purchaser, while construction typically takes place
on location for a known client. Construction as an industry comprises six to nine percent of the gross domestic product of developed countries. Construction starts with planning,[citation
needed] design, and financing and continues until the project is built and ready for use.Large-scale construction requires collaboration across multiple disciplines. An architect normally
manages the job, and a construction manager, design engineer, construction engineer or project manager supervises it. For the successful execution of a project, effective planning is
essential. Those involved with the design and execution of the infrastructure in question must consider zoning requirements, the environmental impact of the job, the successful scheduling,
budgeting, construction-site safety, availability and transportation of building materials, logistics, inconvenience to the public caused by construction delays and bidding, etc. The
largest construction projects are referred to as megaprojects.In general, there are three sectors of construction: buildings, infrastructure and industrial. Building construction is
usually further divided into residential and non-residential (commercial/institutional). Infrastructure is often called heavy/highway, heavy civil or heavy engineering. It includes large
public works, dams, bridges, highways, water/wastewater and utility distribution. Industrial includes refineries, process chemical, power generation, mills and manufacturing plants. There
are other ways to break the industry into sectors or markets.Engineering News-Record (ENR) is a trade magazine for the construction industry. Each year, ENR compiles and reports on data
about the size of design and construction companies. They publish a list of the largest companies in the United States (Top-40) and also a list the largest global firms (Top-250, by amount
of work they are doing outside their home country). In 2014, ENR compiled the data in nine market segments. It was divided as transportation, petroleum, buildings, power, industrial,
water, manufacturing, sewer/waste, telecom, hazardous waste plus a tenth category for other projects. In their reporting on the Top 400, they used data on transportation, sewer, hazardous
waste and water to rank firms as heavy contractors.The Standard Industrial Classification and the newer North American Industry Classification System have a classification system for
companies that perform or otherwise engage in construction. To recognize the differences of companies in this sector, it is divided into three subsectors: building construction, heavy and
civil engineering construction, and specialty trade contractors. There are also categories for construction service firms (e.g., engineering, architecture) and construction secrs (firms
engaged in managing construction projects without assuming direct financial responsibility for completion of the construction project)Building construction is the process of adding
structure to real property or construction of buildings. The majority of building construction jobs are small renovations, such as addition of a room, or renovation of a bathroom. Often,
the owner of the property acts as laborer, paymaster, and design team for the entire project. Although building construction projects typically include various common elements, such as
design, financial, estimating and legal considerations, many projects of varying sizes reach undesirable end results, such as structural collapse, cost overruns, and/or litigation. For
this reason, those with experience in the field make detailed plans and maintain careful oversight during the project to ensure a positive outcome.
```

Pseudocode:

Generating the term frequency for the words in the dataset.

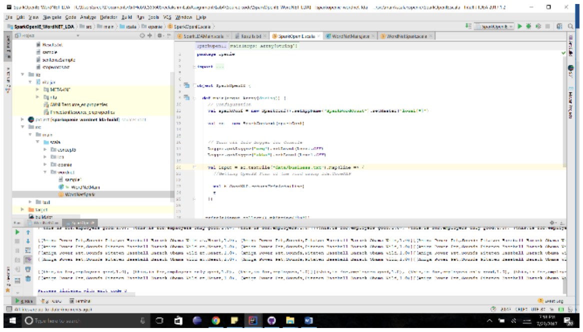


```
import java.io.{File, FileOutputStream, PrintWriter}

import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.mllib.feature.{HashingTF, IDF}

import scala.collection.immutable.HashMap

/**
  * Created by Mayanka on 19-06-2017.
  */
object TF_IDF {
  def main(args: Array[String]): Unit = {

    System.setProperty("hadoop.home.dir", "D:\\Mayanka Lenevo F Drive\\winutils")

    val sparkConf = new SparkConf().setAppName("SparkWordCount").setMaster("local[*]")

    val sc = new SparkContext(sparkConf)

    //Reading the Text File where each sentence is a textfile
    val documents = sc.textFile("data/sample1")
```

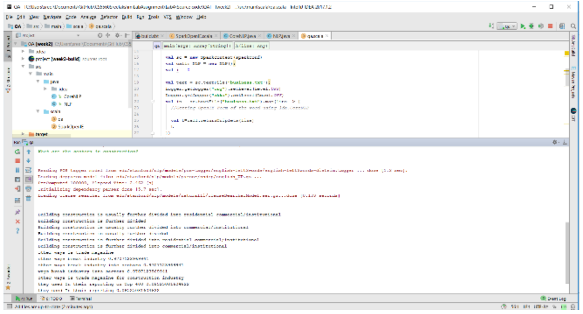
```
(Parliament,1.0986122886681098)
(health,1.0986122886681098)
(Digital,1.0986122886681098)
(unemployment,1.0986122886681098)
(Life,1.0986122886681098)
(Australia,1.0986122886681098)
(American,1.0986122886681098)
(charitable,1.0986122886681098)
(Corporation,1.0986122886681098)
(Art,1.0986122886681098)
```

All files are up-to-date (moments ago)

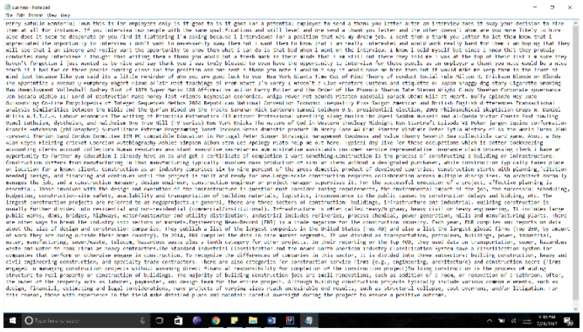
OpenIE for our dataset for triplet extraction:



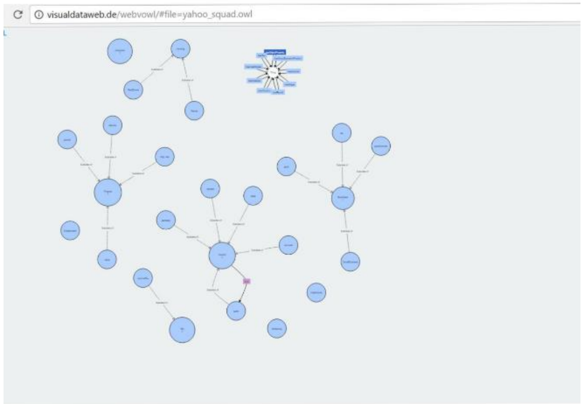
QA for our dataset:



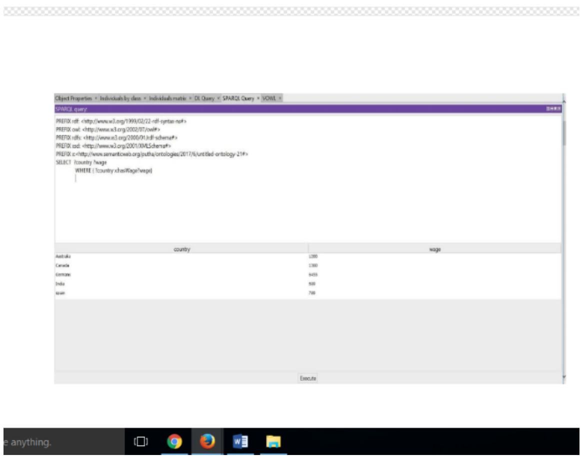
We have stored the results into the text file.



Visualization using the webowl plugin



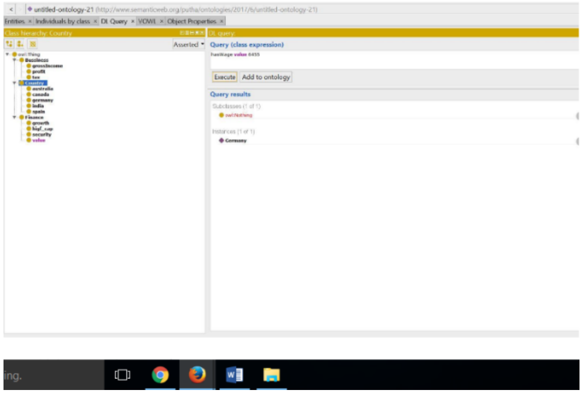
Below mentioned pictorial representation indicates the SparQL and DL queries respectively.



Object Properties × Individuals by class × Individuals matrix × DL Query × SPARQL Query × VOWL ×
SPARQL query:
Execute
e anything.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX x:<http://www.semanticweb.org/putha/ontologies/2017/6/untitled-ontology-21#>
SELECT ?country ?wage
	WHERE { ?country x:hasWage?wage}
```

country	wage
Australia	1200
Canada	1300
Germany	6455
India	900
spain	700



hasWage value 6455

Video Link: https://github.com/sujithaPuthana/TechCharmProject/blob/master/Documentation/PPT/Project%20Video.mp4

4.2 Evaluation and Results:

We have the performance of our Q/A system after running like how much accuracy we got for the correct answer given by the system. The below table indicates the count of triplets, classes etc. those were obtained.

Classes	12
Triplets	867
Individuals	359
Object Properties	158
Data Properties	180

4.2.1 Evaluation plan:

4.2.1.1 Datasets:

Stanford Dataset: Stanford question answer dataset is the collection of the data from Wikipedia. All this data was collected from real time question answer from google. This data set consists of more than 100000 pairs of question answer from more than 500 articles regarding the business domain. https://rajpurkar.github.io/SQuAD explorer/explore/1.1/dev/Construction.html

Yahoo Dataset: Yahoo dataset are the collection of question answer pairs from yahoo community forum.

https://www.yahoo.com/?err=404&err_url=https%3a%2f%2fanswers.yahoo.com%2f%29

WikiRef220:WikiRef220 is the collection of the news article, taken from the Wikipedia pages. This dataset includes the information in the form of text data. The articles included in this dataset are developing construction business, materials used, tax.

http://mklab.iti.gr/files/WikiRef_dataset.zip

4.2.1.2 System Specifications:

Environment: Integrated development environments, helps in easy development of software with the facility of comprehensive integrated environment. IntelliJ, Pycharm and Protégé.

Languages: We have collaborated various languages in the development of the project and in building the application. Some of them are, Java, Scala and Spark.

Dataset size: 4 MB

4.2.1.3 Measurements:

We have tested approximately 30 questions from the dataset and noticed all the measurements related to time, speed and accuracy. Constructed knowledge graph dynamically for the 4MB data. We didn't faced any huge issue while running the algorithms with the memory of our system. We have 12 classes, 867 triplets, 359 individuals,158 object properties,180 data properties. The accuracy of the dataset while implementing the Naïve Bayes algorithm is 0.5

4.2.2 Comparative Evaluation for Q/A approach:

The below table indicates the evaluation for our dataset.

Questions tested	30
Exact result obtained	17
Processing Time	13sec
Accuracy obtained	60%

4.2.3 Comparative evaluation with others: The system we have developed is unique in Business domain. It will give the results for the queries existing in this particular domain. We can perform any queries regarding tax, construction issues, wages, profits etc. So we can get more accurate domain related answers when compared to the other systems which give non-domain specific answers. We have extracted the triplets dynamically and also manually which helps in getting accurate results.

5. DISCUSSION AND LIMITATIONS

As the accuracy estimation of the system is 60% which indicates there are some irrelevant information obtained while querying process. When we query using the knowledge graph which was constructed dynamically the results are not accurate as there is a completeness issue. In order to address this we need to add the properties manually. However, when we query on the Stanford dataset the results are more accurate when compared to the yahoo set because the data in Stanford is more precisely mentioned. We need to include the classification algorithms for better accuracy and also topic discovery to categorize the answers for all relevant retrievals and we need to work more on perfect dynamic ontology creation.

6. CONCLUSION

From this paper we developed a dynamic intelligent question answering system for business and construction related domain. With the Convention of the NLP, TF-IDF then OpenIE and Wordnet it is easy to build a Q/A approach. We also built knowledge graph using protégé which helps in easy understanding of the system by visualizing it.

REFERENCE

1. https://blog.algorithmia.com/introduction-natural-language-processing-nlp/ https://en.wikipedia.org/wiki/Question_answering
2. https://nlp.stanford.edu/
3. http://visualdataweb.de/webvowl/
4. https://rajpurkar.github.io/SQuAD-explorer/
5. https://cogcomp.cs.illinois.edu/page/resource_view/89
6. https://protegewiki.stanford.edu/wiki/Main_Page
7. http://ai.stanford.edu/~ang/papers/nips01-lda.pdf
8. https://www.cs.umd.edu/~mount/Projects/KMeans/pami02.pdf
9. https://nlp.stanford.edu/pubs/chen2017reading.pdf
10. https://nlp.stanford.edu/software/
11. http://stanford.edu/~cpiech/cs221/handouts/kmeans.html
12. https://nlp.stanford.edu/pubs/llda-emnlp09.pdf
13. https://nlp.stanford.edu/software/openie.html
14. https://plato.stanford.edu/entries/logic-ontology/
15. http://protege.stanford.edu/publications/ontology_development/ontology101.pdf