

**CS5560 Knowledge Discovery and Management**

**Dynamic Intelligent Q/A System**

**PROJECT 3 REPORT**

**SUMMER 2017**

**Team – 3  
 Nageswara, Rao Nandigam – 18  
 Chakilam, Revanth – 2**

**Syed, Moin – 28**

**Sankar, Pentyala – 22**

Motivation:

We the Team Innovators 2.0 are in a search of data and knowledge. But Importantly we have a lot of difference between information, data, statistics and knowledge. When we speak about Information intruding and a semantic google search is more improvised and we can retrieve relevant information sitting in home. “Question Answering” is a dynamic way of retrieving Information, which learns knowledge. The main focus is particular in obtaining the respective documents but also particular in obtaining the respective response to query we post. Question and Answer system is capable of performing the NLP Processing, Information gathering, topic discovery, Machine Learning and Sematic Search. Question and answer system itself is the beauty of NLP and same instance have a bit of science in its essence. Question and Answer System is required every aspect, let be in the field of health and sciences, an intelligent learned system for children at schools, professional assistant etc. So, this is necessary in each case when we require some help from computer as well. It goes without saying that it is worth exploring the exciting field of question answering.

Objective:

Our Project critically deals with the building of typical model of information knowledge retrieval, named Question & Answering model. If suppose any given query asked in natural human language, Our question and answer system is implemented in such a way to extract the reality possible answer in the form of a pre-defined named-entity type, that is a human person, or may be an organization, a location, etc. Thus, we are connecting the question objects with live entities in a given radius is important in question and answer system. The projects main motto is to enhance the performance of the Question Answering system by using the knowledge and data from the natural grouping of word in the document files.

Significance:

We are constructing a knowledge graph such a way to build the question and answering system to deliver answers very effectively. To give better the results from the system designed we are applying different techniques such as NLP operations, Information retrieval, topic discovery and knowledge discovery.

Q/A System:

For this project, we have taken the Data set from BBC sports concentrating on the sport Cricket. From this Data set we try to construct knowledge graph and making system dynamic to answer all possible questions on sports questions.

Related Work:

*Knowledge graph refinement:*

Firstly, Knowledge graph is the one of the advanced and trending concepts today. By making use of it, we can build dynamic systems which can precisely answer our questions contrary to the existing search engines which gives us information and a set of related links which makes the user to spend more time on the web for the obtaining the information one was looking for.

Almost as every technology with the usage it needs refinement the same applies to the knowledge graph too. In this paper, the technologies used to evaluate knowledge graph are Partial gold standard, Silver standard, Retrospective evaluation. Each of this technology is a trade-off between reliability and cost. Completeness aims at increasing the coverage of the knowledge graph & various models are used to evaluate completeness and correctness which cannot be both achieved at the same time. This paper is based on the survey results.

*Knowledge vault:*

The knowledge Vault is a database of what google would call the facts that has been scraped from across the entire web. Unlike the traditional ranking system which relies on the incoming links and number of those links to determine the quality of a source, the knowledge vault allows google to develop this system whereby they count the number of true or false facts, where a true or false is determined by how often that appears on the rest of the web. So, its kind of like querying our collective consciousness to ask us what’s true or not.

Controversial facts are going to trip up algorithms like this and create even more controversy about what should be surfaced at the top and what should be lowered down depending on sort of objective opinions. The knowledge vault is mainly of extracting triplets and giving the score to them and it has advantages of to automatically crawl, index and organize information across the web.

*Knowledge base Construction:*

We are at that point of life where Data is growing day by day. There is a lot of unstructured data in various organizations and medical fields. A need for dark data extraction & KBC - Knowledge base construction is increasing every moment. So Deep Dive is a solution to establish a SQL database with data from various unknown sources like images, emails, static web pages etc. In every Q/A system its been a long standing problem and Deep dive provides a better solution to it. Currently in this project, Deep dive gives a choice to domain experts to design each ones own KBC systems. Gibbs sampling is used in the processing of data which has high accuracy. The output is the probability of the words which are needed to be present in the data set.

*Semantic data integration for knowledge graph construction:*

The introduction of Web documents in to a Webservices and data has definitely reulted in the increment availabity of data from each type of domain networks. A new Semantic-Data-Integration approach called "FUHSEN" has been invented which this system can exploit use the key words and structured strengths of webdata sources and can generate quality know;edge graphs by amalgaming the data collected from various data sources.This system FUHSEN depends on two things. i) RESOURCE DESCRIPTION FRAMEWORK (RDF) : for semantically describing the collection of entities

ii) Semantic similarity measures among the collected entities and establishing the relation between them.

The results of the Fuhsen integration system arfe evaluated from DBpedia knowledge base. So in the current project, using the Fuhsen data intergration technique, we can accurately integrate the various similar data entities semantically and transform them in to quality knowledge graphs.

*Knowledge Base Construction from Richly Formatted Data:*

In this framework. the entities , relation between them and the attributes are related via tabular, textual, structured and visualized expression. This FONDUER intriduces a uniuqe model for KNOWLEDGE BASE CONSTRUCTION built on a unified data representation. This is a new KBC system for the RFIE - richly formatted information data extraction and it also use human loop algorithm in order to train machine learning systems. In our project, we can use FONDUER to ease the burden of traditional approaches. This model in addition with data programming allows the end users to supervise and help to implement Knowledge Base Construction process over the ricjly formatted data.

*Clustering Approach :*

The blooming of World WIde Web 2.0 Community Question Answering such as Yahoo! Answers , WikiAnswer etc., have emerged as other approches for knowledge and information extraction and acquisition. Over a period of time, a huge number of question and answers with good quality has been devoted by human intelligence have been acheived as a knowledge base. Besides the search engines, which return long queries, finding in the CQA services can achieve proper answers to the questions by automatically finding similar questions that has been answered by different users. It improves the efficiency of information retrieval. However, if given a question, finding the similar answered query is a difficult task. The main task is the word mismatch between question query and candidate query for retrieval query.

*Knowledge Base Completion:*

In recent years, huge amounts of world knowledge have been in publicly available knowledge bases, such as Freebase and YAGO. Having huge databse they look incomplte.Here we propose a way to leverage old Web based search question answering technology to fulfil the gaps in knowledge bases. In general, for every entity attribute learn the best set of queries to ask, from that the answer snippets returned will contain the correct value for the attribute. The system learns and add disambiguating terms, in order to make it more likely that the search results contain correct mentioning. The system will also learns how many queries to ask for each attribute, since in many situations, asking many can reduce accuracy, predictions for possible values for each attribute.Lastly we evaluate our system and show that it is able to extract a large number of facts with high accuracy.

*Predictive Analysis* :

An expert in domain can analyze huge data to make meaningful predictions from the data. For example, by scanning through the research papers and patent records, a domain expert can predict the geographic location and time , where and when the technology will become useful. However, this is an tedious task. This paper presents an system that acquaints the heterogeneous data into a knowledge graph in the RDF (Resource Description Framework) format using ontology method. Hence, the user can easily query the knowledge graph to get the required data from the predictive analysis tools. The system extracts the detailed data from public sources including research papers and patent records. Next, the system an ontology-based data deiled method to generate knowledge in the RDF format to enable quickly between machine learning for predictive analytic tasks.

*ConceptNet 5.5:*

Machine learning language can be improved by providing with knowledge and resources of external information. ConceptNet is particularly suited to be used with modern NLP techniques such as word embeddings. It is a knowledge graph that mainly connects words and phrases of natural language with labeled edges. Its knowledge is collected from many sources that include expert created resources and games with a purpose. It is designed to represent the basic knowledge involved in processing the natural languag by improving its applications by allowing the application to understand the synonyms behind the words that we use. When this approach is combined with word embeddings which is acquired from semantics such as word2vec, it provides application that they would not require from semantics alone, nor from the resources such as DBPedia and WordNet.

Datasets:

*BBC Sports domain:*

We have taken BBC sports dataset as one of our dataset for the design of Question and Answering system. This dataset consists of all kind of sports which includes athletic, cricket, football ,rugby and tennis.

We have mainly focused on the cricket dataset for designing and extracting information using Natural language processing, word2vec, N-gram and TF-IDf techniques

<http://mlg.ucd.ie/datasets/bbc.html>

*BBC News:*

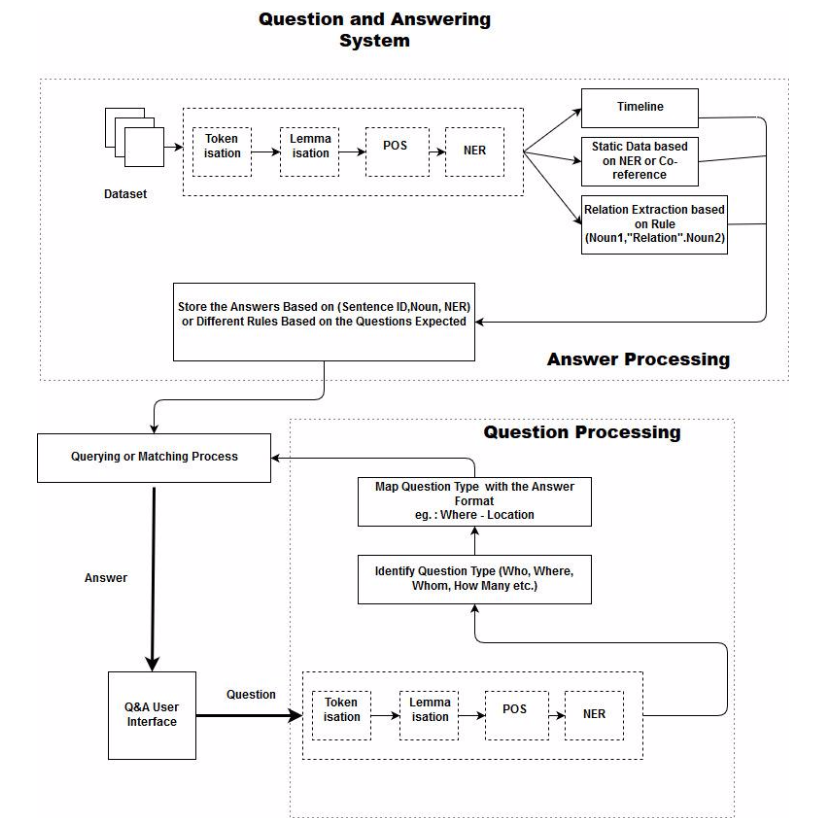
We have taken BBC news dataset as one of second dataset for the design of Question and Answering system. This dataset consists of all kind of news which includes business, entertainment, politics, sports and technology.

We have mainly focused ont the sports dataset for designing and extracting information using Natural language processing, word2vec, N-gram and TF-IDf techniques.

<http://mlg.ucd.ie/datasets/bbc.html>

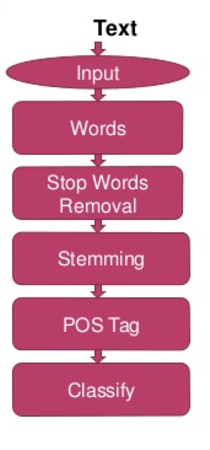
Design:

Workflow:



NLP:

Open NLP is a toolkit available in variety of programming languages which supports most of the Natural Language Processing tasks like tokenization, pos tagging, chunking, name recognition, sentence segmentation. These tasks are generally required to produce services that require more advanced text processing.

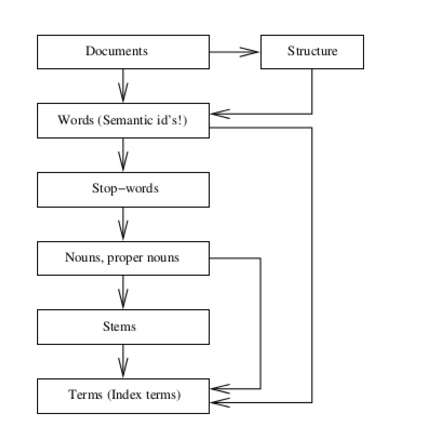


Information Retrieval:

Skip-gram negative sampling (or Word2Vec) is an algorithm based on a shallow neural network which aims to learn a word embedding. It is highly efficient, as it avoids dense matrix multiplication and does not require the full term co-occurrence matrix. Given some target word wt, the intermediate goal is to train the neural network to predict the words in the c-neighborhood of wt: wt − c, …, wt − 1, wt + 1, …, wt + c. First, the word is directly associated to its respective vector, which as used as input for a (multinomial) logistic regression to predict the words in the c-neighborhood. Then, the weights for the logistic regression are adjusted, as well as the vector itself (by back-propagation). The Word2Vec algorithm employs negative sampling: additional k noise words which do not appear in the c-neighborhood are introduced as possible outputs, for which the desired output is known to be false. Thus, the model does not reduce the weights to all other vocabulary words but only to those sampled k noise words. When these noise words appear in a similar context as wt, the model gets more and more fine-grained over the training epoch.

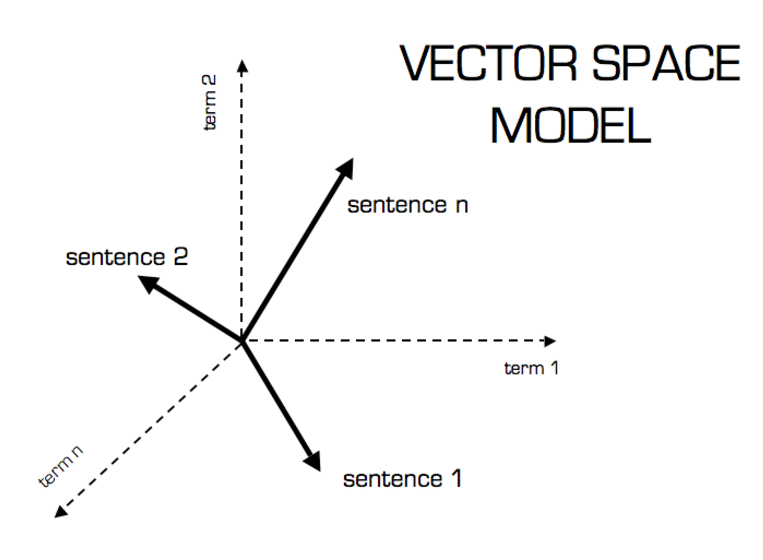
* TF – IDF

Term Frequency Inverse document frequency produces an output with the word and term weight. It gives how importance a word is to a document collection/corpus.



* W2V

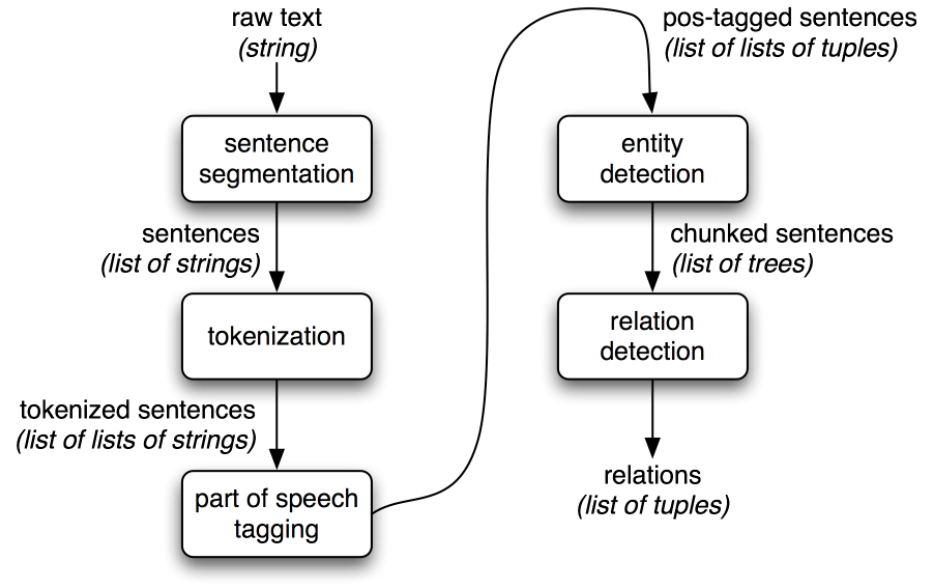
The input to this model is a text corpus. It is a 2-layer neural network that processes the text data. The output is a series of vectors which are feature vectors for the words in the input corpus.



Information Extraction:

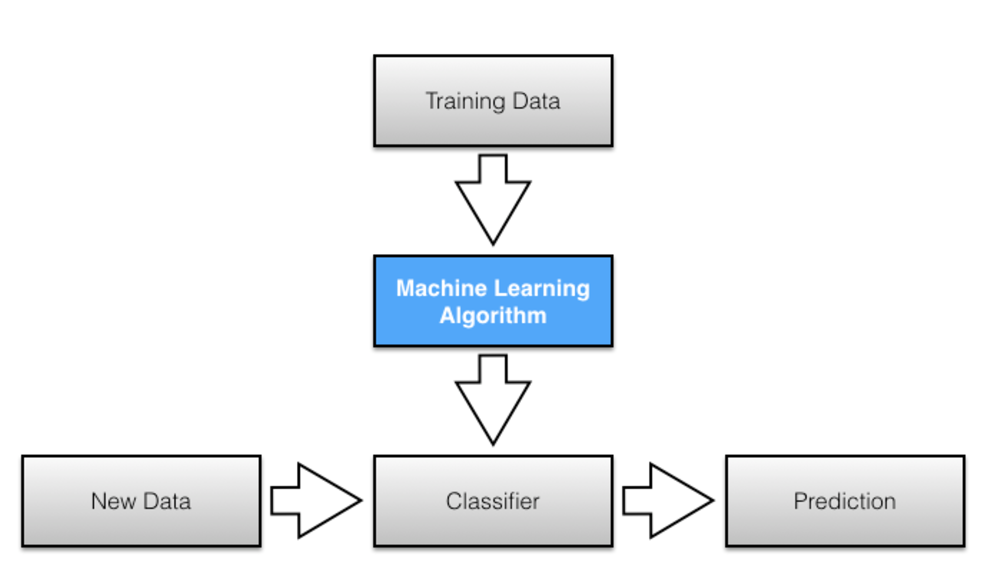
Word net is a software package developed at Princeton University providing a lexical data base of English words meaning and categorization. It provides the means to determine the categorization of a word which is used in answer ranking phase.

IE is a process of extracting a structured data from an unstructured data through a means of Natural Language Processing.

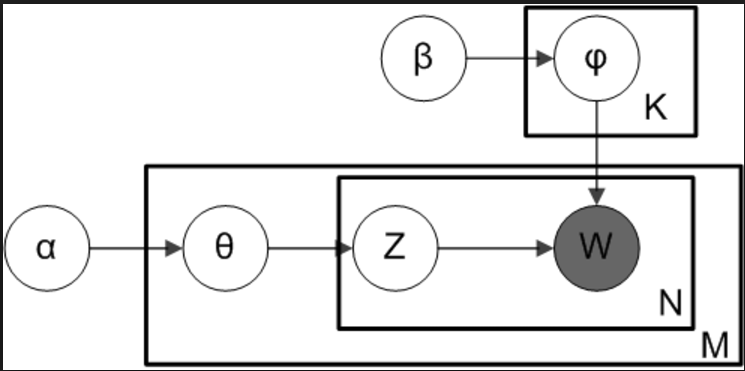


Machine Learning:

Classification is the result of supervised learning which means that there is a known label that you want the system to generate. For example, if you built a fruit classifier, it would say “this is an orange, this is an apple”, based on you showing it examples of apples and oranges.

Clustering is the result of unsupervised learning which means that you’ve seen lots of examples, but don’t have labels. In this case, the clustering might return with “fruits with soft skin and lots of dimples”, “fruits with shiny hard skin” and “elongated yellow fruits” based not simply showing lots of fruit to the system, but not identifying the names of different types of fruit.

LDA:



Knowledge Graph:

A knowledge graph is mainly organized as a graph, which is not always true of knowledge bases. The primary benefits of a graph are that relationships in the data are first-class citizens, you can easily connect new data items as they are injected into the data pool, and, finally, you can easily traverse links to discover how remote parts of a domain relate to each other there’s a huge value in linking information. A graph is one of the most flexible formal data structures, so you can easily map other data formats to graphs using generic tools and pipelines.

<Won>

<Scored>

<Has\_Name>

<Sets\_up>

<Has name>

<hits >

<Named>

<Has >

<Named>

<Named>

<Named>

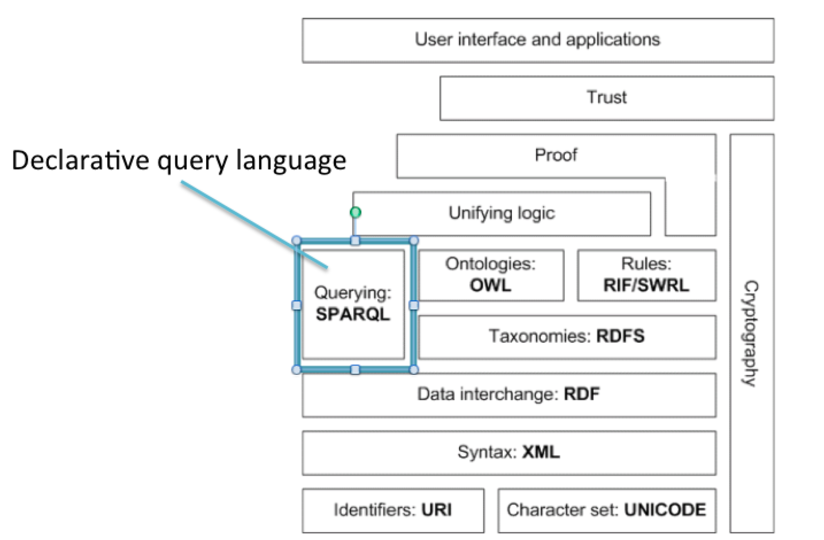
<took >

<Scored >

<Scored >

SPARQL:

It is the extension of spark query language for operating not only on Resource Descriptive Framework but also on the documents/corpus data on arbitrary format.



Questions and Answers:

1. Who all played between nez vs aus

Answer:

S P Fleming, N J Astle, M S Sinclair, J Wilson, C D McMillan, H J H Marshall, C L Cairns, B B McCullum, K D Mills, D L Vettori, DTuffey

M L Hayden, A C Gilchrist, R T Ponting, D R Martyn, A Symonds, M J Clarke, M E K Hussey, G B Hogg, B Lee, J Gillespie, G D McGrath

1. When nez vs Aus match happened

Answer:

March 1993

1. Where nez vs aus match was held

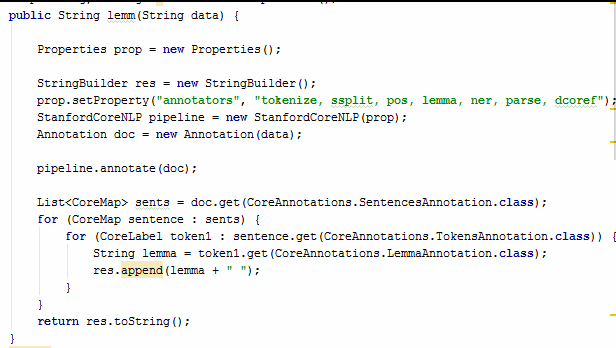
Answer:

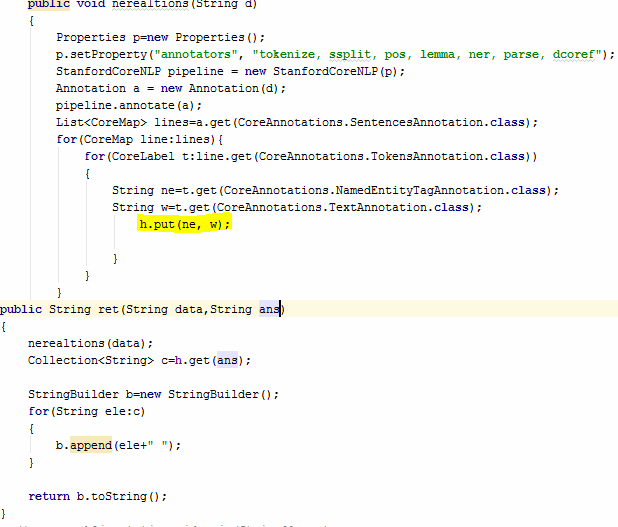
Jade Stadium

Implementation, Results and Evaluation

NLP Process:

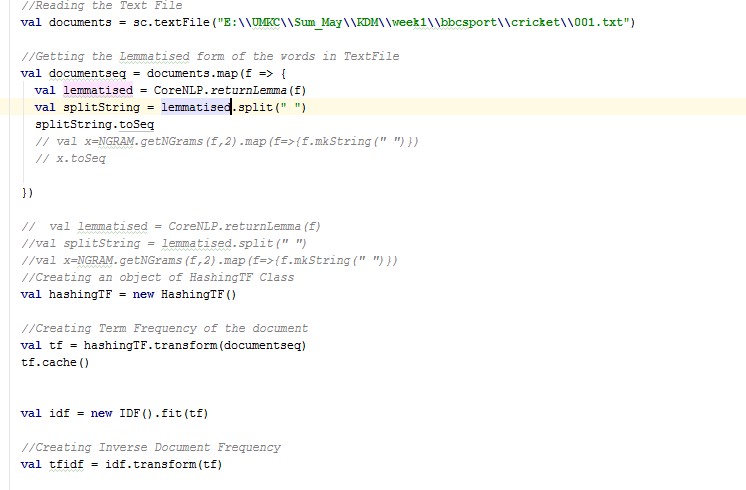
1. Doing NLP processing of tokenizing, lemmatization and extracting named entity relations and storing it in HashMap.

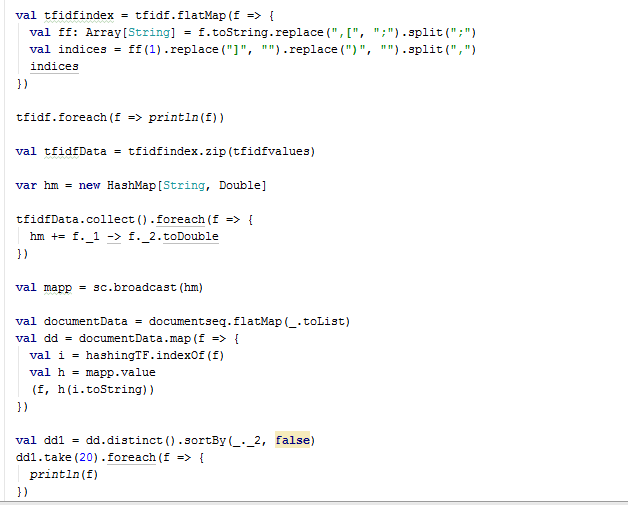




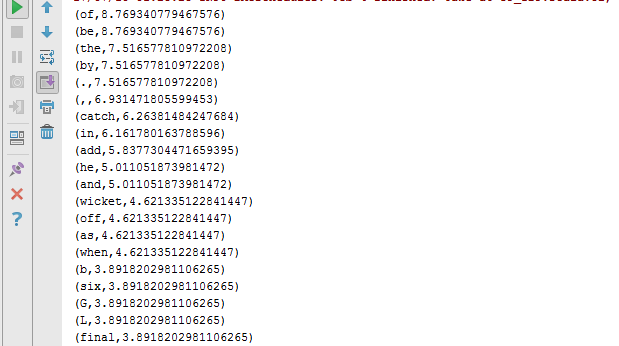
Information Retrieval:

TF-IDF:





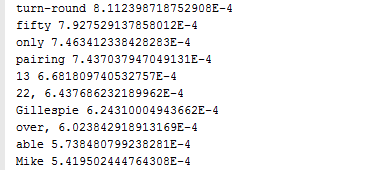
Output:



Word2Vec:

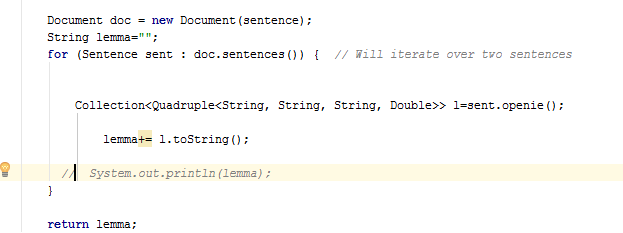


Output:

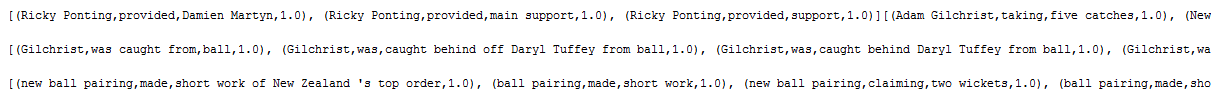


Information extraction:

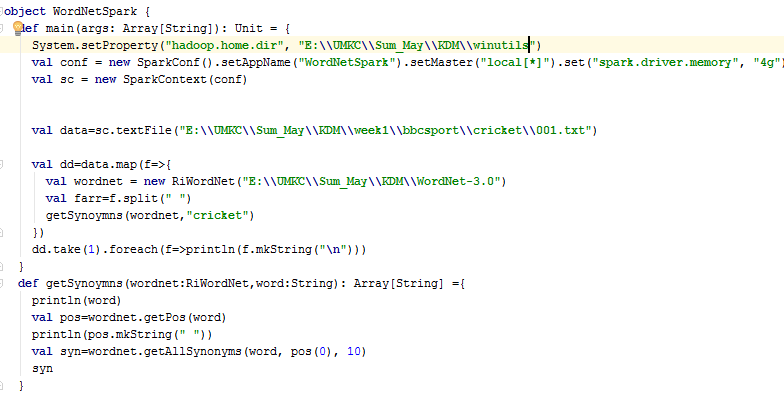
OpenIE:



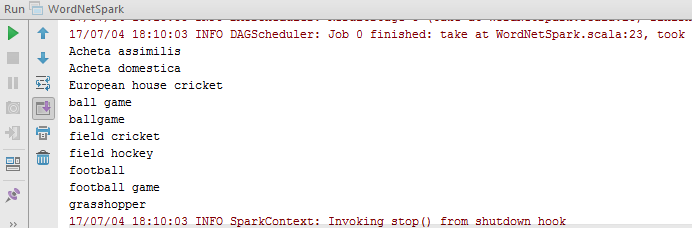
Output:



Wordnet:



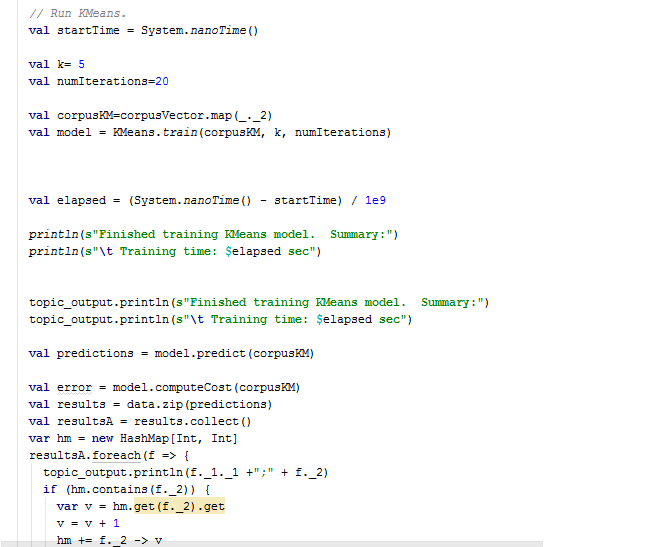
Output:



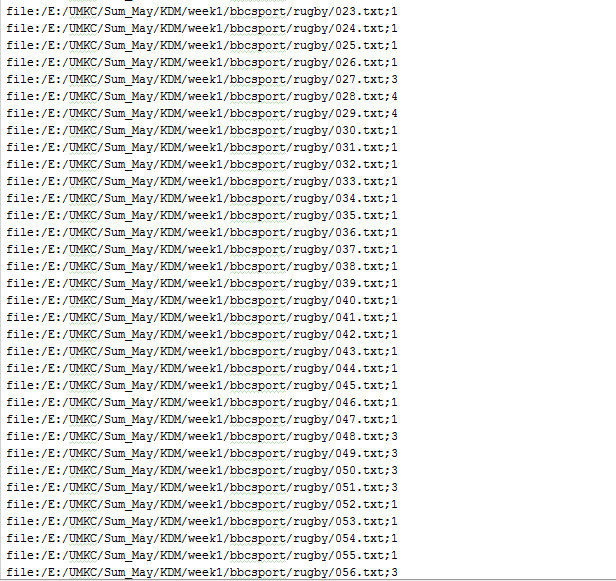
Machine Learning:

Clustering:



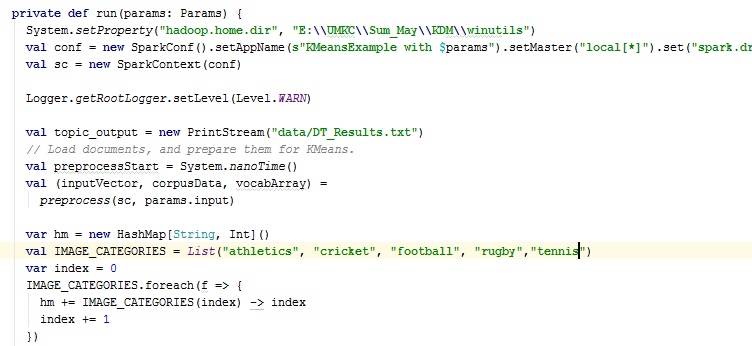


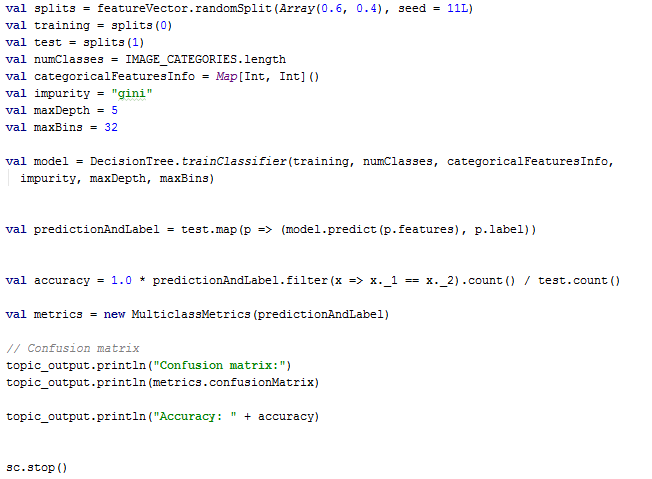
Output:



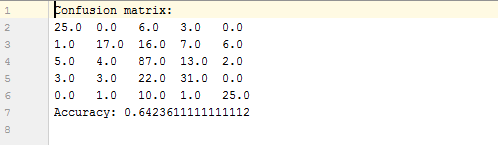
Classification:

Decision tree:

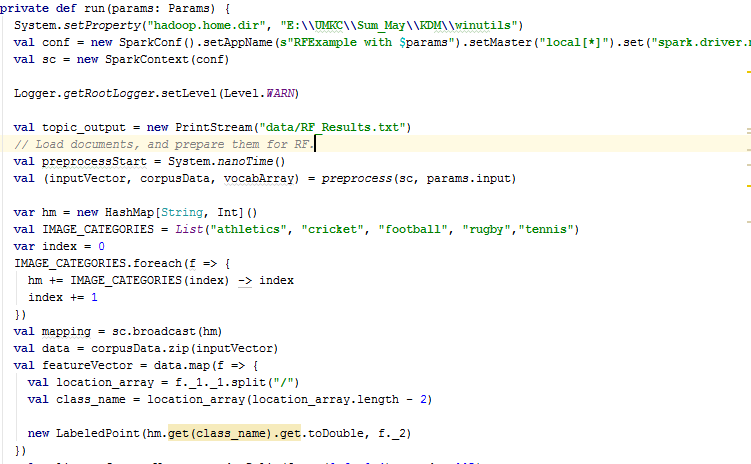


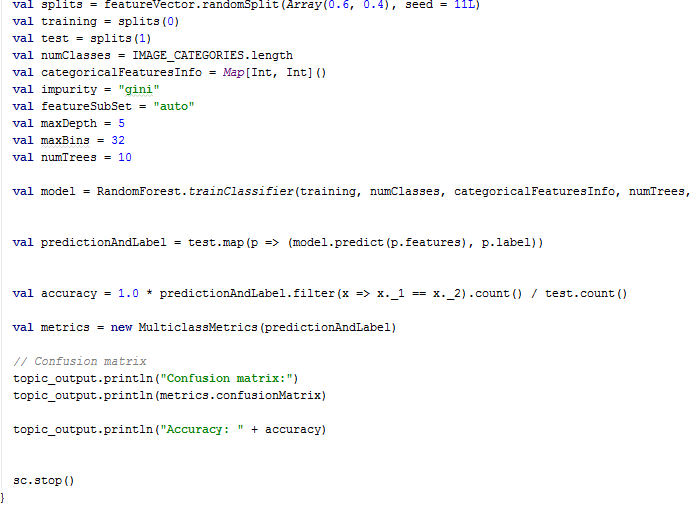


Output:

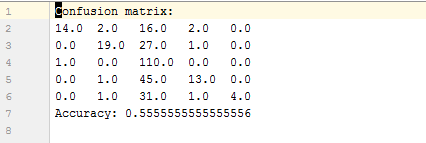


RandomForest:

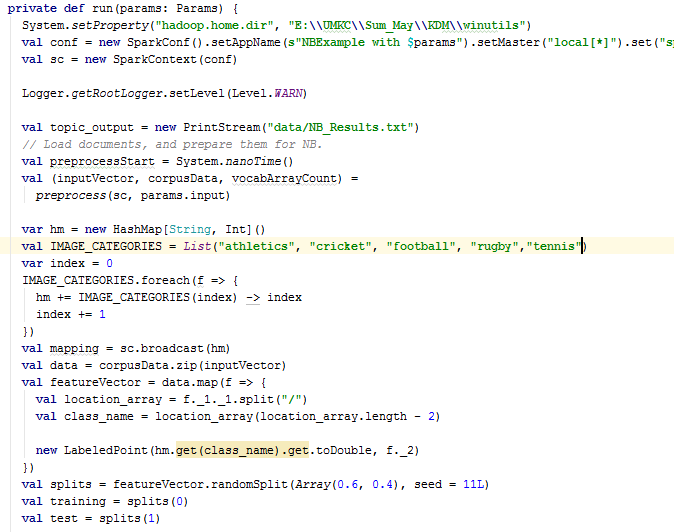


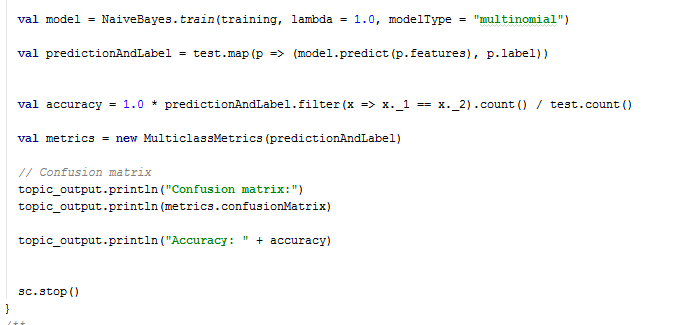


Output:

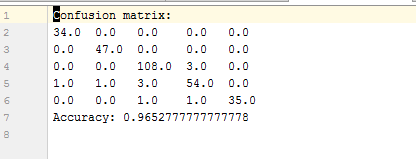


Naïve Bayes:

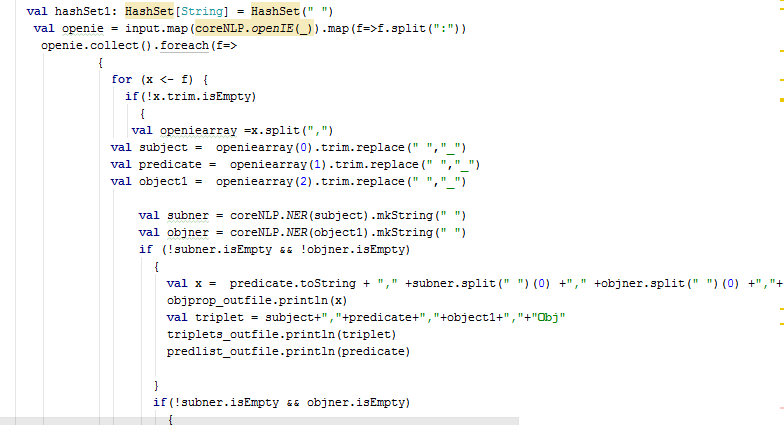


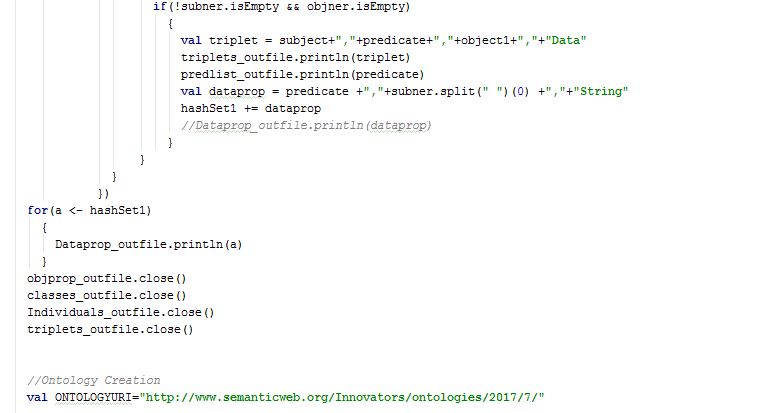


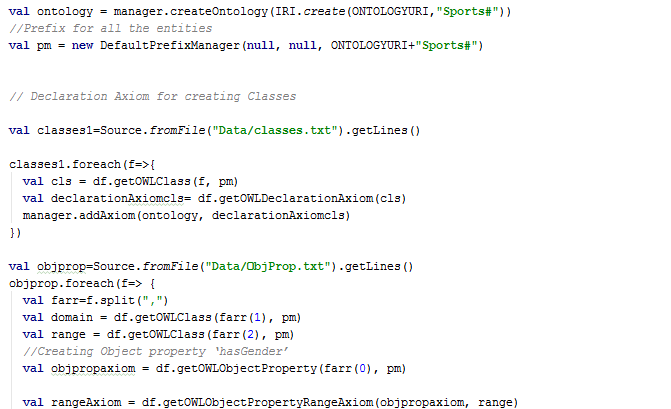
Output:

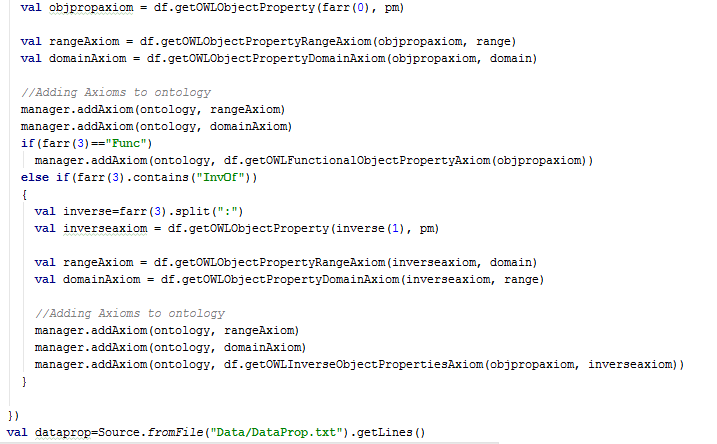


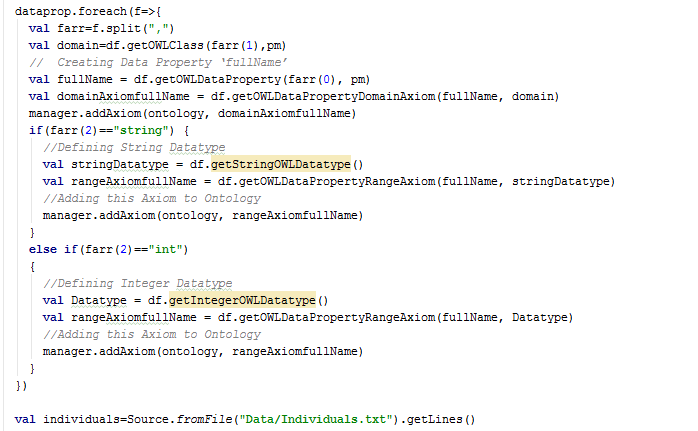
Knowledge Graph (KG):

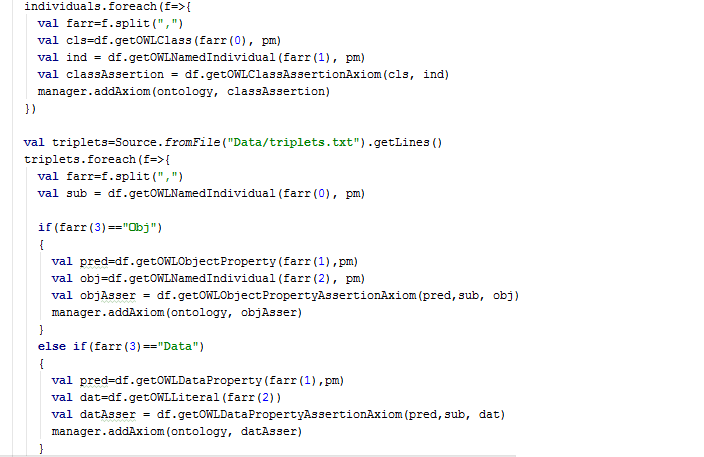


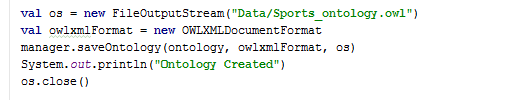












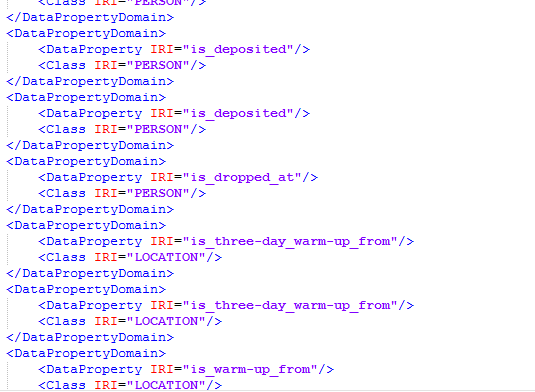
Output:







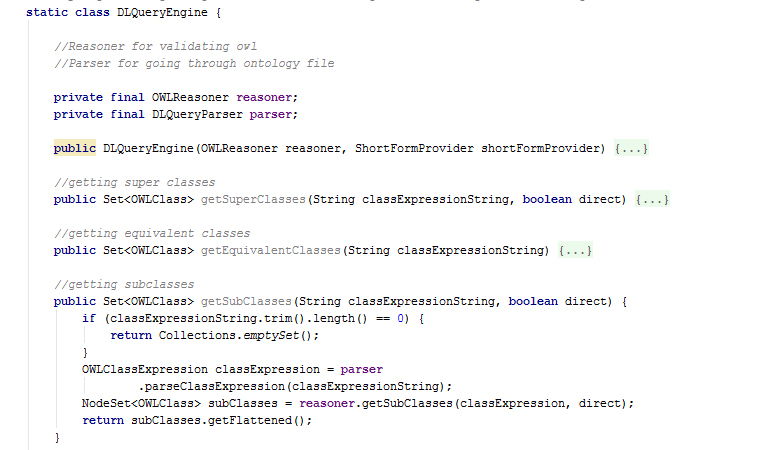


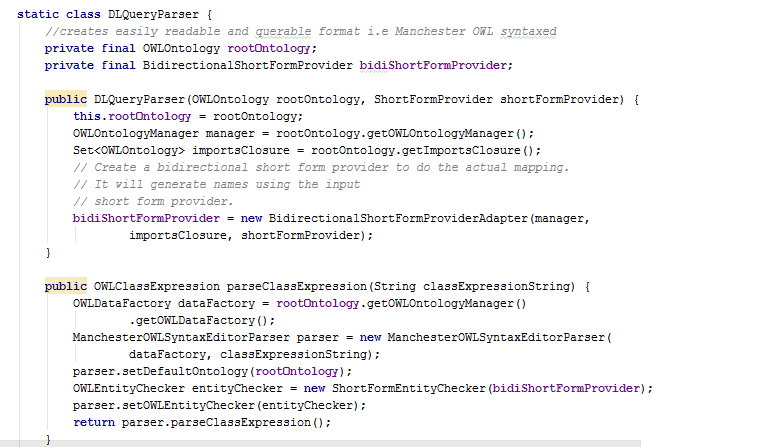


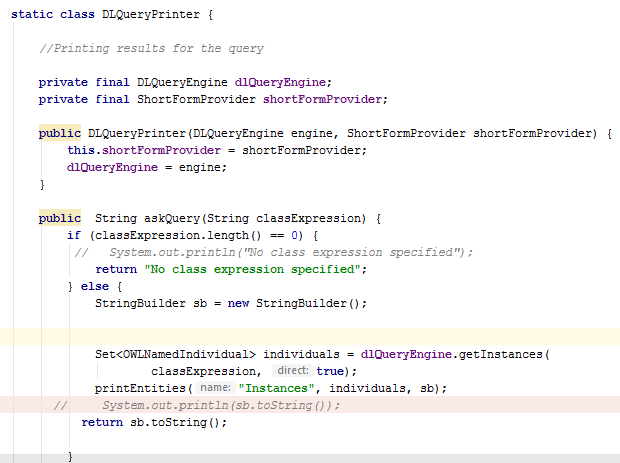


Generation of SPARQL/SWRL:

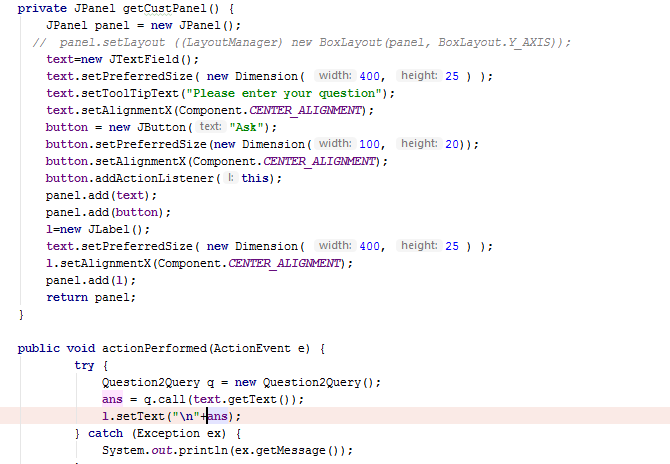




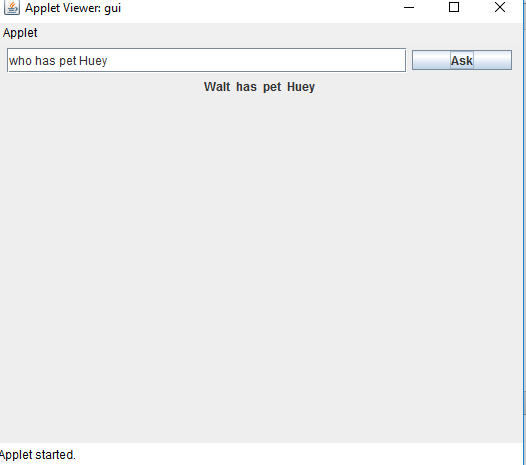




Question Answering using your KG and SPARQL/SWRL:



Output:

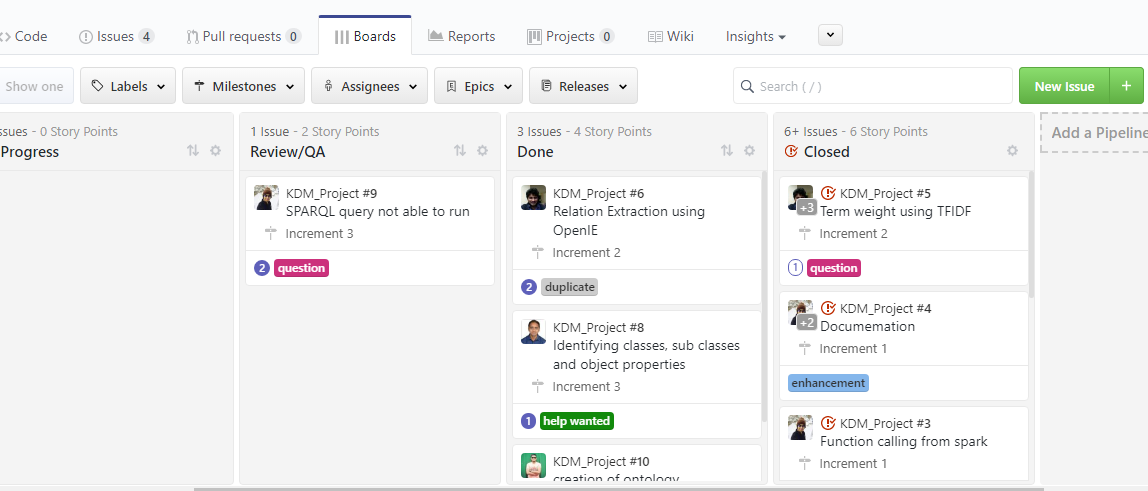


Project Management:

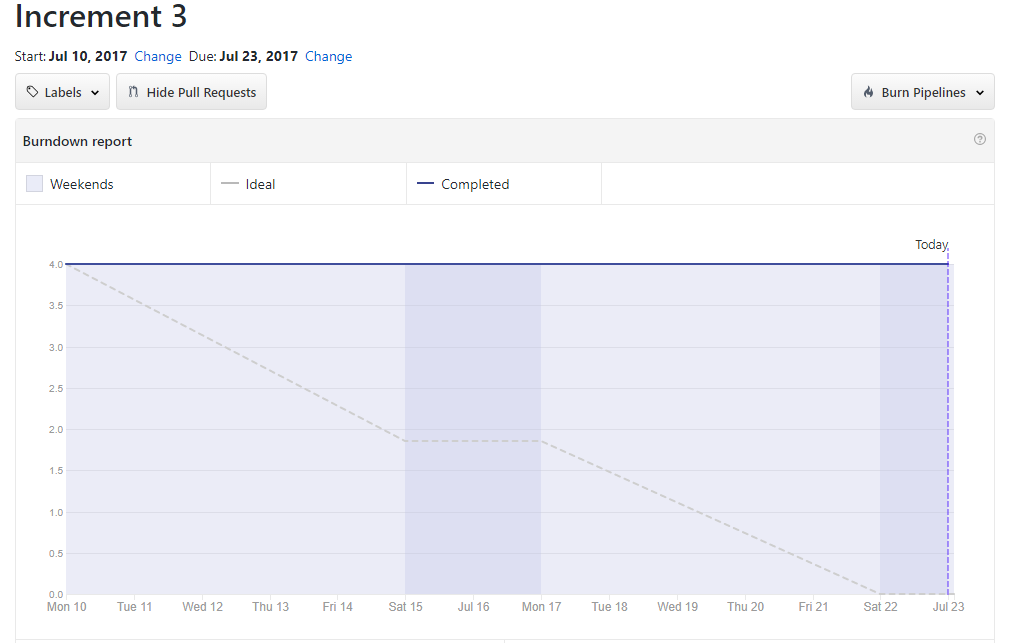
Contribution:

GitHub screens:

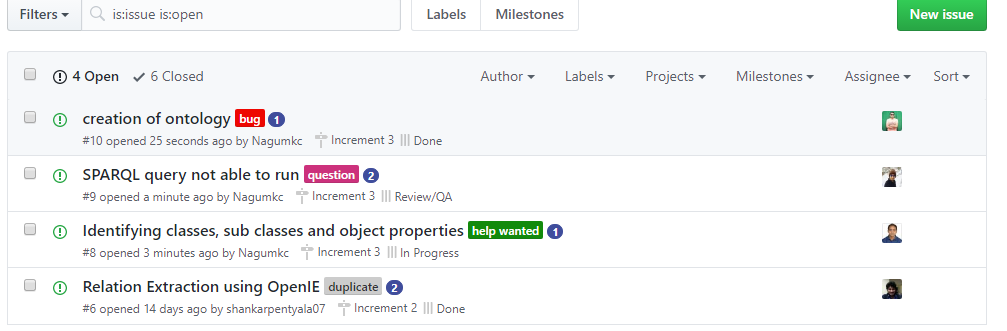
Board:



Burndown report:



Issues:



Future work:

Till now, we have created the ontology and the data statically from the Data set using the NLP techniques and have created question and answering system based on the extracted data.

The Future scope is to generate the Question and Answering based on the Knowledge graph dynamically by parsing data and finding out the entities and relationship between the entities. The main Entity extraction information is implicitly done using Natural language and explicitly done using structured data markup.