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*

Project-2

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.

2. 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.

3. Significance

Using Question Answer system searching become easy and it will give 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.

4. 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.

5. 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.

Finally the Q/A Application give the answers for the questions posed by the humans in their language. This will improve the search results in that documents.

6. 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.

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

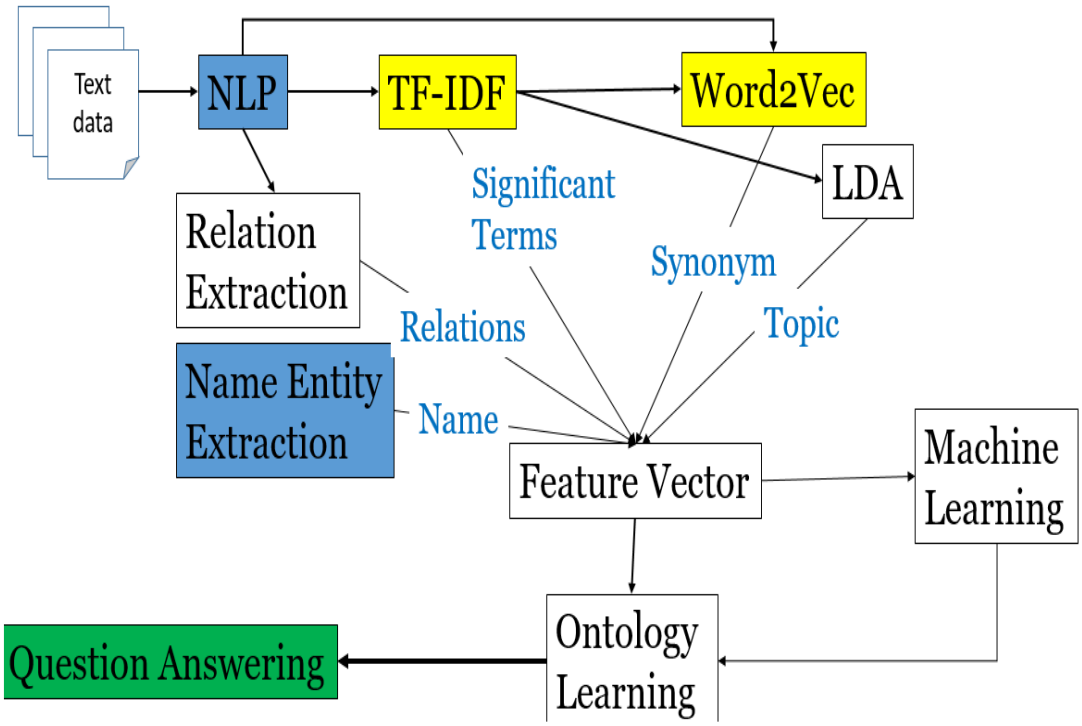
the next problem, where we have to find the relation between the entities. By using the Deep Dive concept we can get the solution for the above problem. Deep Dive helps for extracting and integrating the data in order to make the data training procedure easy.

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.

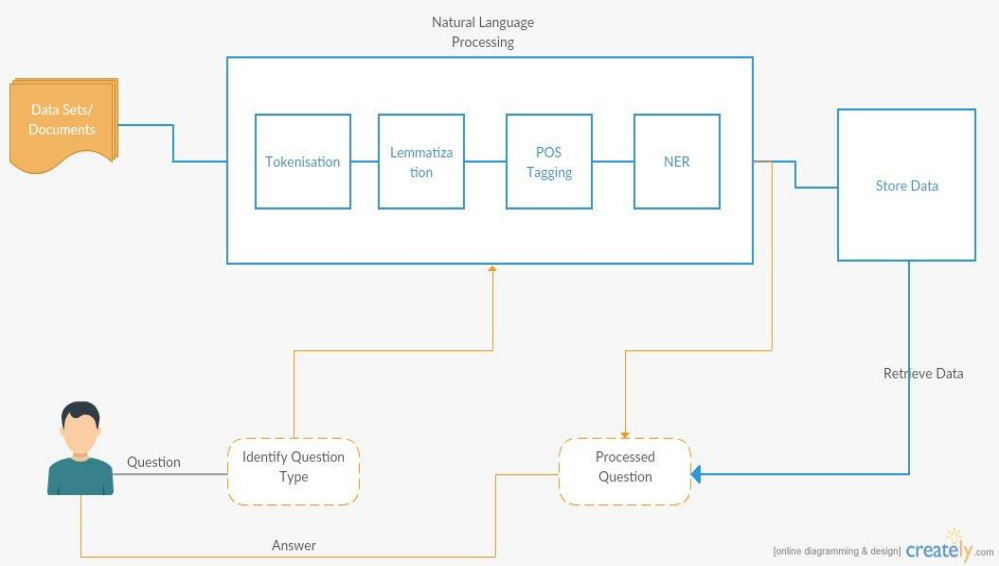
Founder is an framework for constructing the knowledge graph for richly formatted data. By using this framework we can handle the structured data and we can get the solutions for the data completeness and data correctness issues.

7. Design

a. Workflow

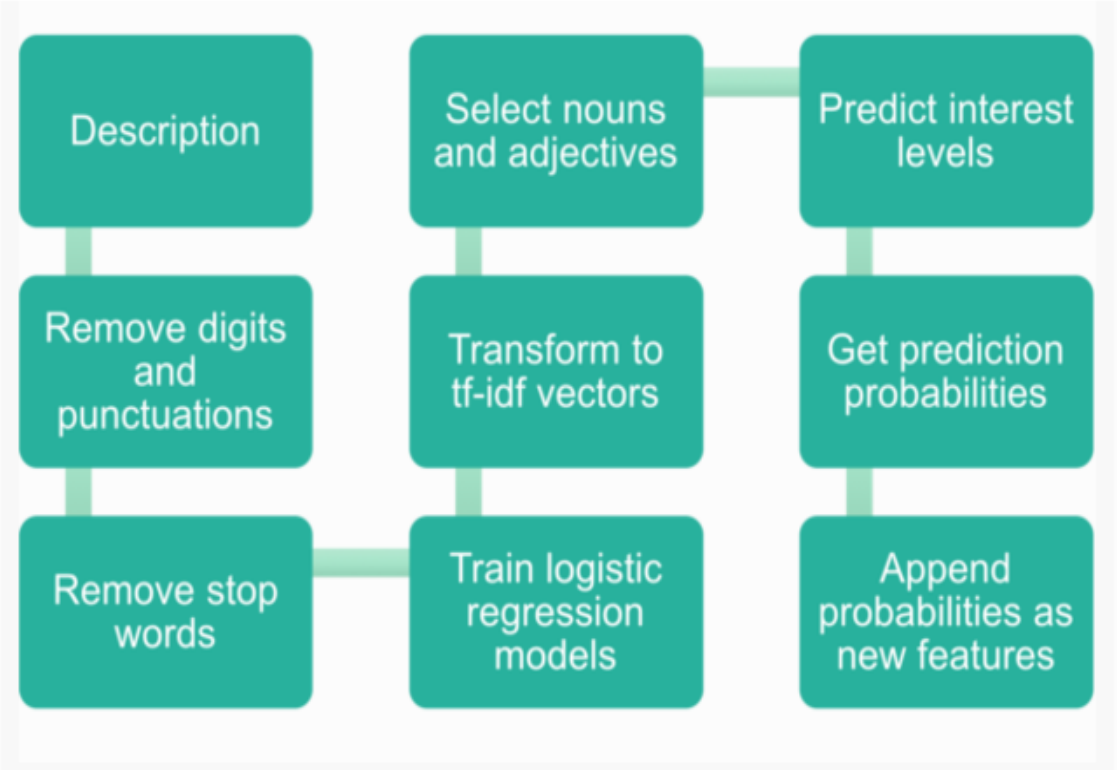


b. NLP



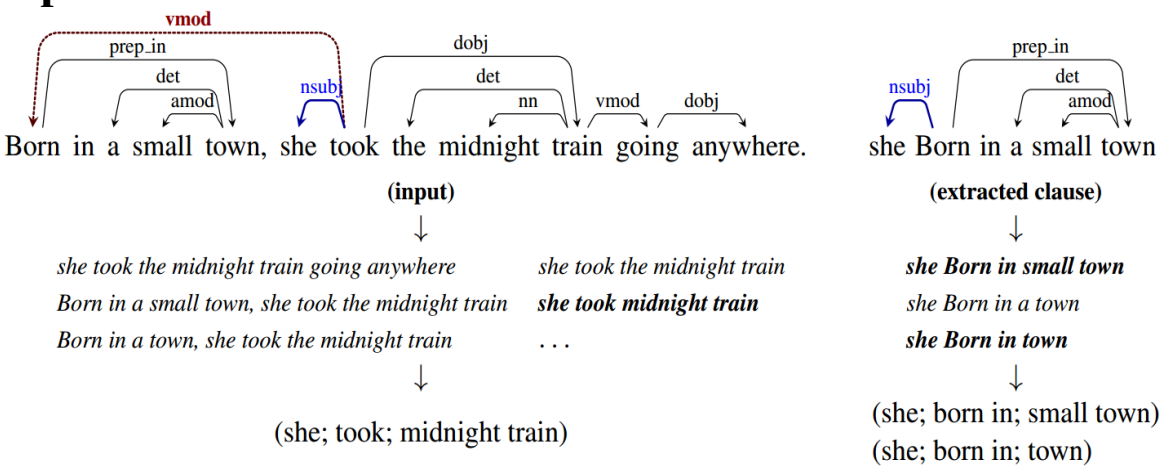
c. Information Retrieval (TFIDF, Word2Vec)

TFIdF:



d. Information Extraction (OpenIE, WordNet)

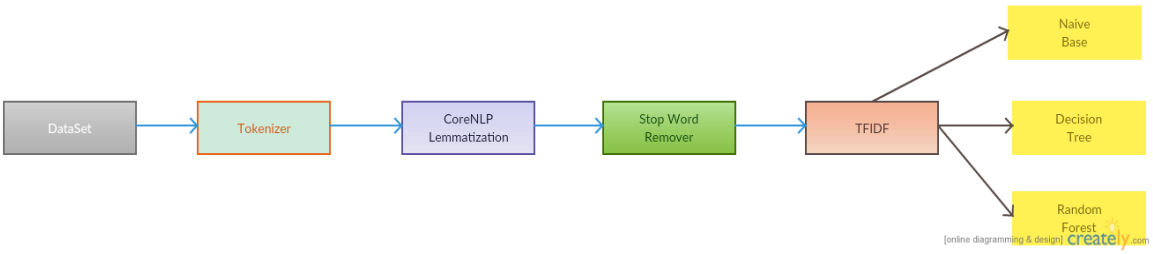
OpenIE:



e. Machine Learning

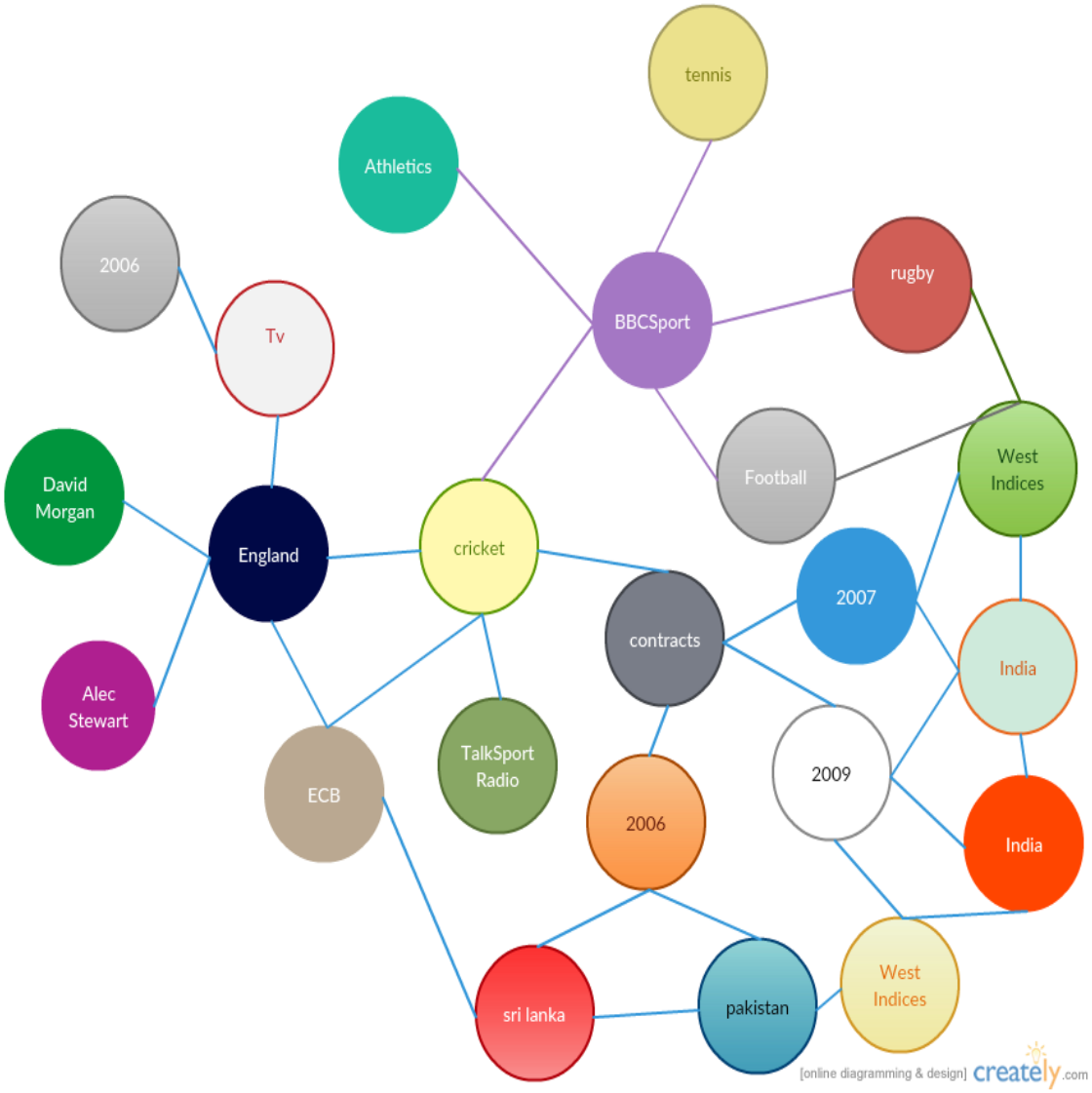
KMeans Clustering:

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

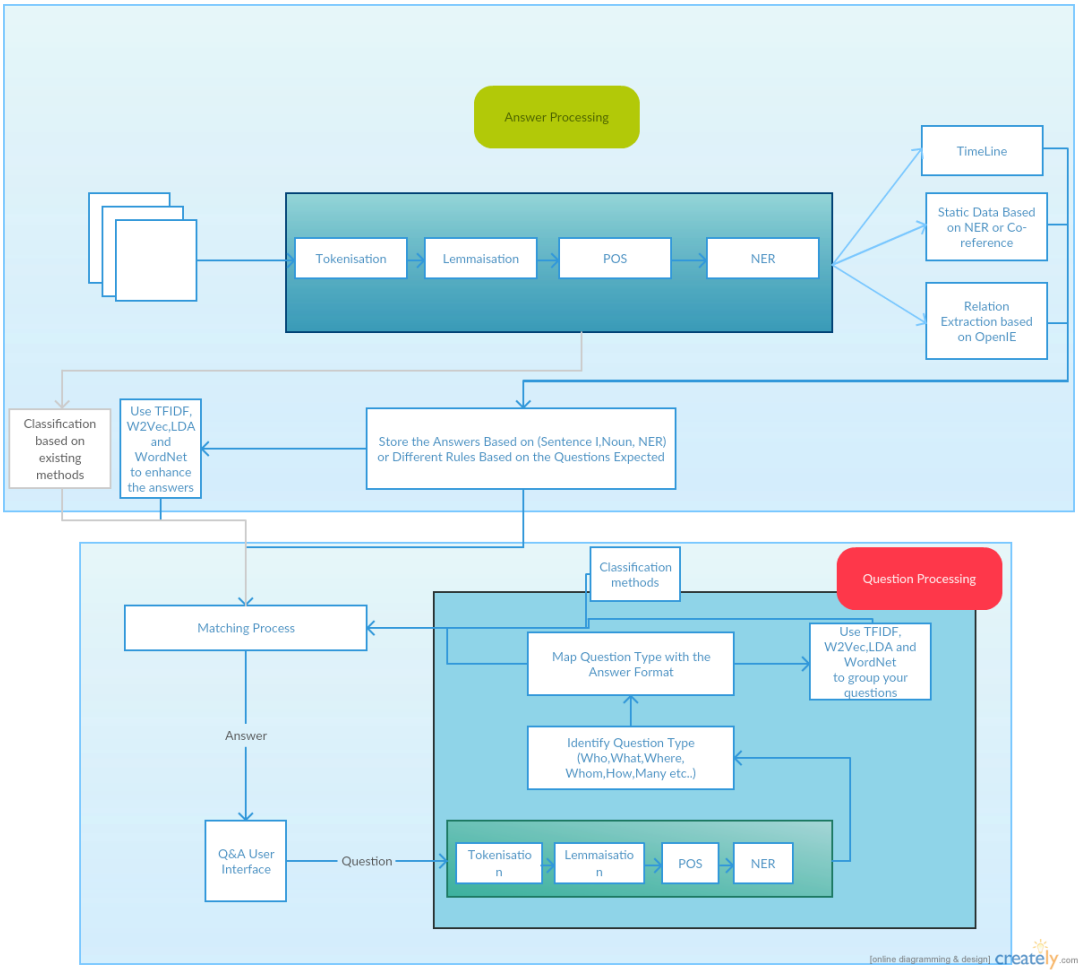


f. Knowledge Graph:

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



g. Questions and Answers

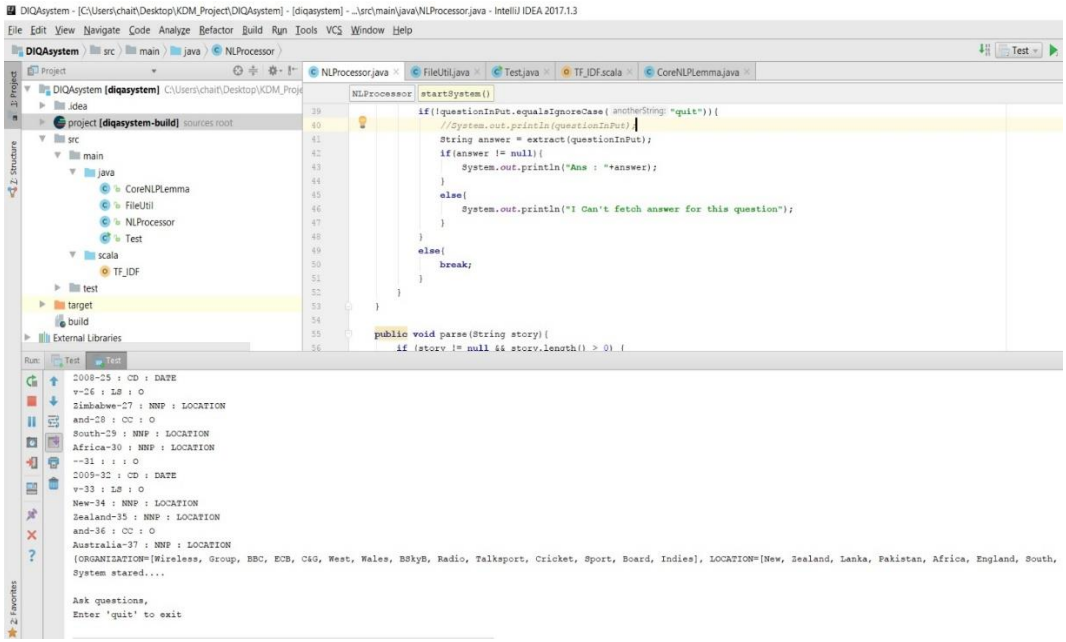


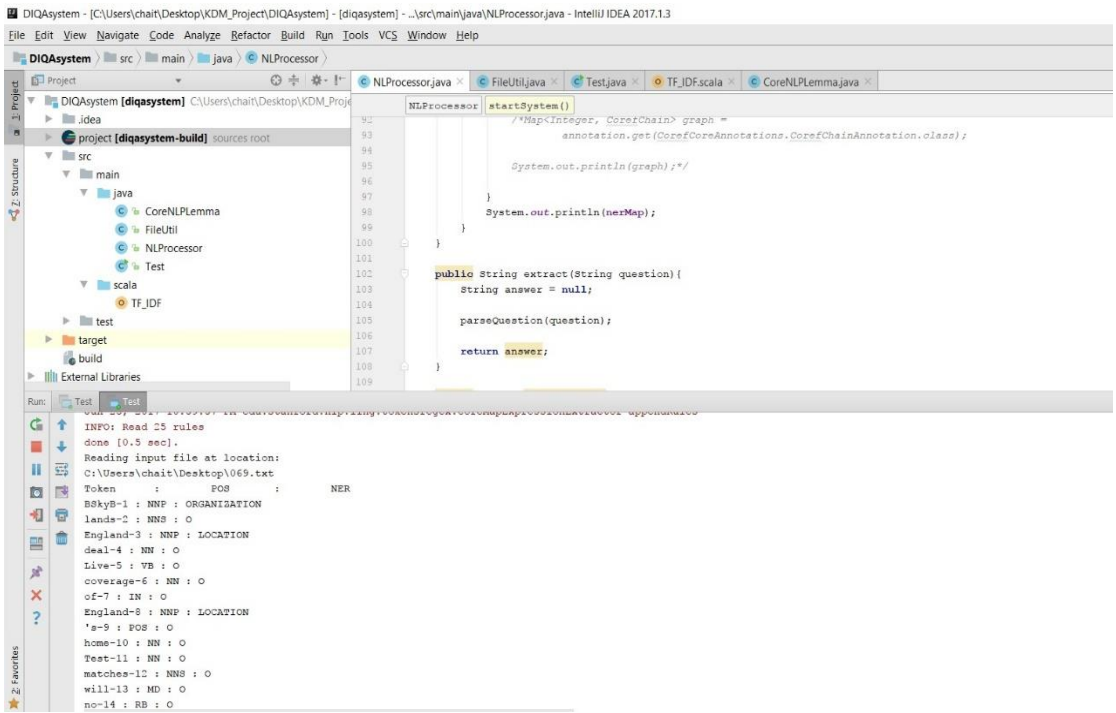
8. Implementation

a. Using NLP:

NLP is a procedure which we have used for processing information given in natural language by the user. Here, NLP takes the users natural language data as input and converts that data into the machine readable format.

Here are the screen Shots of the program and execution of the NLP for the chosen dataset:





b. Information Retrieval:

Using 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:

(Normalized) term frequency of term k_i in Document d_j

$$tf(i,j) = \frac{freq(i,j)}{\max_{k \in T} freq(k,j)}$$

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d}, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

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

Inverse document frequency of term k_i

$$idf(i) = \log(\frac{n}{n_i}) \in [0, n]$$

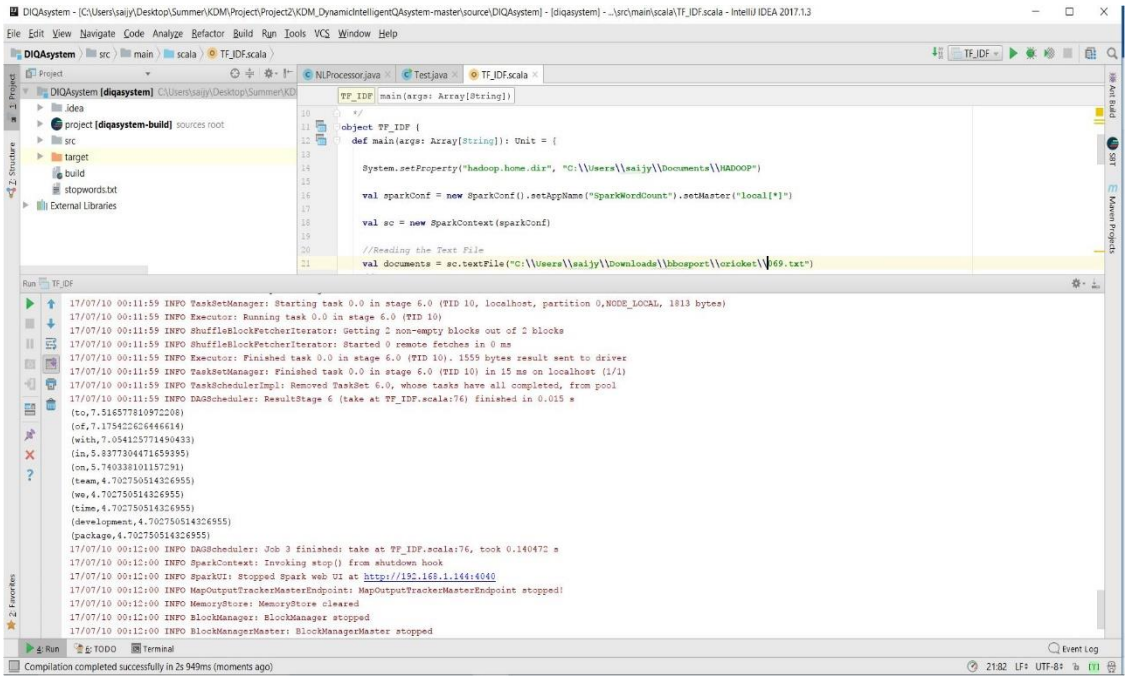
n_i number of documents in which term k_i occurs

Term weight $w_{ij} = tf(i,j)\ idf(i)$

TFIF weight of term:

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log \; N / \mathrm{df}_t$$

Here are the screen Shots of the program and execution of the Information retrieval using TFIDF for the chosen dataset:



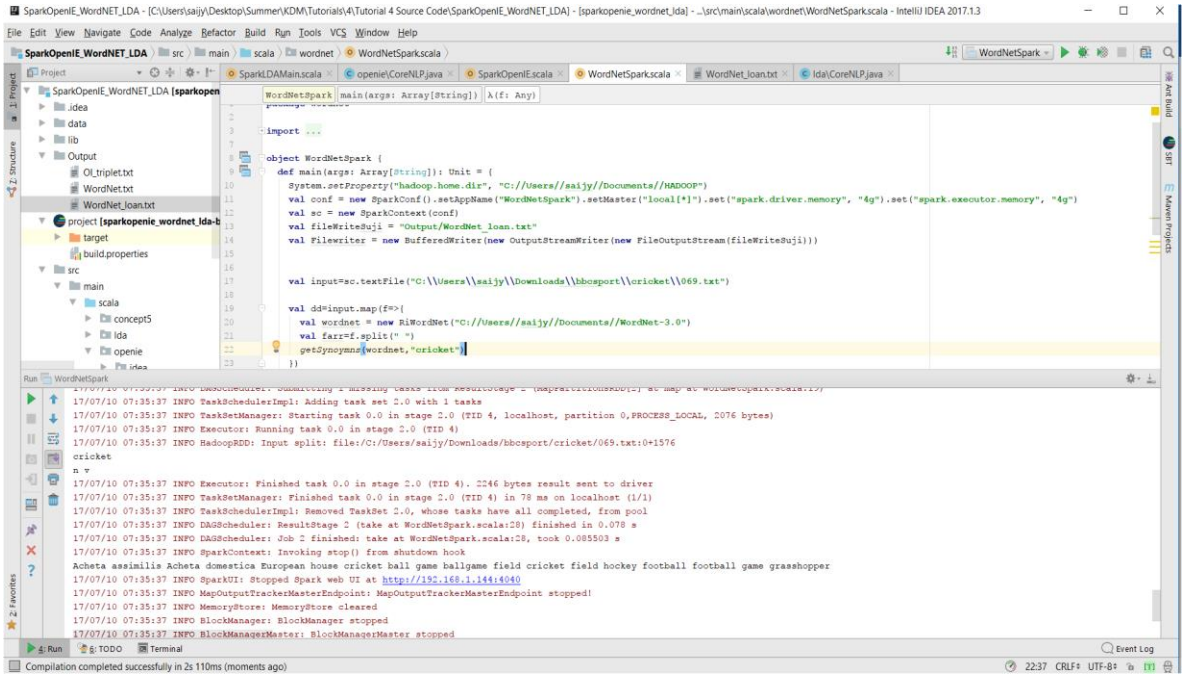
Using Word2Vec:

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.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

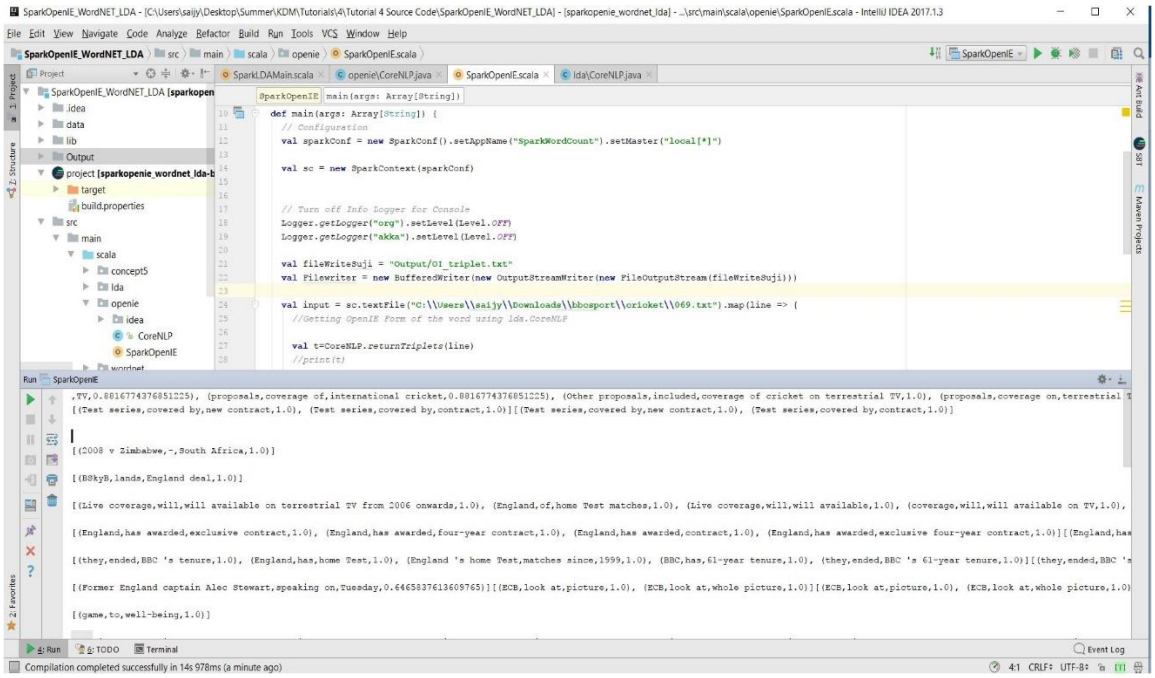
Here are the screen Shots of the program and execution of the Information retrieval using Word2Vec for the chosen dataset:



c. Information Extraction:

Using OpenIE:

Here are the screen Shots of the program and execution of the Information extraction using OpenIE for the chosen dataset:



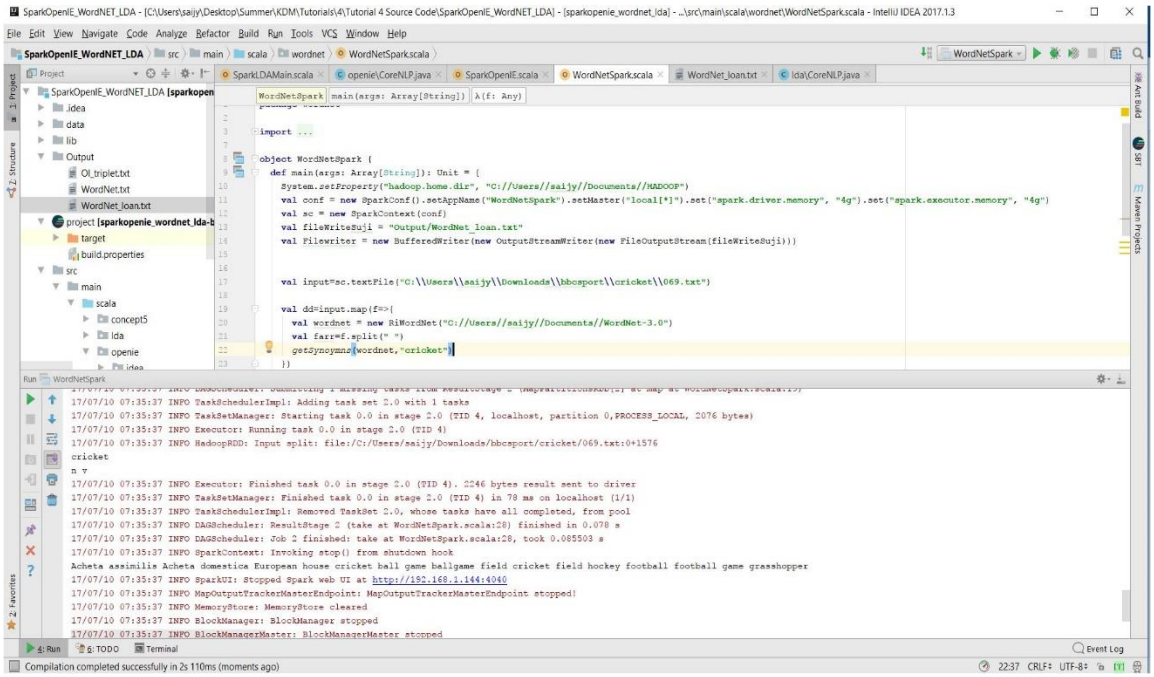
Using WordNet:

WordNet is a great English verbal catalogue. Nouns, verbs, adjectives and adverbs are assembled into groups of perceptive substitutes (synsets), separately stating a different idea. WordNet is implemented using RitaJS, which is a toolkit designed for experimentations in normal language. Languages used for the implementation are, a combined API for Java and JavaScript.

In this program, input is text file and we will specify a word in the program itself, then program use synsets, pos retival mechanisms and finally gives the output as the parts of

speech of that specified word and some of the definitions of the given according to the given dataset. For example If we give the word as eat, then it will give the output as V N and definitions for the eat as take in solid food, eat meal like that.

Here are the screen Shots of the program and execution of the Information extraction using WordNet for the chosen dataset:



d. Machine Learning:

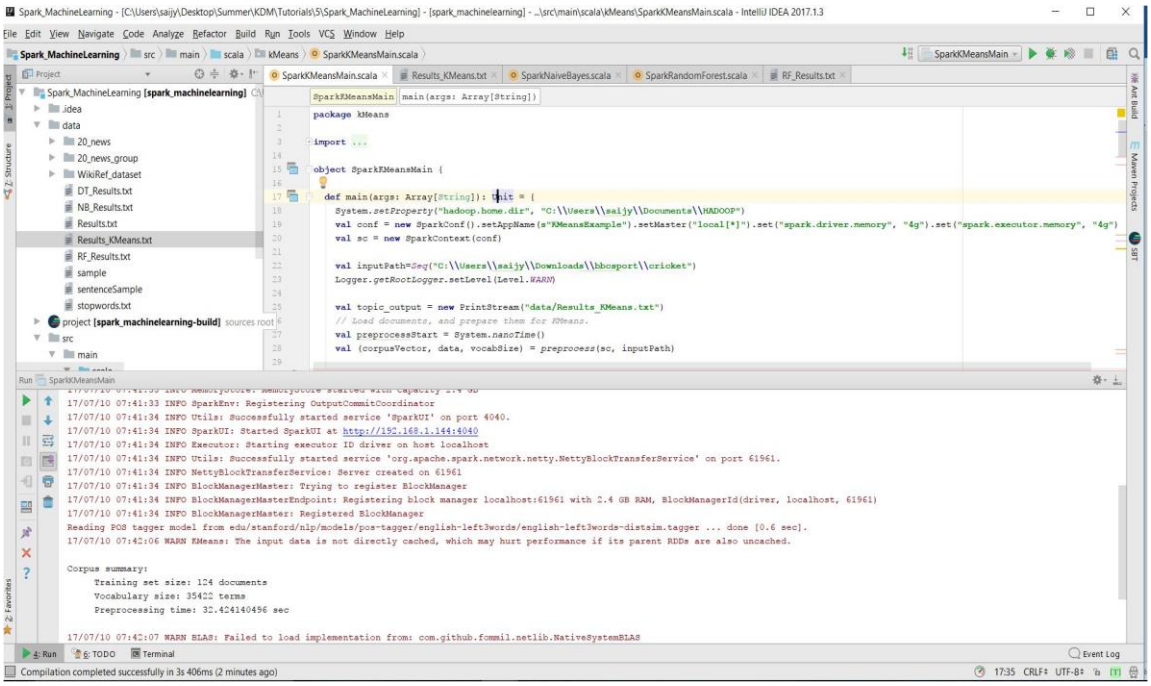
Clustering:

In this program we used KMeans for clustering. 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 inorder 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.

Here are the screen Shots of the program and execution of the Information extraction using WordNet for the chosen dataset:



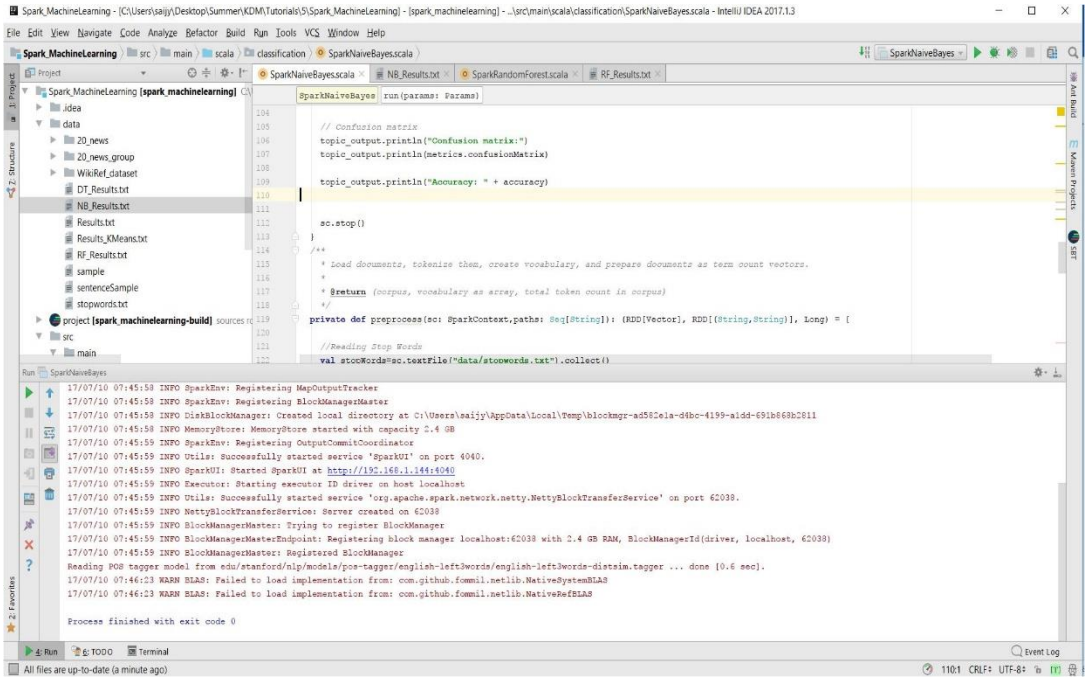
Classification:

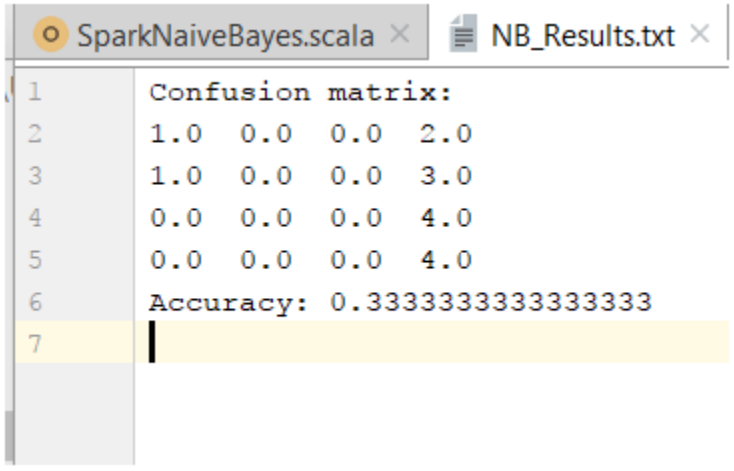
Classification is nothing but a problem of classifying the observations, to which category it goes, on the base of the training dataset.

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

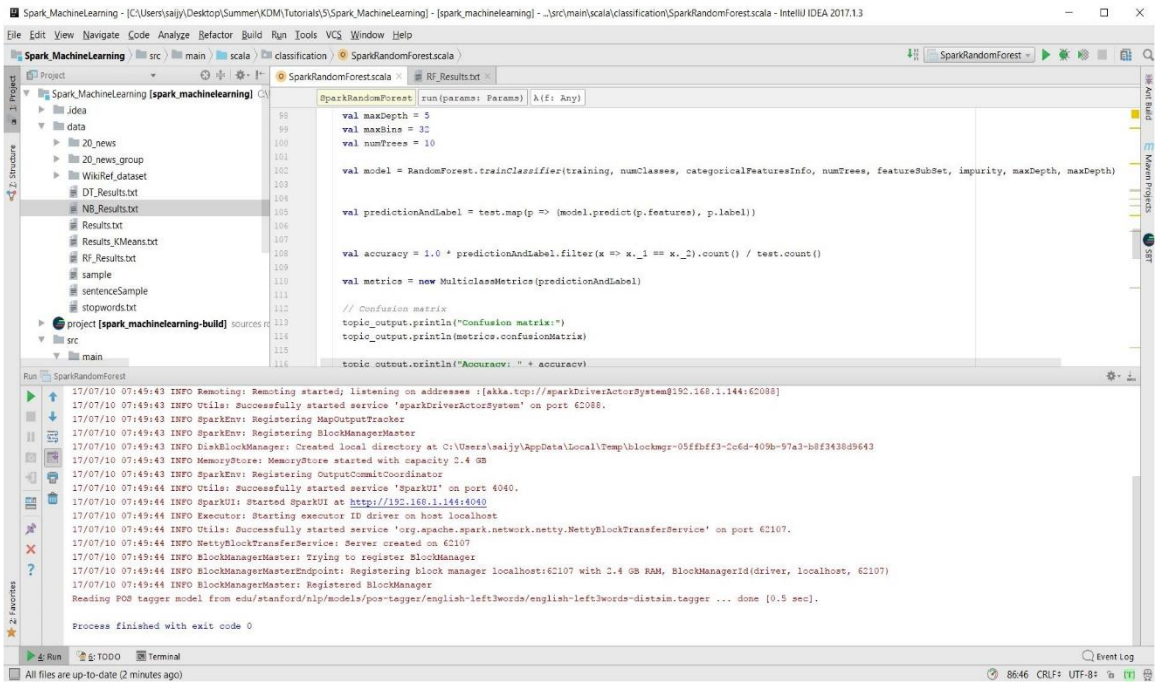
Here are the screen Shots of the program and execution of the Machine Learning clustering and classification for the chosen dataset:

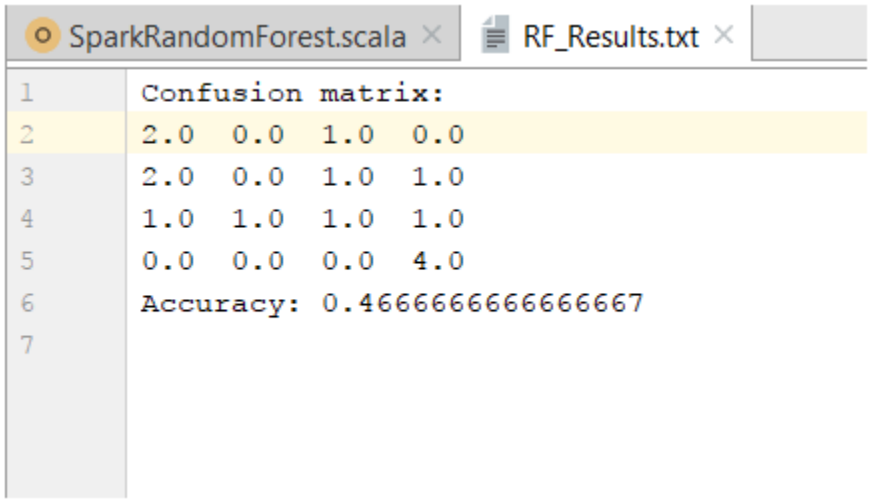
Naïve Base:



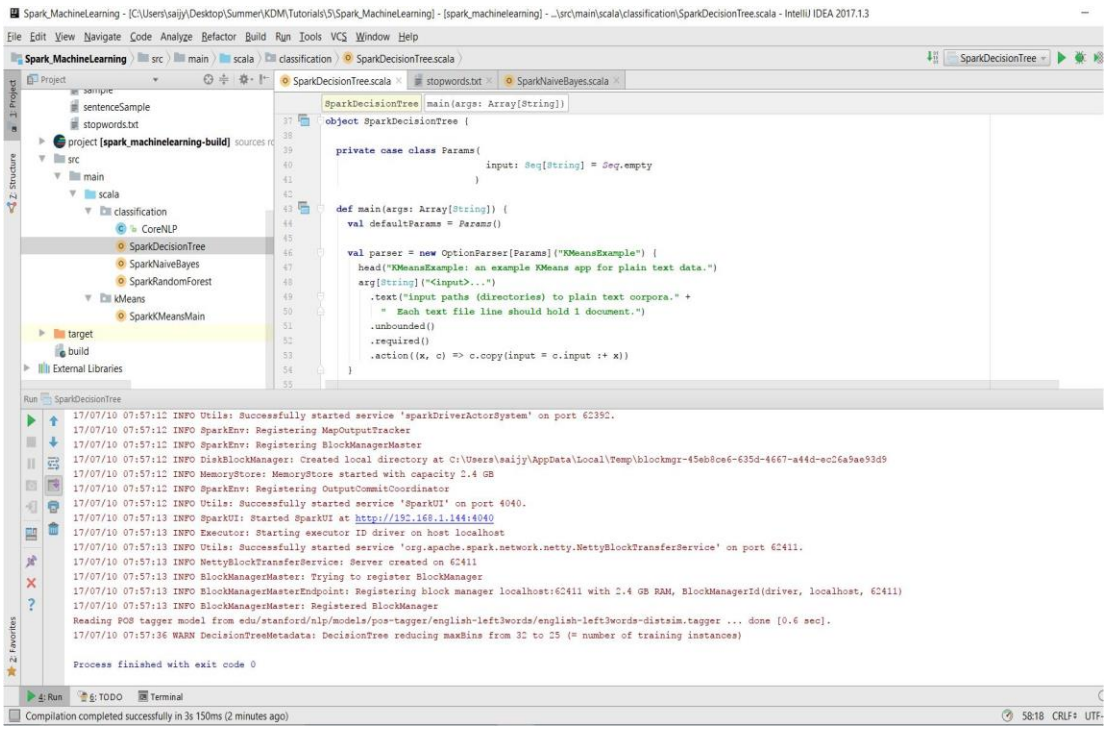


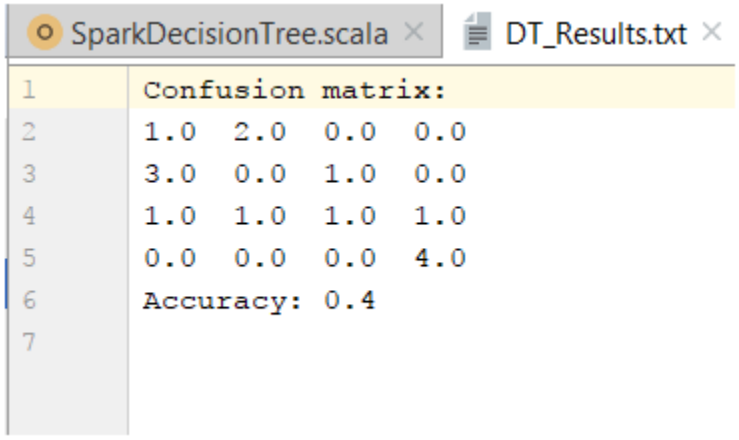
Random Forest:





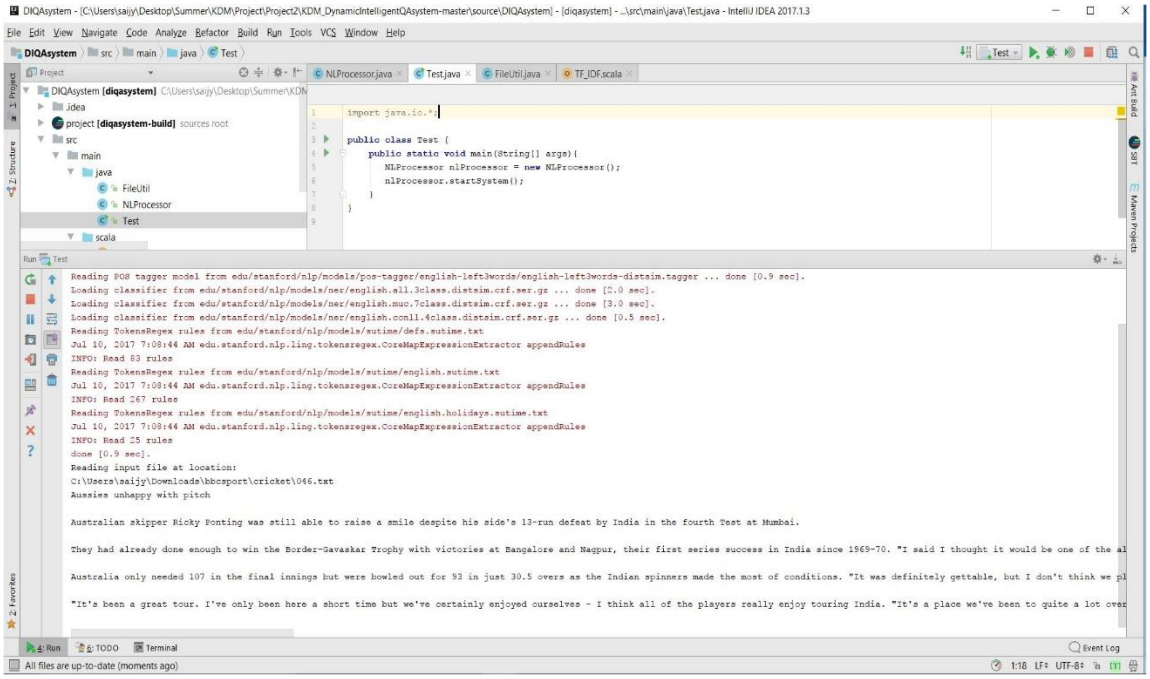
Decision Tree:

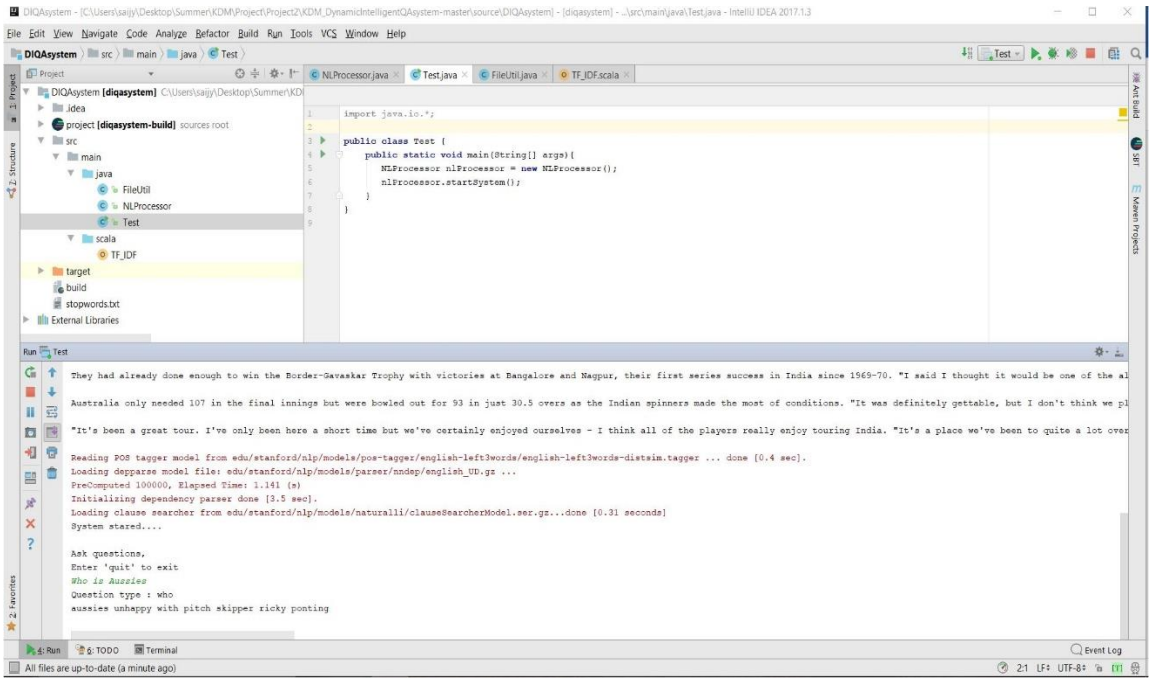




e. Question Answering:

Here are the screen Shots of the program and execution of the Question Answering for the chosen dataset:





9. Project Management

a. Work Completed

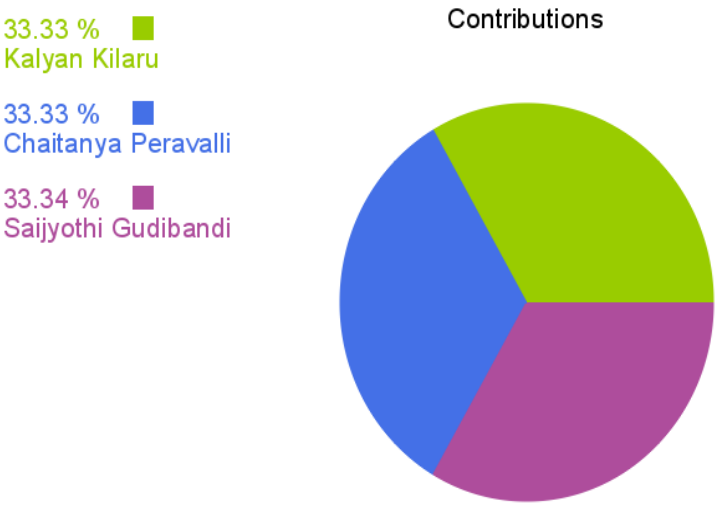
Work completed on NLP, TFIDF, Word2Vec, OpenIE, WordNet, Clustering, Classification, Question Answer system for the chosen dataset.

b. Contributions

SaiJyothi Gudibandi

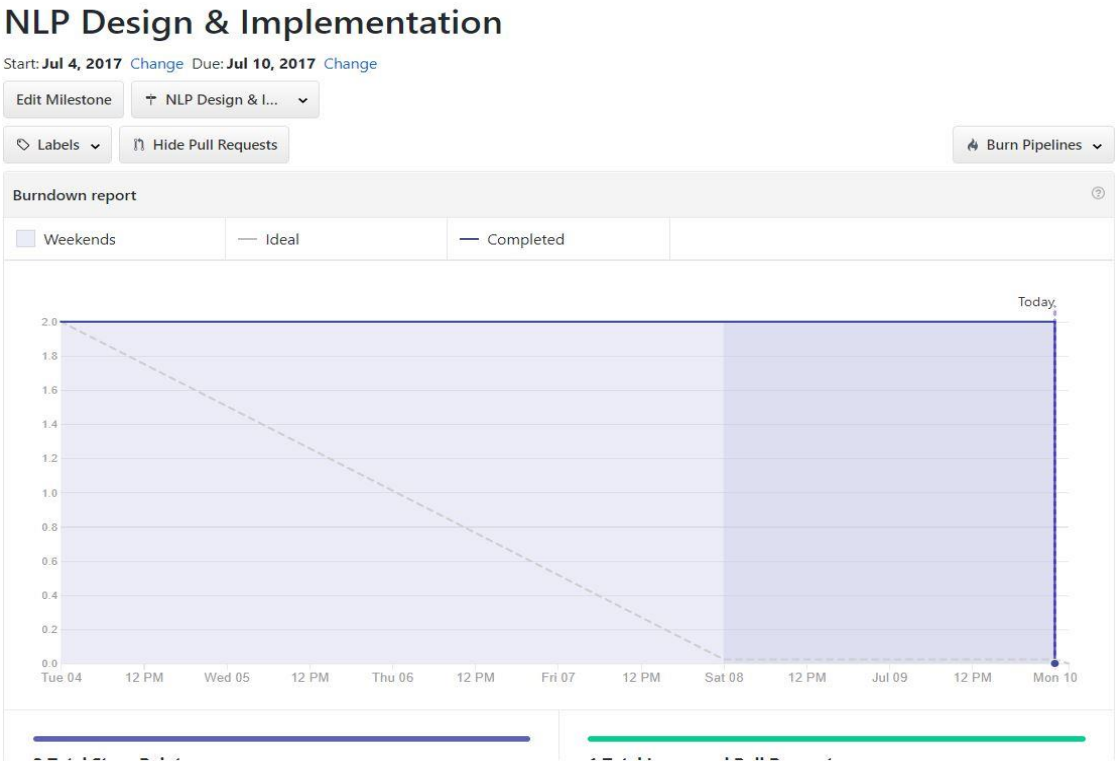
Kalyan Kilaru

Chaitanya Kumar Peravalli



c. Statics

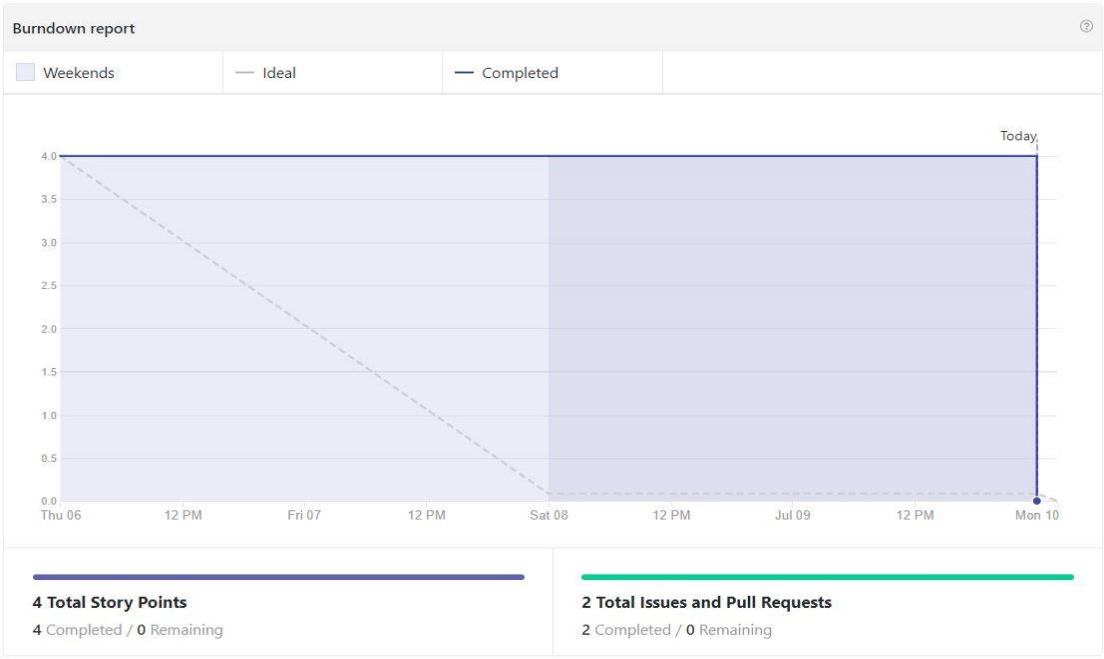
Here are the screen Shots Project Management:



Knowledge Graph & Questions and Answers

Start: **Jul 6, 2017** Change Due: **Jul 10, 2017** Change

Edit Milestone | Knowledge Gra... | Labels | Hide Pull Requests | Burn Pipelines



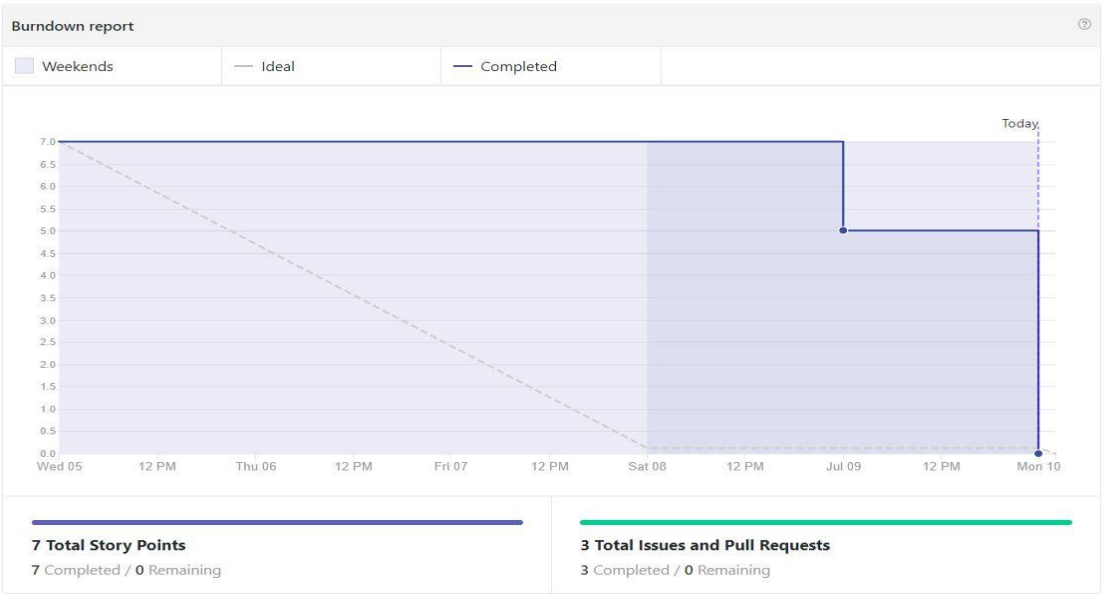
4 Total Story Points
4 Completed / **0** Remaining

2 Total Issues and Pull Requests
2 Completed / **0** Remaining

Information Retrieval, Extraction & Machine Learning

Start: **Jul 5, 2017** Change Due: **Jul 10, 2017** Change

Edit Milestone | Information Ret... | Labels | Hide Pull Requests | Burn Pipelines



7 Total Story Points
7 Completed / **0** Remaining

3 Total Issues and Pull Requests
3 Completed / **0** Remaining

d. Concerns/Issues

NA

e. 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.