

# Dynamic Intelligent Q/A Systems

**CS5560 Knowledge Discovery and Management**

**Project 1**  
**Summer 2017**

**Team 7: TechGeeks**



## **Team members:**

- *Sai Jyothi Gudibandi*
- *Kalyan Kilaru*
- *Chaitanya Kumar Peravalli*

# **Project-1**

## **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.itι.gr/files/WikiRef\\_dataset.zip](http://mklab.itι.gr/files/WikiRef_dataset.zip)

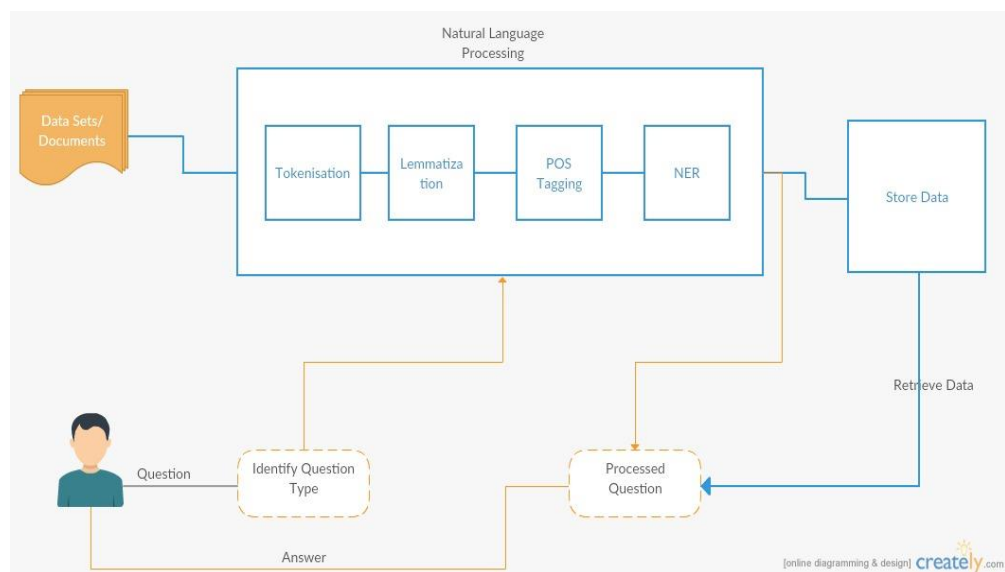
In BBCSports, we have so many interesting topics on sports like athletics, Cricket, Football, Rugby and Tennis. Here, we mainly focused on the Cricket topic because that topic has many interesting things to know.

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.

## 7. Design

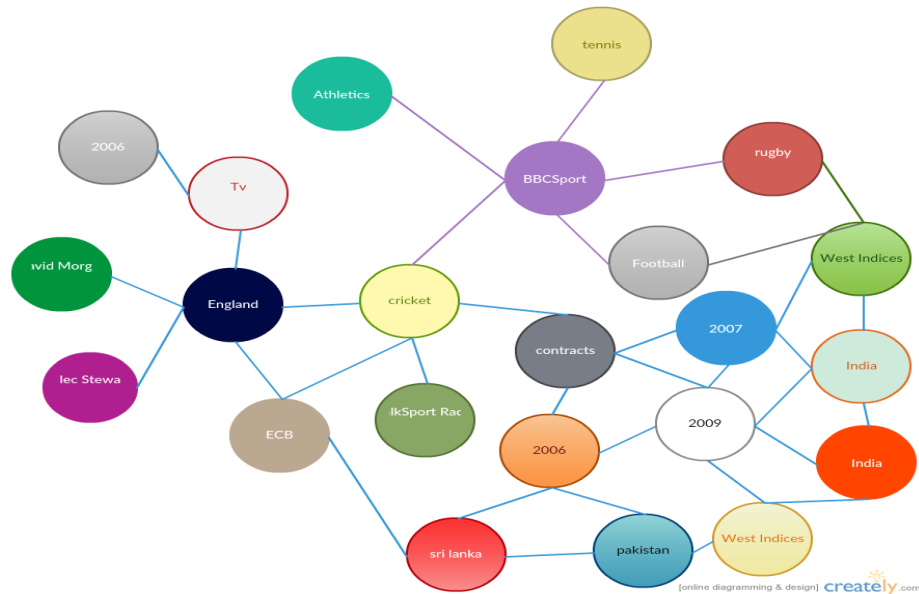
---

### a. Workflow for NLP

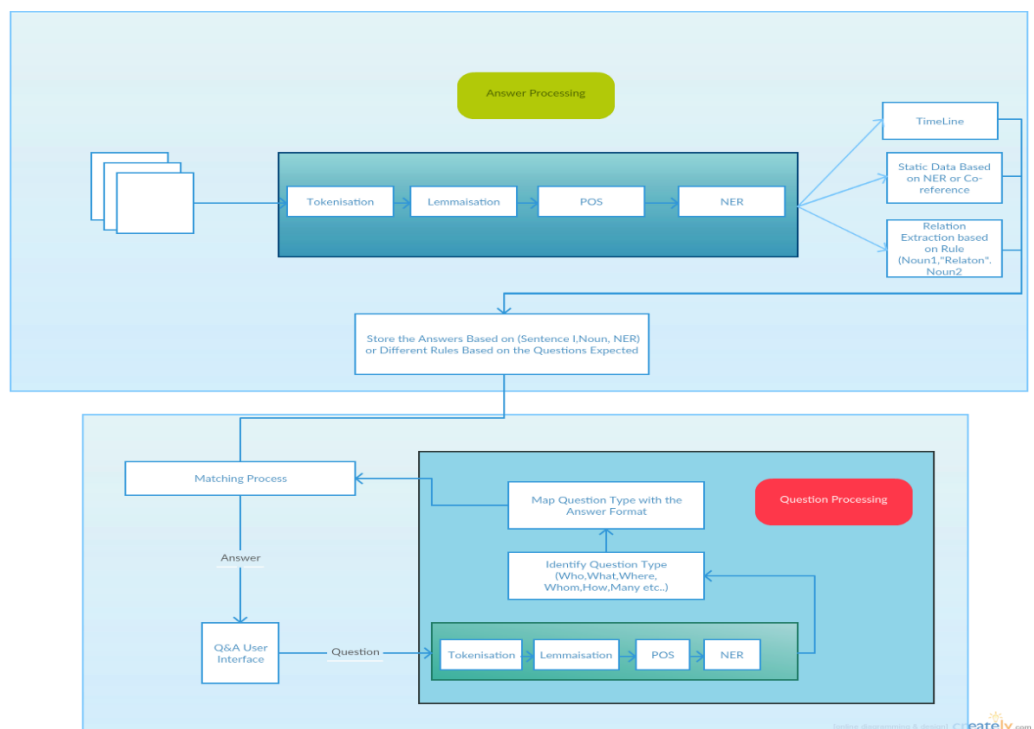


## b. Knowledge Graph

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



## c. Question and Answers

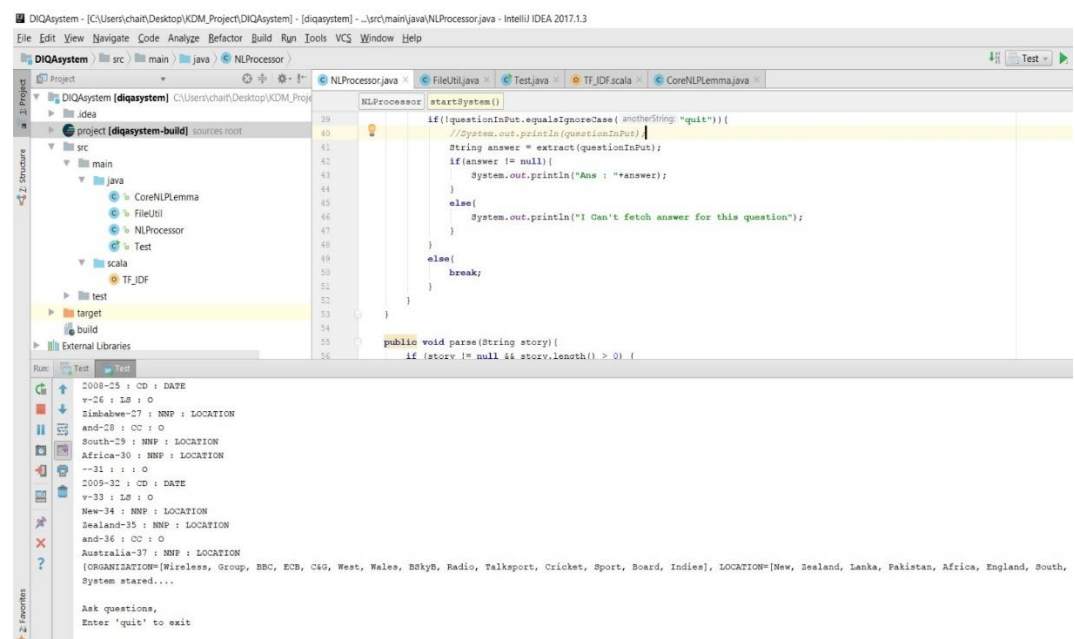


## 8. Implementation

### a. Using NLP

**Natural Language Processing:** 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.

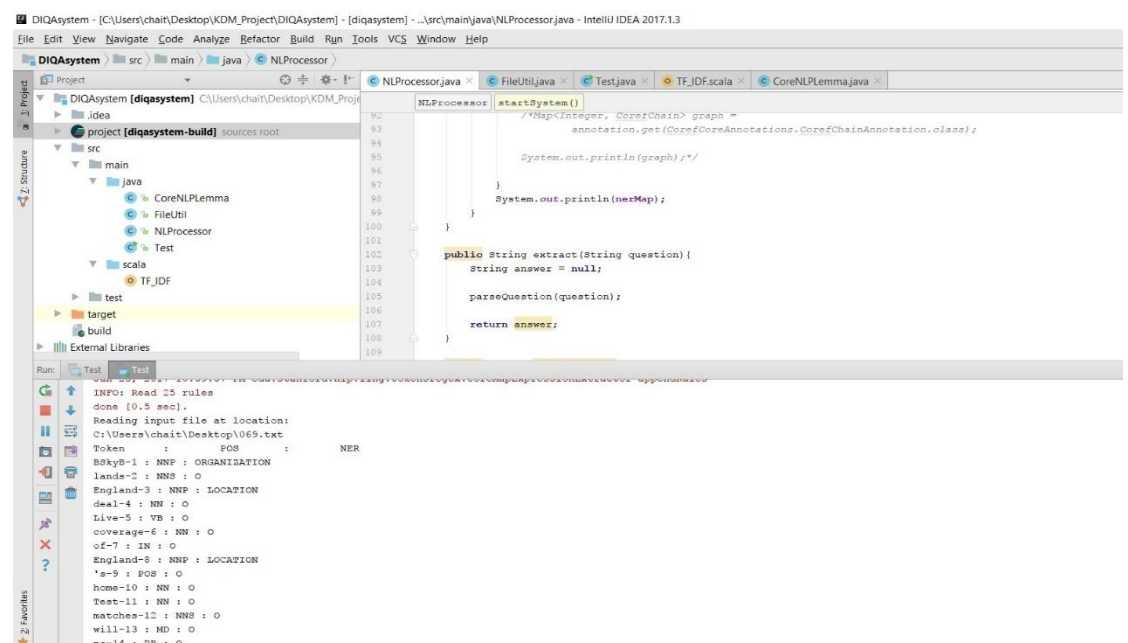
*Here are the screen Shots of the program and execution of the NLP for the chosen dataset:*



The screenshot shows the IntelliJ IDEA interface with the `NLProcessor.java` file open. The code defines a `startSystem()` method that interacts with the user and a `parse()` method that processes the input. The Run window at the bottom shows the execution output, which includes a list of entities and their types extracted from a dataset.

```
2008-25 : CD : DATE
v-26 : LB : O
Zimbabwe-27 : NNP : LOCATION
and-28 : CC : O
South-29 : NNP : LOCATION
Africa-30 : NNP : LOCATION
--31 : : : O
2009-32 : CD : DATE
v-33 : LB : O
New-34 : NNP : LOCATION
Zealand-35 : NNP : LOCATION
and-36 : CC : O
Australia-37 : NNP : LOCATION
(ORGANISATION=[wireless, Group, BBC, ECB, C&G, West, Wales, B&K&B, Radio, Talksport, Cricket, Sport, Board, Indies], LOCATION=[New, Zealand, Lanka, Pakistan, Africa, England, South,
System started....

Ask questions,
Enter 'quit' to exit
```



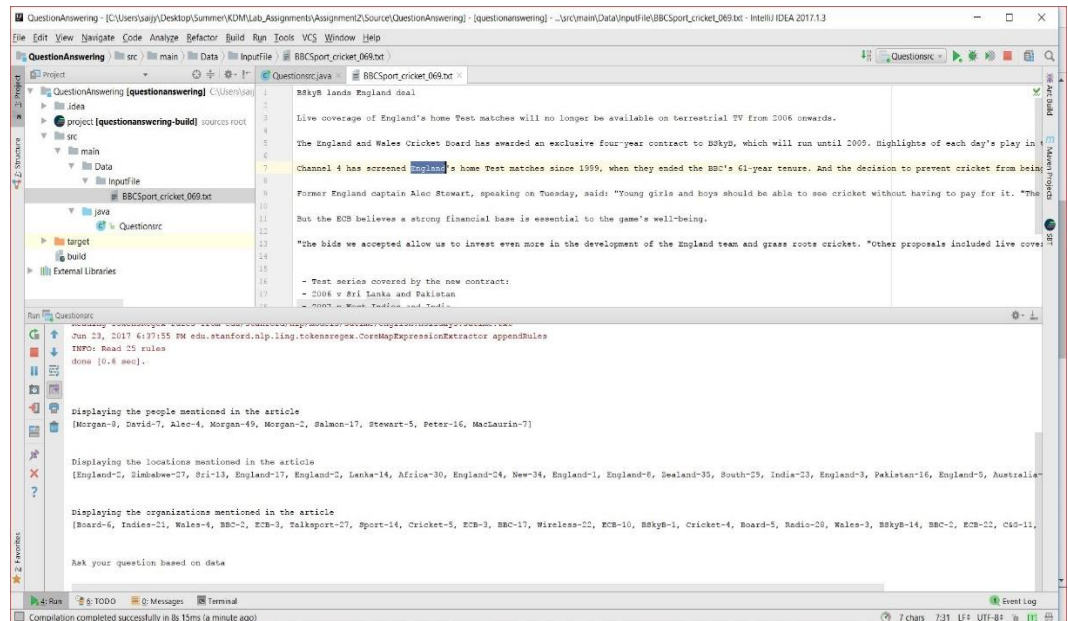
This screenshot shows the `NLProcessor.java` file with the `extract()` method implemented. The Run window displays the output of the Named Entity Recognition (NER) process, showing a list of tokens and their corresponding entity types.

```
INFO: Read 25 rules
done [0.5 sec].
Reading input file at location:
C:\Users\chait\Desktop\069.txt
Token : POS : NER
BskyB-1 : NNP : ORGANISATION
lands-2 : NNS : O
England-3 : NNP : LOCATION
deal-4 : NN : O
Live-5 : VB : O
coverage-6 : NN : O
of-7 : IN : O
England-8 : NNP : LOCATION
'a-9 : POS : O
home-10 : NN : O
Test-11 : NN : O
matchee-12 : NNS : O
will-13 : MD : O
no-14 : RB : O
```

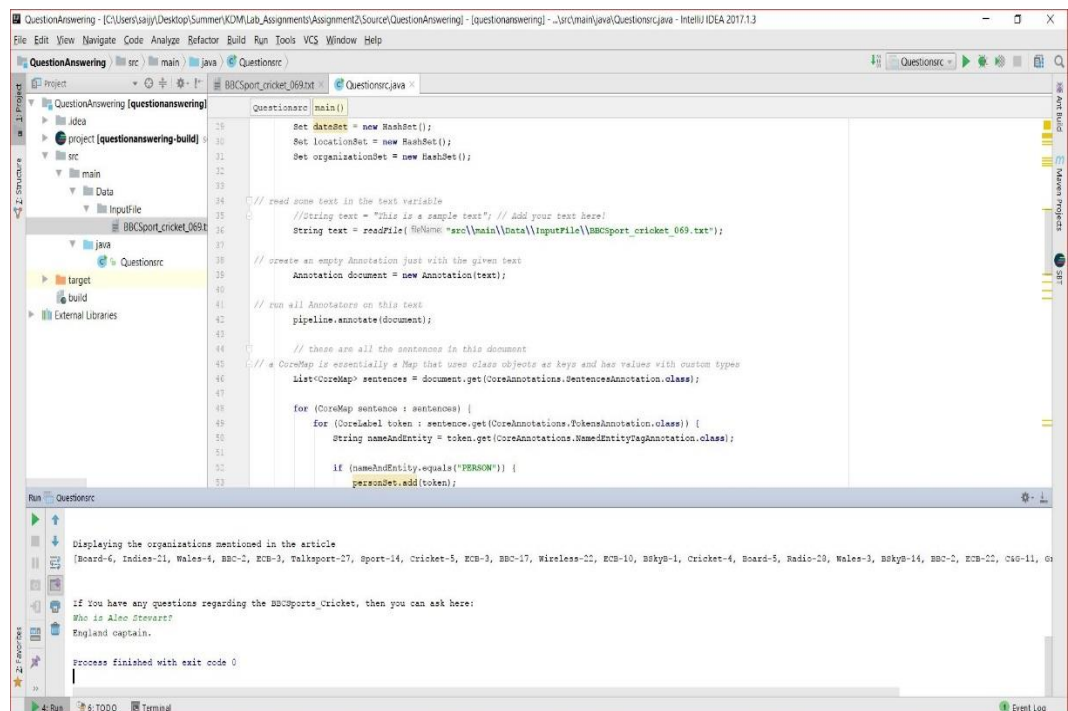
## b. Question Answer System

**Q/A System:** Here by using CoreNLP we will process the user given question and the Q/A system automatically produce the answer for that question.

*Here are the screen Shots of the program and execution of the Question Answer system for the chosen dataset:*

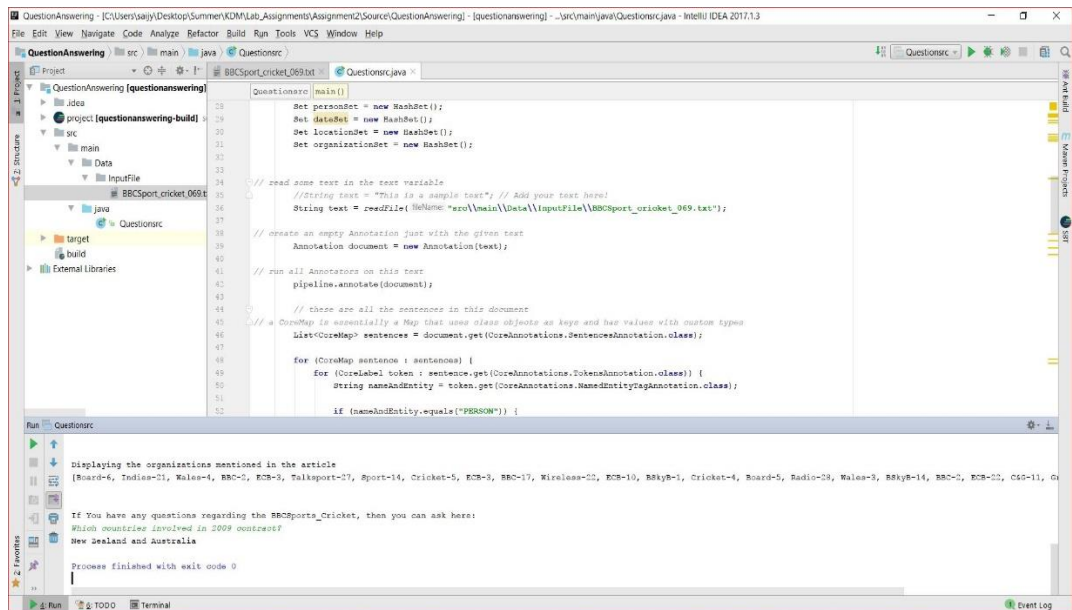


```
QuestionAnswering - [C:\Users\sai\Desktop\Summer\KDM\Lab_Assignments\Assignment2\Source\QuestionAnswering - (questionanswering) - ...src\main\Data\inputFile\BBCSport_cricket_069.txt - IntelliJ IDEA 2017.1.3]
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
QuestionAnswering [src | main | Data | InputFile | BBCSport_cricket_069.txt]
Project [questionanswering] [C:\Users\sai\... | QuestionAnswering.java | BBCSport_cricket_069.txt]
Run [QuestionAnswering] [C:\Users\sai\... | QuestionAnswering.java | BBCSport_cricket_069.txt]
INFO: Read 25 rules
done [0.8 sec].
Displaying the people mentioned in the article
[Morgan-8, David-7, Alec-4, Morgan-49, Morgan-2, Salmon-17, Stewart-5, Peter-16, MacLaurin-7]
Displaying the locations mentioned in the article
[England-2, Zimbabwe-27, Sri-13, England-17, England-2, Lanka-14, Africa-30, England-24, New-34, England-1, England-6, Zealand-35, South-25, India-23, England-3, Pakistan-16, England-5, Australia-11]
Displaying the organizations mentioned in the article
[Board-6, Indies-21, Wales-4, BBC-2, ECB-3, Talksport-27, Sport-14, Cricket-5, ECB-3, BBC-17, Wireless-22, ECB-10, BSkyB-1, Cricket-4, Board-5, Radio-20, Wales-3, BSkyB-14, BBC-2, ECB-22, C4G-11]
Ask your question based on data
Compilation completed successfully in 8s 15ms (a minute ago)
```



```
QuestionAnswering - [C:\Users\sai\Desktop\Summer\KDM\Lab_Assignments\Assignment2\Source\QuestionAnswering - (questionanswering) - ...src\main\java\Questioner.java - IntelliJ IDEA 2017.1.3]
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
QuestionAnswering [src | main | java | Questioner]
Project [questionanswering] [C:\Users\sai\... | Questioner.java | BBCSport_cricket_069.txt]
Run [Questioner] [C:\Users\sai\... | Questioner.java | BBCSport_cricket_069.txt]
Displaying the organizations mentioned in the article
[Board-6, Indies-21, Wales-4, BBC-2, ECB-3, Talksport-27, Sport-14, Cricket-5, ECB-3, BBC-17, Wireless-22, ECB-10, BSkyB-1, Cricket-4, Board-5, Radio-20, Wales-3, BSkyB-14, BBC-2, ECB-22, C4G-11, G]
If you have any questions regarding the BBCSports_Cricket, then you can ask here:
Who is Alec Stewart?
England captain.
Process finished with exit code 0
```



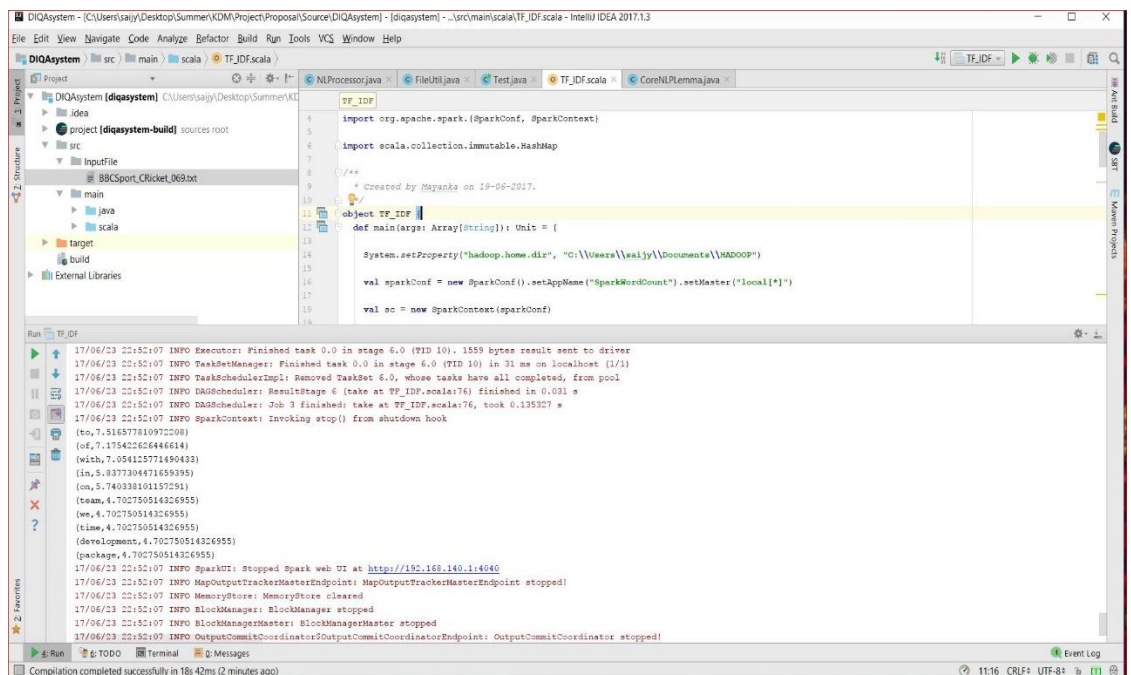


## c. TF-IDF

**TF-IDF:** Term Frequency-Inverse Document Frequency is used for information extraction. By this we can get know the importance of the words to that particular document.

In this program we will get the output word and including with the TF-IDF of that particular word.

*Here are the screen Shots of the program and execution of the TF-IDF for the chosen dataset:*





## 9. Project Management

---

### a. Work Completed

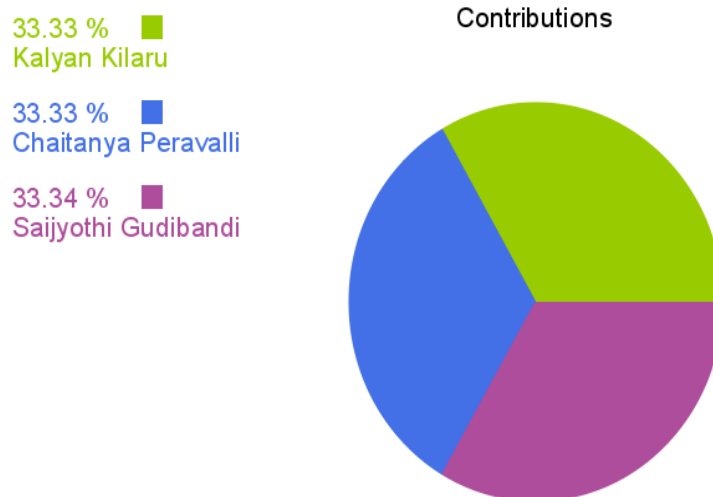
Design and Architecture of the application.

### b. Contributions

*SaiJyothi Gudibandi*

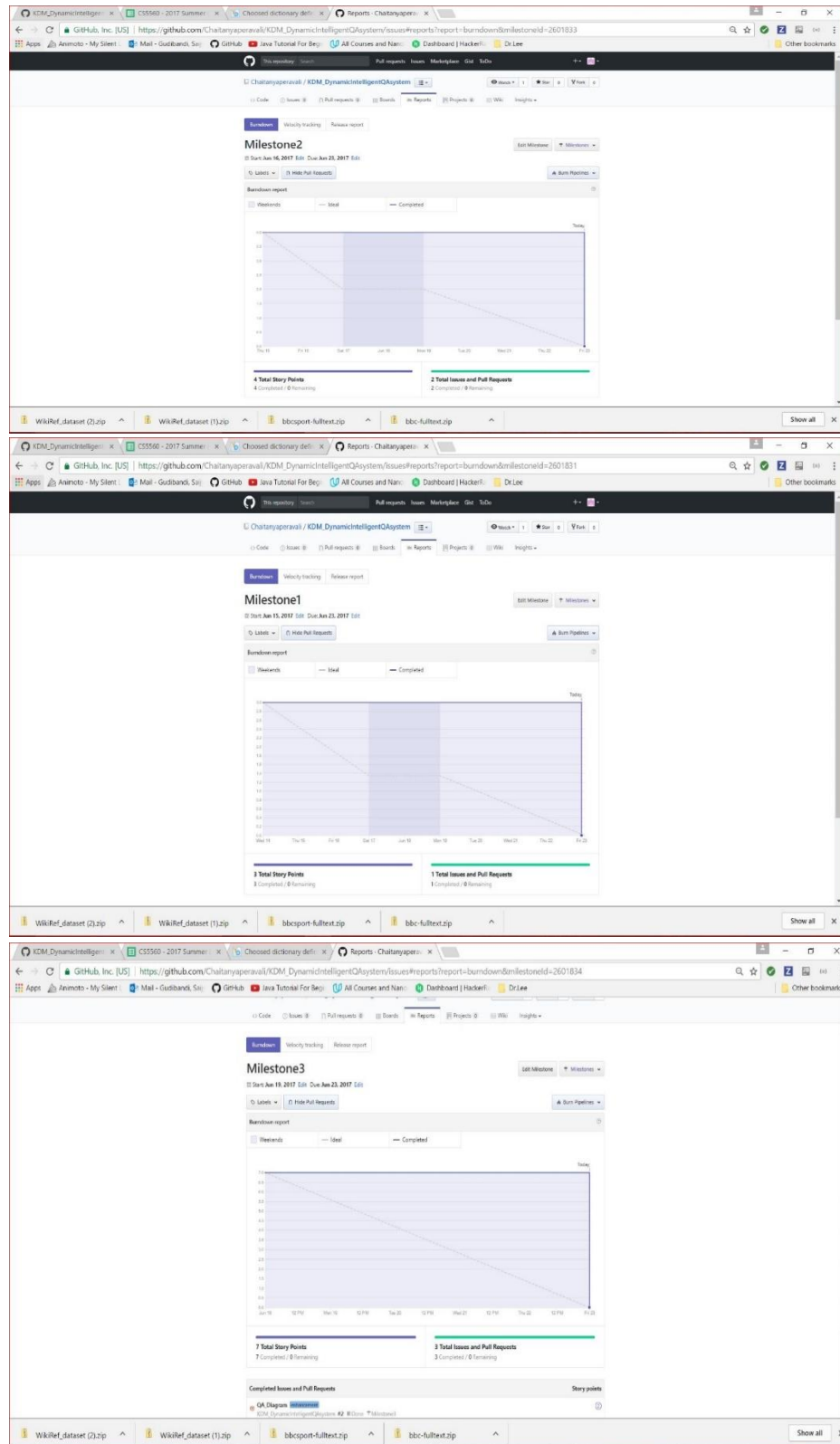
*Kalyan Kilaru*

*Chaitanya Kumar Peravalli*



## c. Statics

*Here are the screen Shots Project Management:*



#### **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.