Submitted by

Prudhvi Suggula|SADANAND KALLAKURI|NIKITHA KONA|SAI SRIHARSHA |

Intelligent QA System

Increment 1

# **1.Motivation**

Humans are always in quest of knowledge. Information retrieval has become a form of knowledge discovery where we search on web to get relevant answers. Question Answering can be considered as an intersection of Natural Language Processing, Machine learning and artificial intelligence. QA systems are required in all the fields such as Health & Medical care, Education, Customer service as its always good to have assistance from computers.

# **1.1. Objective**

Our main objective is to design and implement a question answering system which can answer the posed questions accurately. We would start off with basic approach where our system can list the people, locations and organizations in the dataset, the we would refine our architecture by implementing word2 vector, TF-IDF, n-gram techniques where the system can answer way better than the previous approach, we will look at the drawbacks of this system and refine our architecture to yield the best results.

# **1.2. Significance**

Question and Answering systems are widely used now and they are trending in the field of Artificial intelligence and Machine learning. Few such examples would be Siri, IBM Watson, Amazon Echo that have done tremendously well and beat the national champions in many Atari games. Their significance lies in understanding the user queries and giving the best relevant results in short span of time. There has been a tremendous improvement that has taken place and still we are trying to minimize the gap between completeness and correctness.

# **2.Question Answering Application**

We are implementing a question and answering system on Sports where the user can ask about the current matches held on various games, live scores , locations of the games, ticket availability, weather conditions. We will start with the basic questionnaire system and incrementally build an enriched QA system.

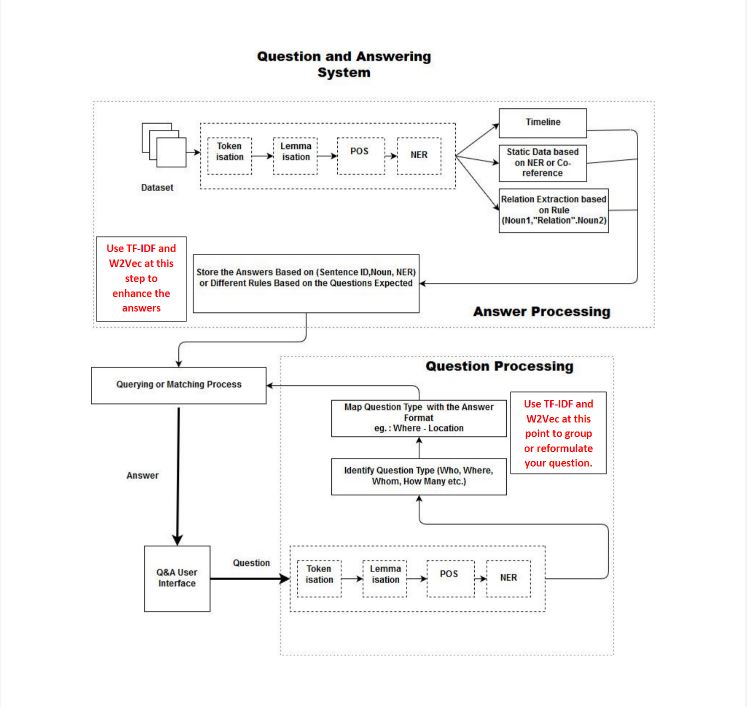
# **3.Dataset**

We have chosen **wikiref150 and BBC Sports** datasets for our project.

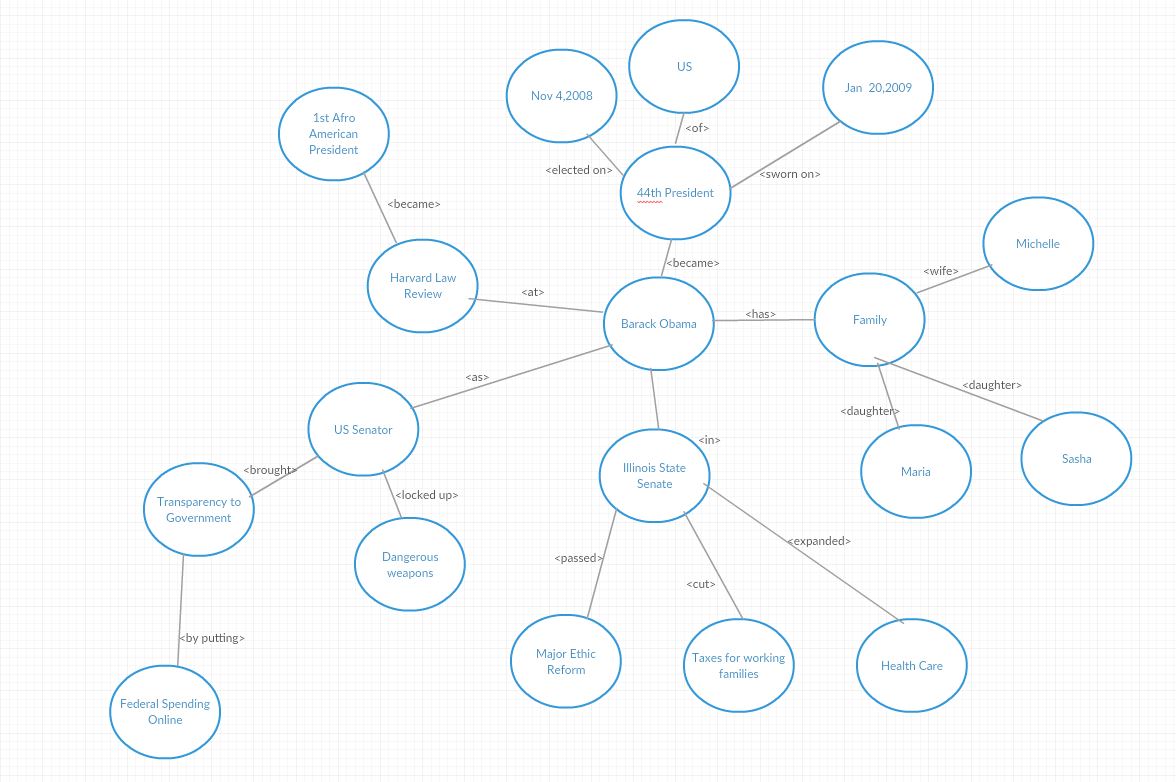
**BBC Sports Dataset Link:** <http://mlg.ucd.ie/datasets/bbc.html>

**Wikiref150 Dataset Link:** http://mklab.iti.gr/project/web-news-article-dataset

# **4.Workflow**



# **4.1 Knowledge Graph**



## **4.2 Question Answering**

The questions asked would be as follows.

1.Which location is the India vs Pakistan match held?

2. List the people that are playing the match?

3. What is the pitch condition there?

4.Who won the toss?

# **4.3 Related Work**

## **4.3.1 Evolution of Knowledge Graph**

In last few years lot of efforts are laid on development of web knowledge graphs. Noteworthy mentions are knowledge graph developed by google , wikidata , DBpedia, freebase , satori etc. This Paper consists of analysis of survey which emphasis on knowledge graph refinement

**Categorization of Knowledge Graph Refinement Approaches**

Knowledge graph refinement methods can differ along different dimensions. This paper used completion vs correction

**Completion vs. Error Detection**

The two goals of knowledge graph refinement are adding missing knowledge to the graph also known as completion, and identifying wrong information in the graph which is error detection

**Target of Refinement**

 completion and error detection are further distinguished by the target kind of information in the knowledge graph

**Internal vs. External Methods**

While internal approaches only use the knowledge graph itself as input, external methods use additional data, such as text corpora

**Categorization of Evaluation Methods**

**Partial gold standard:**

For completion tasks, this means that all axioms that should exist in the knowledge graph are collected, whereas for correction tasks, a set of axioms in the graph is manually labeled as correct or incorrect

**Silver standard:**

Knowledge graph is not perfect and is of reasonable quality

**Retrospective Evaluation:**

Output of retrospective evaluation is judged by human for annotations. Then suggested completions are labeled.

**Approaches:**

Approaches are for knowledge graph refinement. they help for automatic completion and error detection. Both completion and correction can’t be achieved with a single approach. All approaches have many computations in common and are measured using confidence scores

## **4.3.2 Knowledge Vault**

Knowledge Vault is a probability triple store database. Each entry in the database is of the form subject-predicate-object-probability, the database was built by combining the knowledge base Freebase with the Wikipedia and approximately one billion web pages. Knowledge Vault is the largest database ever created.

Author says that using natural language processing the information from web was extracted ,The information provided using Google knowledge vault that includes : data from web ,prior data from all social websites has huge set of database facts ,the tools include extractors, graph based priors ,knowledge fusion. All the text in these sources is mined from web sources , probability of each extractor assigns where each relation get assigned a score to an extracted

The author mentioned that facts extracted using a evaluation protocol which contains billions of facts has a technique to categorize data into a training set and test set. Implications for these facts uses a local closed world assumption .

Author mentioned that the path ranking algorithm can guess that if two people are parents to the same child, then it is likely that they are married and few other examples which are mentioned in the paper the approaches on automatic knowledge base constructions built on wikipedia and other structured data sourced which is used to open information extraction techniques which is applied to web overall and extract information from entire web but uses a fixed schema to construct hierarchies to knowledge base with various types of predicates.

## **4.3.3 Deep Dive Declarative Knowledge Base Construction**

The paper describes about the data extraction and representing them in the way, so they can be used for further knowledge base construction. Author raised the problem converting the data from unstructured data to well representational knowledge graph. They proposed a end to end system which helps the uses to write the udf which acts a schema from making knowledge graph. then applying machine inference, statistical tools to build decisions upon it . Domain experts try to create a schema that crushes the unstructured data to make a knowledge graph.

This try to analyses the text pdf and figures from and perform basic data cleaning tasks and apply udf to create schemas. Deep dive enables one to extract and integrate and predict the single system. He uses own representation for making relations between entities they are entity, relation and method. they m uses a high-level language to describe inputs and outputs. it also has a sql based query language which help at query time.

He describes about the use cases about the system in pharmaceutical industry, as palenoDB and paleno DeepDIve . this uses case is about biological classify of fossils. They need the domain experts to make the rules which help in making the ontology of the system. He explains about a how they dealt with the system to make for extracting the data from the research papers to make a decision upon it. He also tells about the how this can applicable to any domain.

Challenges faced by them are dealing with the unstructured data and multinomial inputs. He talks about different synonyms that they deal with entity, relation and mention. They try to search the corpus using the entities and which results in relationships. Framework it uses the input corpus and then use the candidate generation and feature extraction then supervision is done by the domain experts. By learning and inference task this helps in building the reasoning capability which results in output.

The candidate generation and feature extraction ­­are helps in mapping the different entities for the making the fasts from data. Feature extraction data then associated with the other feature candidates. the supervision helps to make identities that new relations.

Error analysis of the system that process incorrect queries that they need the supervision from domain experts it updates the udf which updates the ontology. the Author describes about both hardware and statistical efficient.

## **4.3.4 Semantic Data Integration for Knowledge Graph Construction at Query Time**

As a result of the evolution of Web Services from web of documents, there has been a rise in the amount of data available from almost all domains. Consider DBpedia or Wikidata, these general domain knowledge bases allow for accessing of knowledge about a wide variety of entities which include people, organizations and art paintings. These general domain data sources usually publish data in many different formats. Hence, there arises a need to have integration techniques which provide a unified view of the published data.

Fuhsen is devised by taking all these issues into consideration. Hence Fuhsen can be formallhy defined as an data integration approach. It employs keyword and structured search capabilities of Web data sources. Fuhsen cretaes on-demand knowledge graphs, which has the data merged from available web sources. It generally relies on Resource Description Framework(RDF) for semantically describing the collected entities, also depends on semantic similarity measure to define relatedness among different entities that have to be merged. Upon giving a keyword to Fuhsen, it produces an OnDemand knowledge graph at query time, which is generally composed of all the data merged by using the proposed semantic integration techniques. Traditionally, Jaccard coefficient is used to find the similarity between 2 rdf molecules basing on their triplets. Fuhsen receives keyword query Q, and similarity threshold value T. Molecules with materialized facts are then combined into a knowledge graph. An RDF integration module consists of three sub-modules. After the integration into RDF molecules, molecule enrichment takes place. Fuhsen allows provenance info that allows the user to trace the roots of a fact to a source. RDF molecule enrichment is carried out by two tools namely, DBpedia Spotlight and Silk Framework. Threshold is defined as the value which indicates a minimal similarity to be taken into consideration by linking engines. Weight can be described as a degree of importance.

Fuhsen constructs knowledge graphs out of the enriched RDF molecules. The input can be thought of as a set of molecules, while the output is an integrated RDF graph. Jaccard similarity measure computes a similarity score of any two molecules. Gades is semantic similarity measure, which is used to compare entities in a knowledge graph. Gades generally considers three main aspects, which are, class hierarchy, neighbors of the entities and the specificity of entities. The One-to-One weighted perfect matching is performed using the Hungarian algorithm. After performing all these tasks, similar molecules are identified and the final pipeline, integrates them into an RDF knowledge graph. This graph contains all unique facts of analysed set of molecules.

Concluding this, Fuhsen is an approach to integrate RDF molecules which are spread on different WEB data sources. The techniques proposed in the paper are implemented in FUHSEN, a federated hybrid search engine. Results of empirical evaluation suggests that Fuhsen is able to efficiently integrate pieces of data spread in different data sources. Fuhsen can be applied in many use cases, such as in price comparison among different e-commerce web sites and candidate profile building from web data in human resource management.

## **4.3.5 Fonduer Knowledge Base Construction**

Knowledgebase construction is a procedure of populating the database with text,column,images or videos. Broad endeavors have been made to construct extensive, top notch learning bases (KBs, for example, Freebase, Yago,Watson, Knowledge graph. Generally, KBC

have focused on unstructured data, with extraordinary achievement. These KBC frameworks as of now bolster a wide scope of downstream applications, for example, data recovery, question replying, therapeutic analysis, and. In any case, troves of data remain undiscovered in lavishly organized information, where relations and properties are communicated by means of blends of literary, auxiliary, forbidden, and visual signals. In these situations, the semantics of the information are fundamentally influenced by the association and it operates on a variety of features, training data, Similar to knowledge graph there are various entities that play a major role like entities, relationships. Given a set of documents and the knowledge base structure extract the relationship R from D. Data programming experts takes the training datasets by computing the weak sources present. We have user defined schemas that that defines the relationship to be extracted which is used by fonduer to initialize relation database. ot at all like KBC from unstructured content, KBC from lavishly designed information requires supervision from numerous methodology qualities of the information. In luxuriously designed information, valuable examples for KBC are more meager and covered up in contextual data. Furthermore, redundancy in an assortment of examples over different modalities. Fonduer's bound together information display enables clients to specifically express rightness utilizing printed, basic, forbidden, or visual attributes, notwithstanding conventional supervision sources like existing KBs. In the Electronics space, over 70% of marking capacities composed by our clients are based on non-literary signs. It is worthy for these naming capacities to be boisterous and struggle with each other. We have performed experimental results on various categories like ctronics,Advertisements,Genomics and Paleontology. We compare the results using the state of art oracle comparison and based on the experimental results Fonduer’s outperforms the state of art systems,also a detailed comparison between Fonduers and existing knowledge base. Based on the results we can conclude that Fonduer performs well and is the present state of art KB construction.

## **4.3.6 Conclusion**

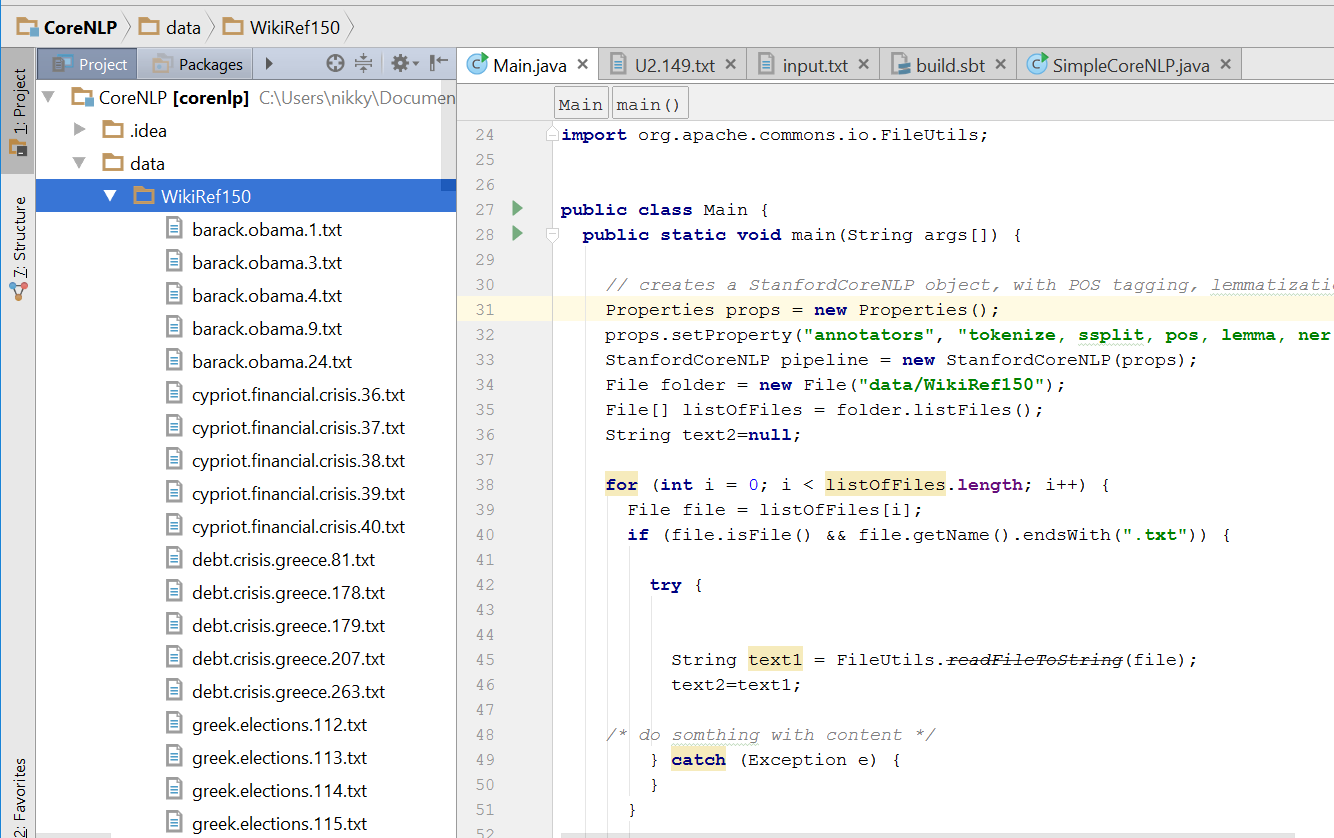
We cannot improve and correct knowledge graph at same time. One should be traded off. It is hard to compare approaches and make generalize statement on performance, scaling is main issue. We have various methodologies in constructing knowledge graphs and its always tough to have both completeness and correctness. With the present state of Art techniques and mechanism we can achieve the best of the results.

# **5.Implementation**

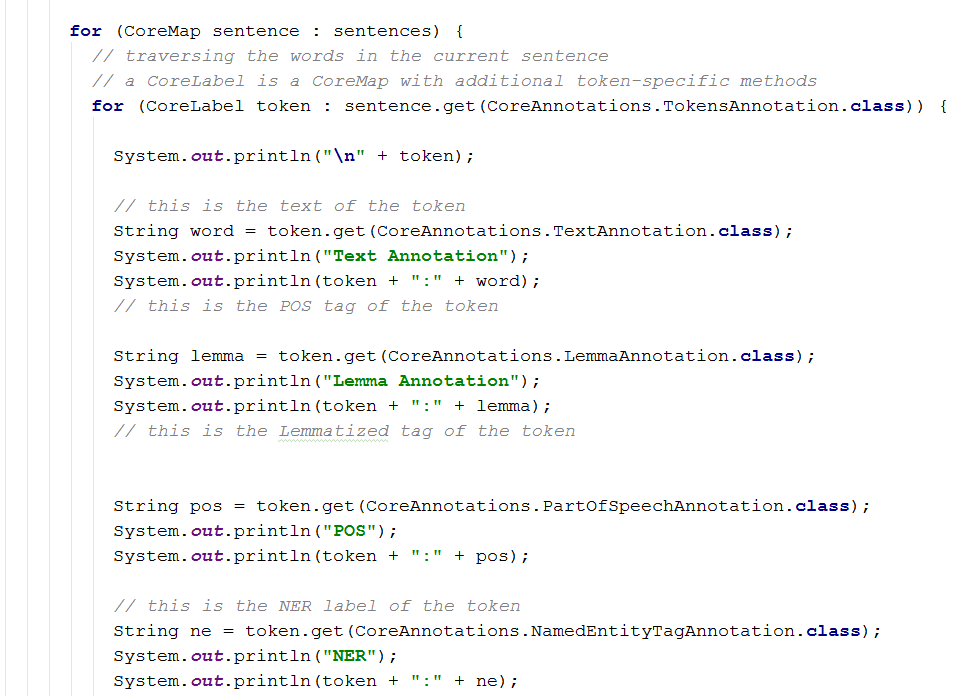
# **5.1 NLP with Dataset**

## **5.1.1Reading the Entire Dataset**

We have taken **wikiref150** dataset. We read the entire dataset and performed the all the NLP operations.

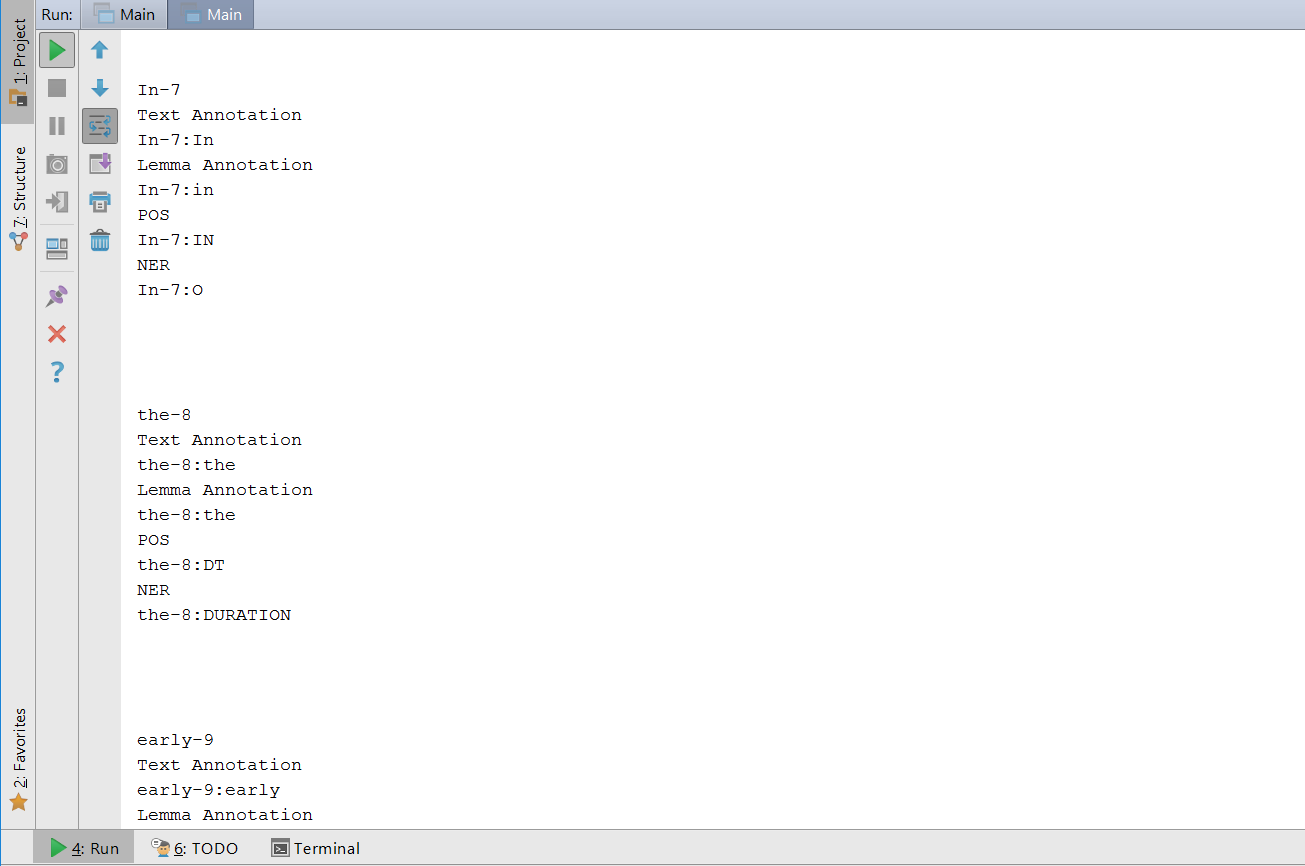


## **5.1.2 Program to Perform NLP Operations**

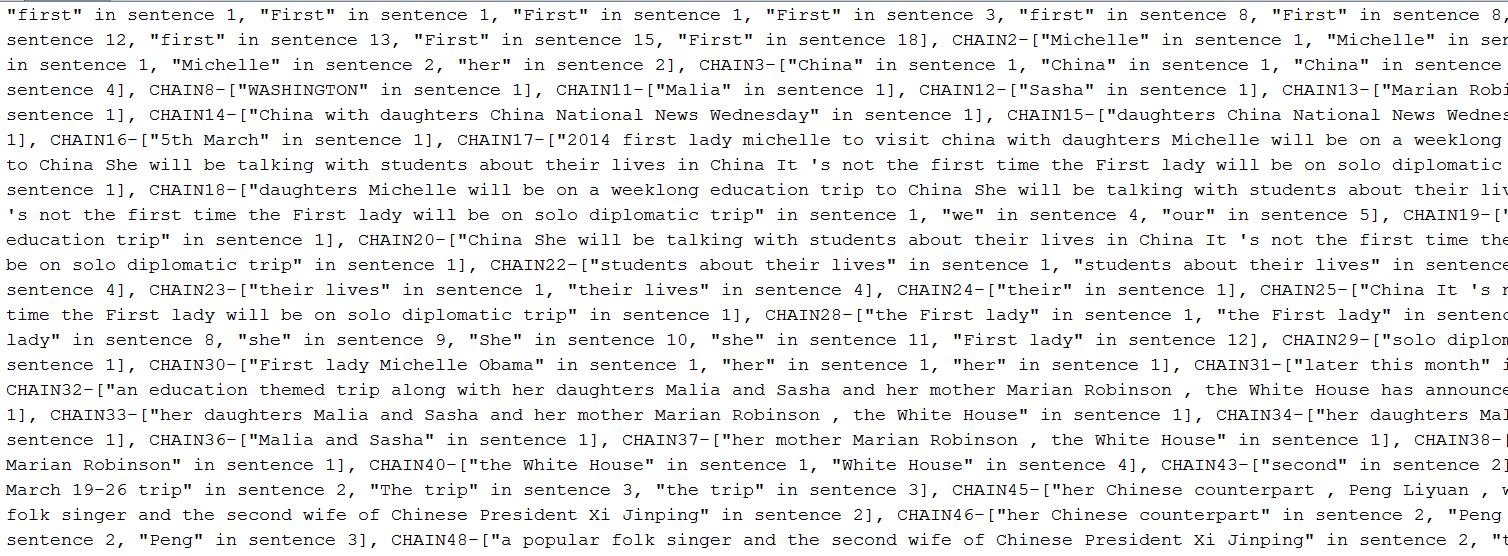


## **5.1.3 Output Screenshots**

### **5.1.3.1 Text Annotation,Lemma Annotation,POS,NER**



### **5.1.3.2 Co-referencing**

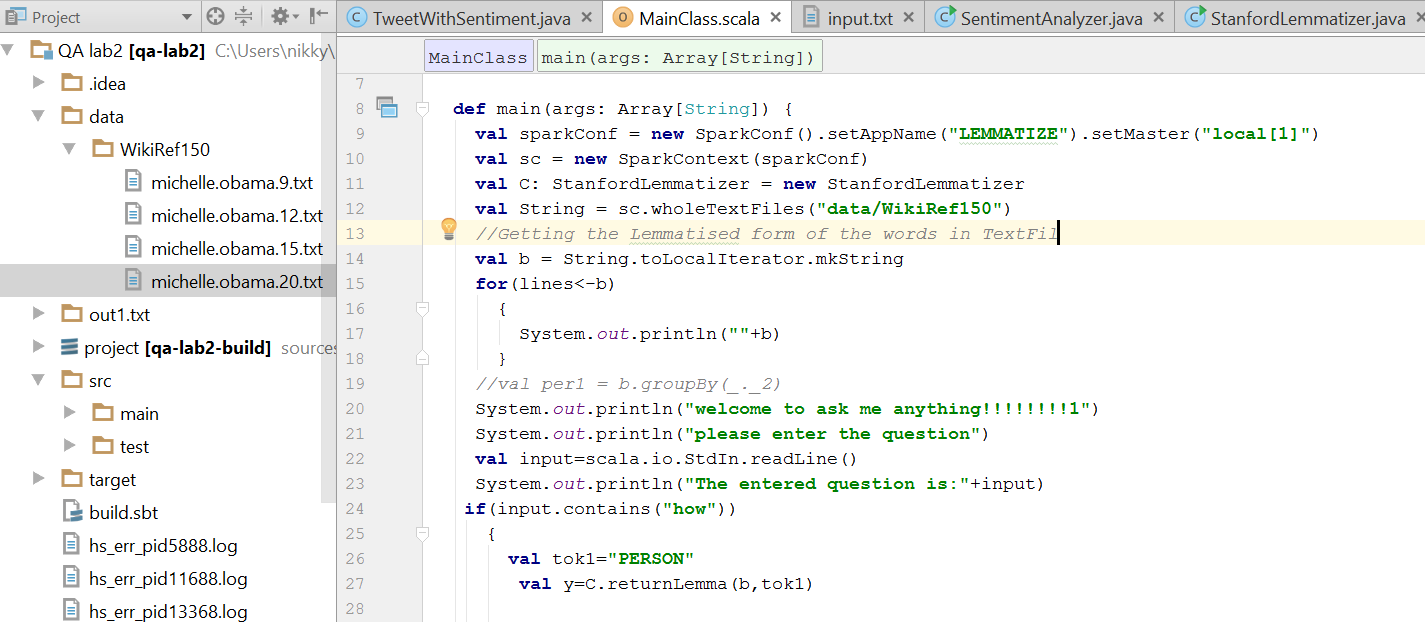


## **5.1.2 Question Answering**

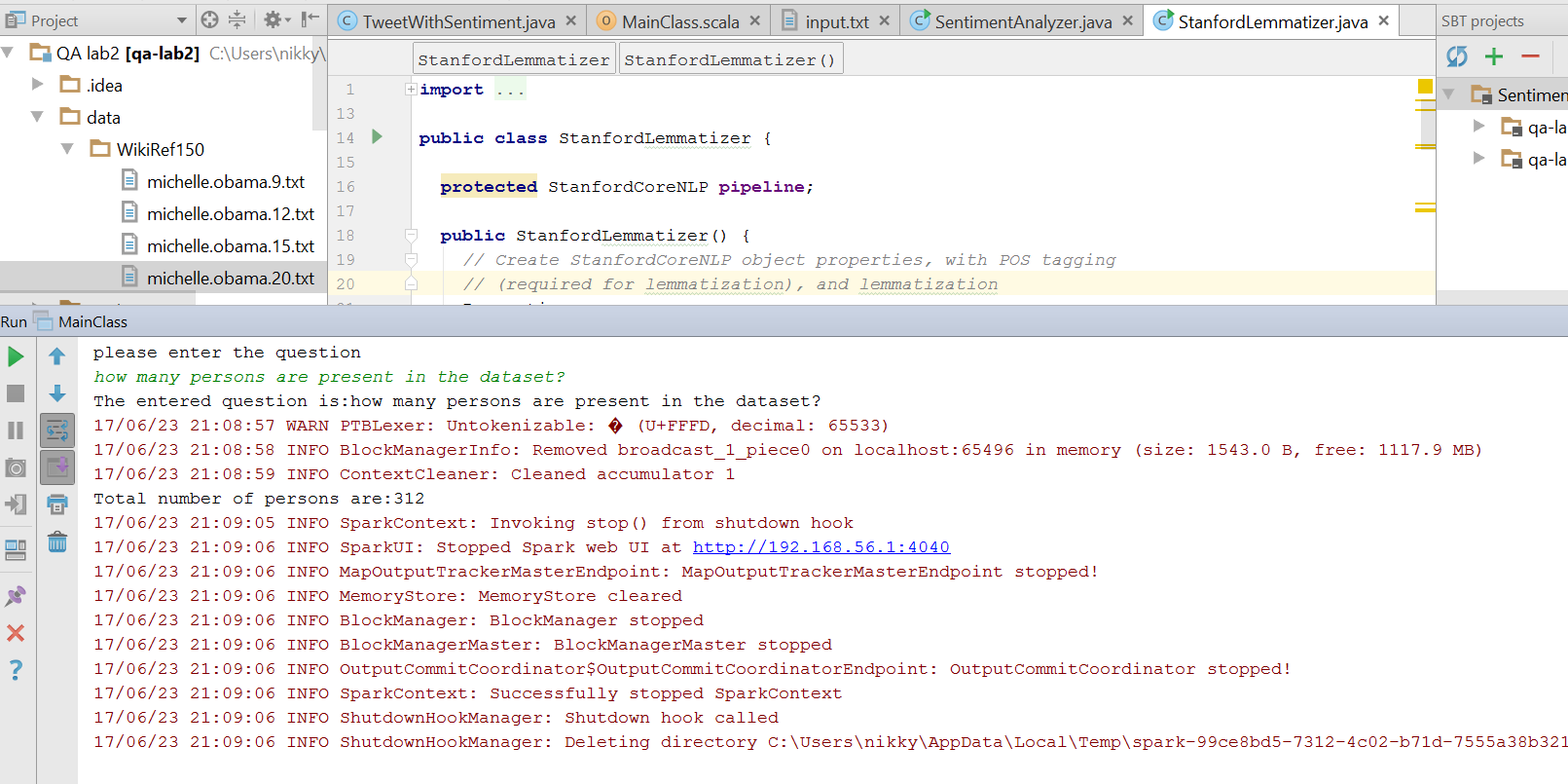
We have implemented a simple question answering system that answers basic questionaries’ asked to the system.

### **5.1.2.1 Scala Program Reading the Entire Dataset**

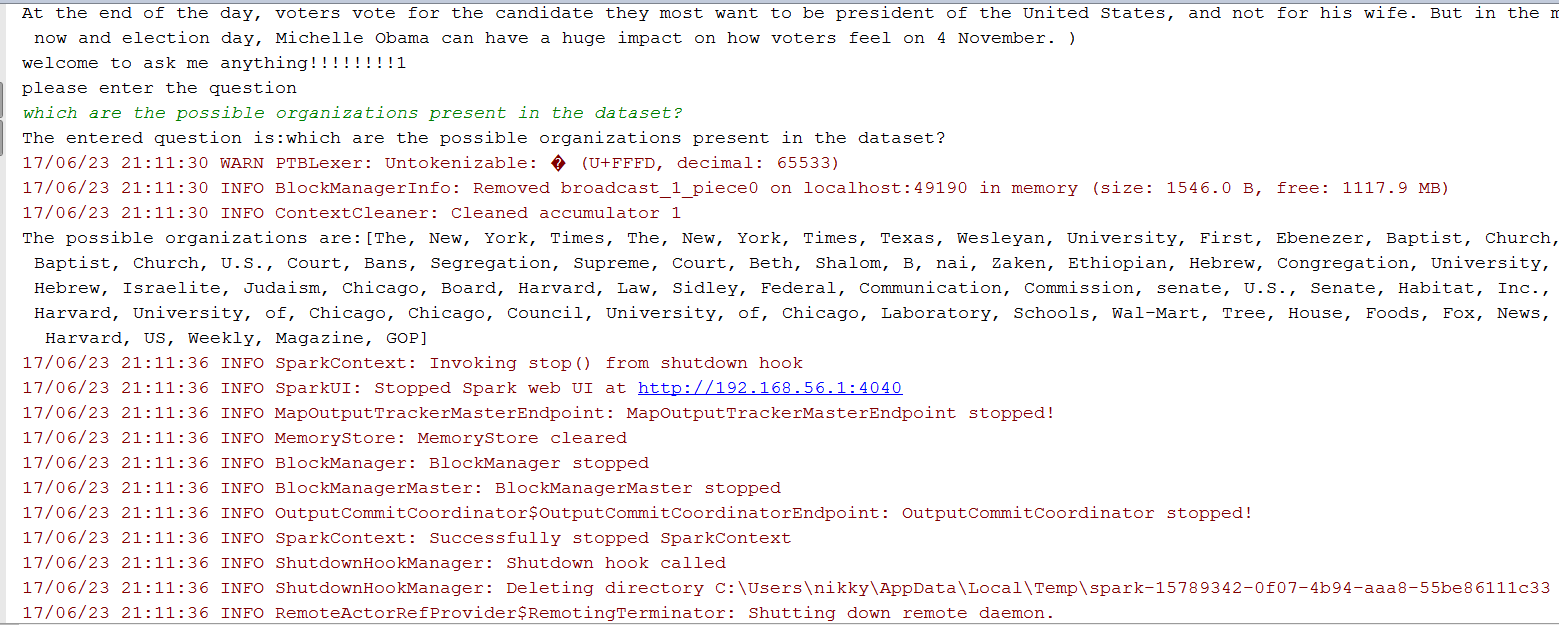
The below scala program takes the entire dataset and performs a basic question and answering.



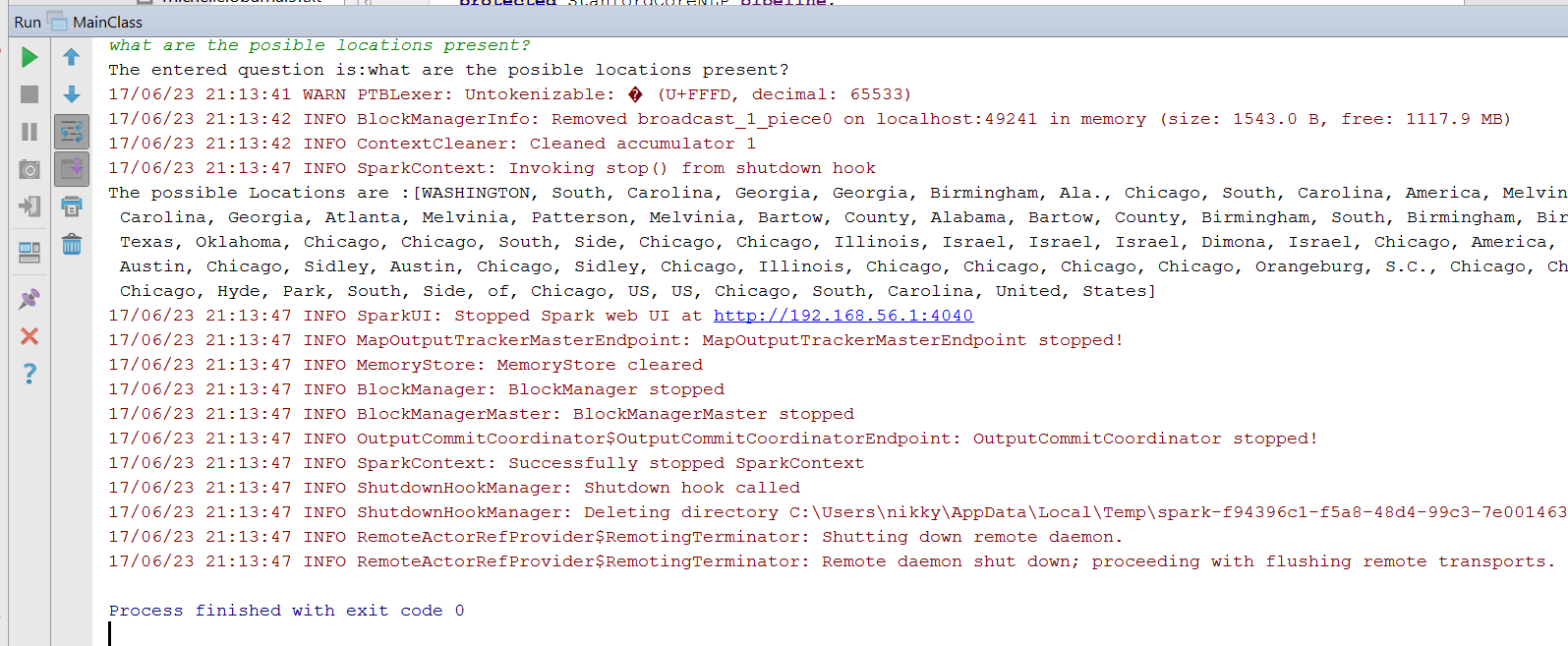
### **5.1.2.2Question1-How**



### **5.1.2.3 Question-Which**



### **5.1.2.4 Question-What**



# **6. Increment 2**

## **6.1 Information Retrieval (TFIDF, Word2Vec)**

To measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either *relevant* or *nonrelevant* for each query-document pair.

The standard approach to information retrieval system evaluation revolves around the notion of *relevant* and *nonrelevant* documents. With respect to a user information need, a document in the test collection is given a binary classification as either relevant or nonrelevant. This decision is referred to as the *gold standard* or *ground truth* judgment of relevance. The test document collection and suite of information needs have to be of a reasonable size: you need to average performance over fairly large test sets, as results are highly variable over different documents and information needs. As a rule of thumb, 50 information needs have usually been found to be a sufficient minimum.

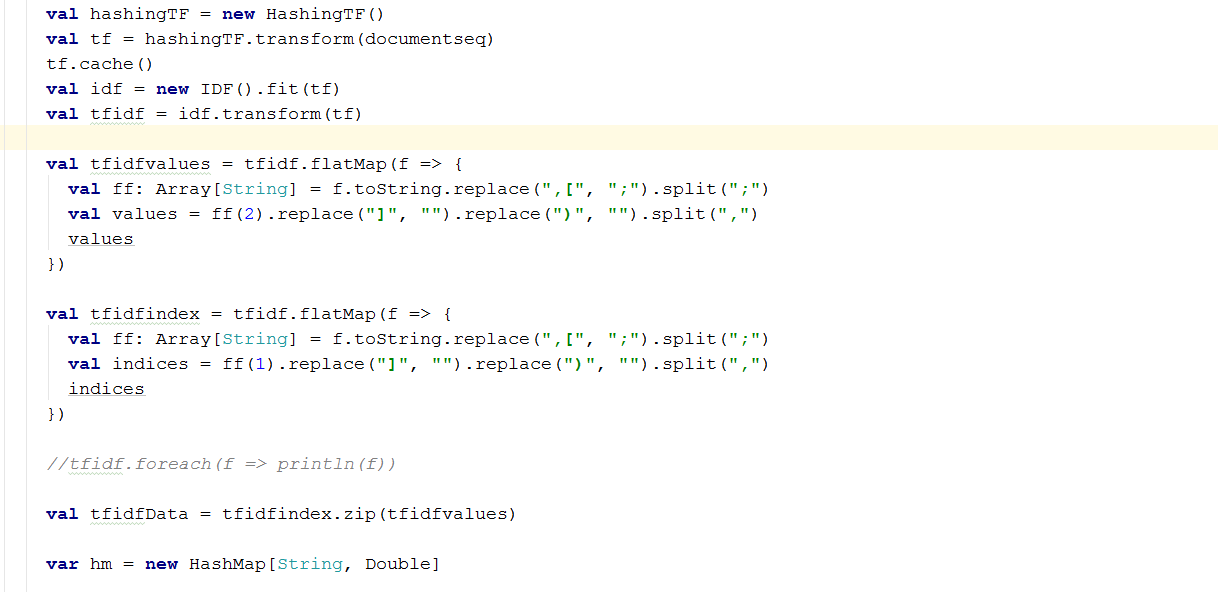
## **6.2 TFIDF**

Tf-idf stands for*term frequency-inverse document frequency*, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

### **6.2.1 Implementation Screenshots**

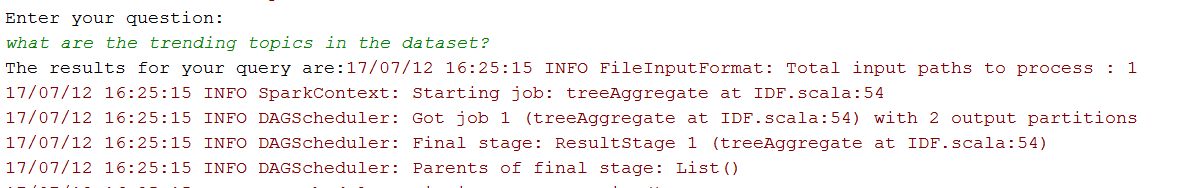
We have taken the given dataset performed stop word removal and then parsed the dataset and have sent the dataset to TFIDF which gives the high frequency words in the dataset.

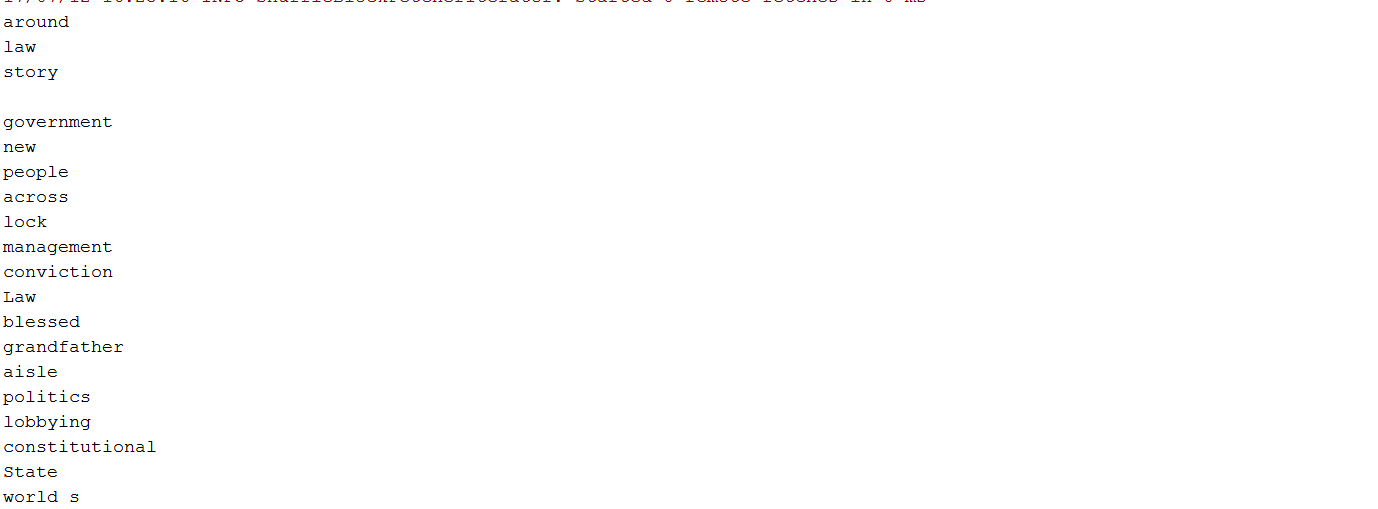




### **6.2.2 Output Screenshots**

The below screenshot shows the high frequency words, or the trending words in the dataset.





### **6.2.3 Stop Words Removal Vs No Stop Words Removal**

When we performed the given Tt-idf with and without stop word removal we have seen a tremendous improvement in the output and we have shown the representation in tabular format

|  |  |
| --- | --- |
| Stop Word with TFIDF | No Stop Word Removal |
| around  law  story  government  new  people  across  lock  management  conviction  Law  blessed  grandfather  aisle  politics  lobbying  constitutional  State  world s | from  who  around  law  story  for  with  to  on  he  He  and  new  plants.  across  lock  management  conviction  Law  blessed |
|  |  |

## **6.3 Information Extraction**

**Information extraction** (**IE**) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be information extraction. Due to the difficulty of the problem, current approaches to IE focus on narrowly restricted domains. An example is the extraction from news wire reports of corporate mergers, such as denoted by the formal relation:

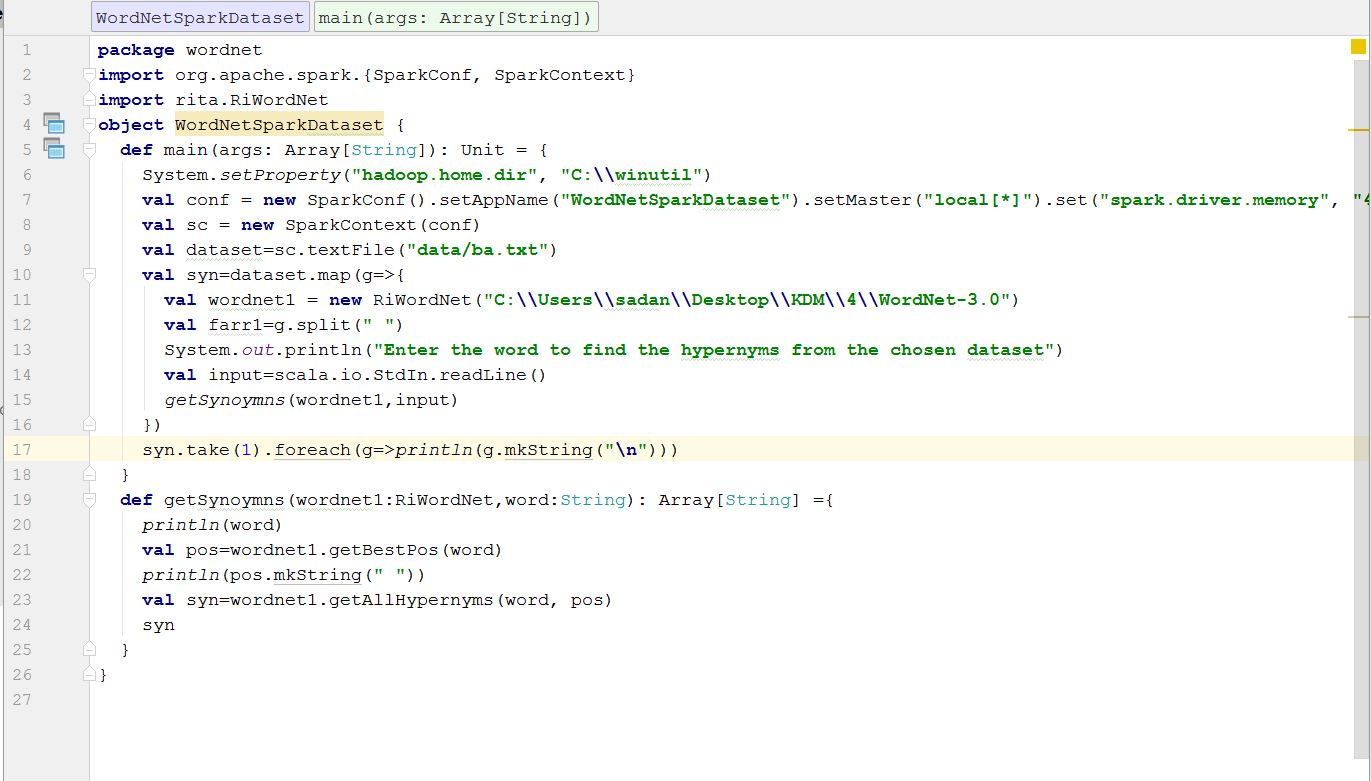
{\displaystyle \mathrm {MergerBetween} (company\_{1},company\_{2},date)}

## **6.3.1 WordNet**

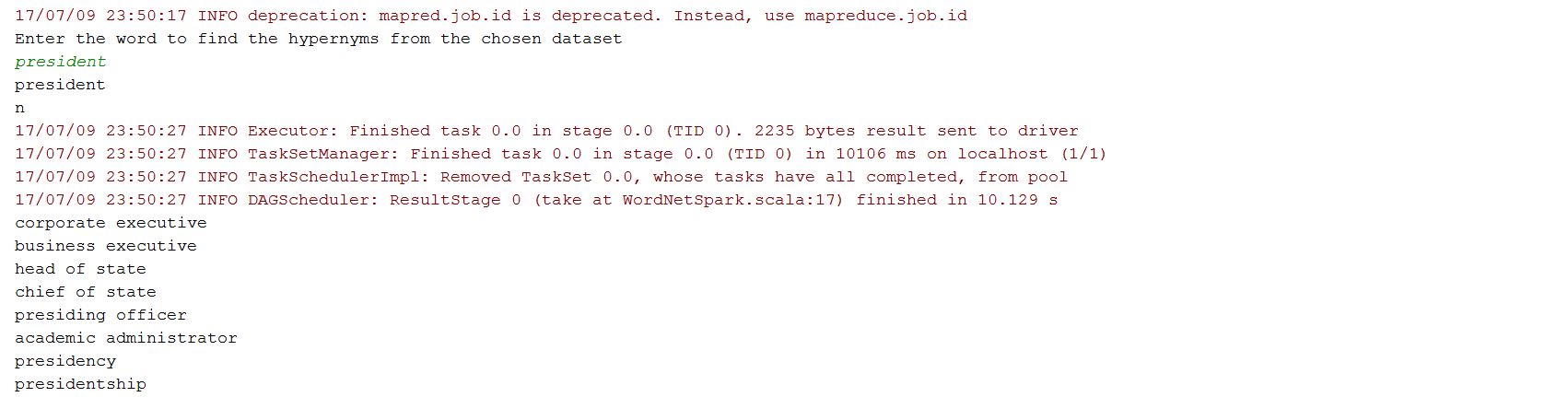
WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

## **6.3.2 Implementation Screenshot**

In the given dataset we have designed a QA system where the user can ask the hypernyms and based on the dataset the model asks the questions. The program and the output screenshots are given below.



## **6.3.3 Output Screenshots**



## **6.4 N-gram Approach to get Full Name and Related Name**

N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward (although you can move X words forward in more advanced scenarios). For example, for the sentence *"The cow jumps over the moon"*. If N=2 (known as bigrams), then the ngrams would be:

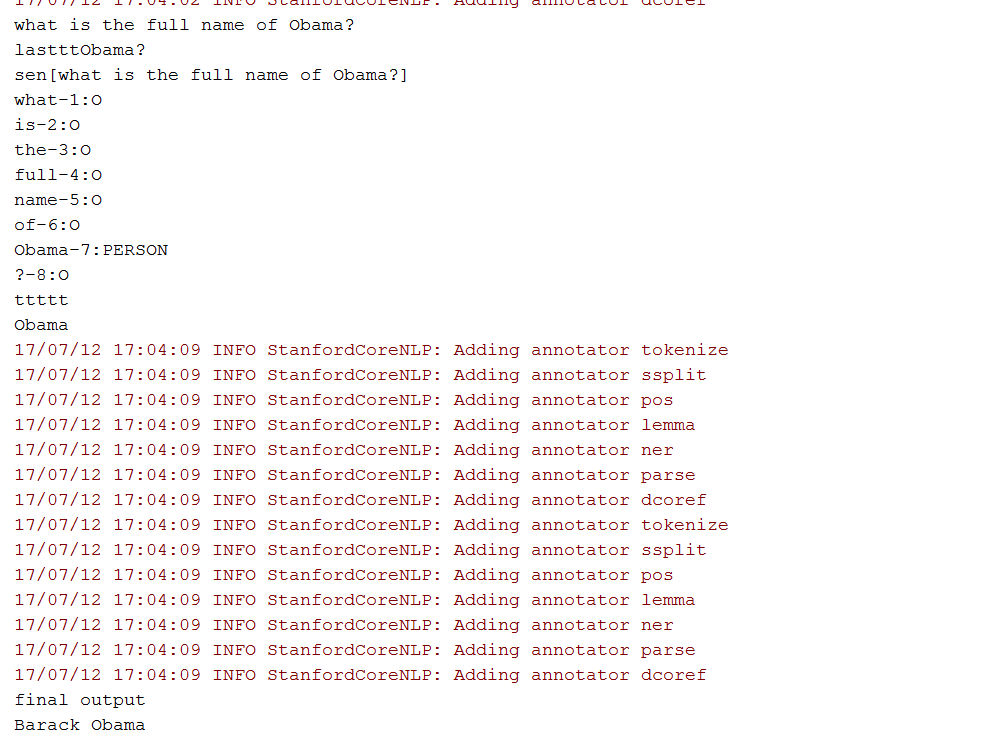
* the cow
* cow jumps
* jumps over
* over the
* the moon

## **6.4.1 Program Screenshots**

In the given dataset, we have taken the given input and the user is prompted to get the name of the person and the model will give the full name of the given person. We have shown the program and implementation screenshots.



## **6.4.2 Implementation Screenshot**

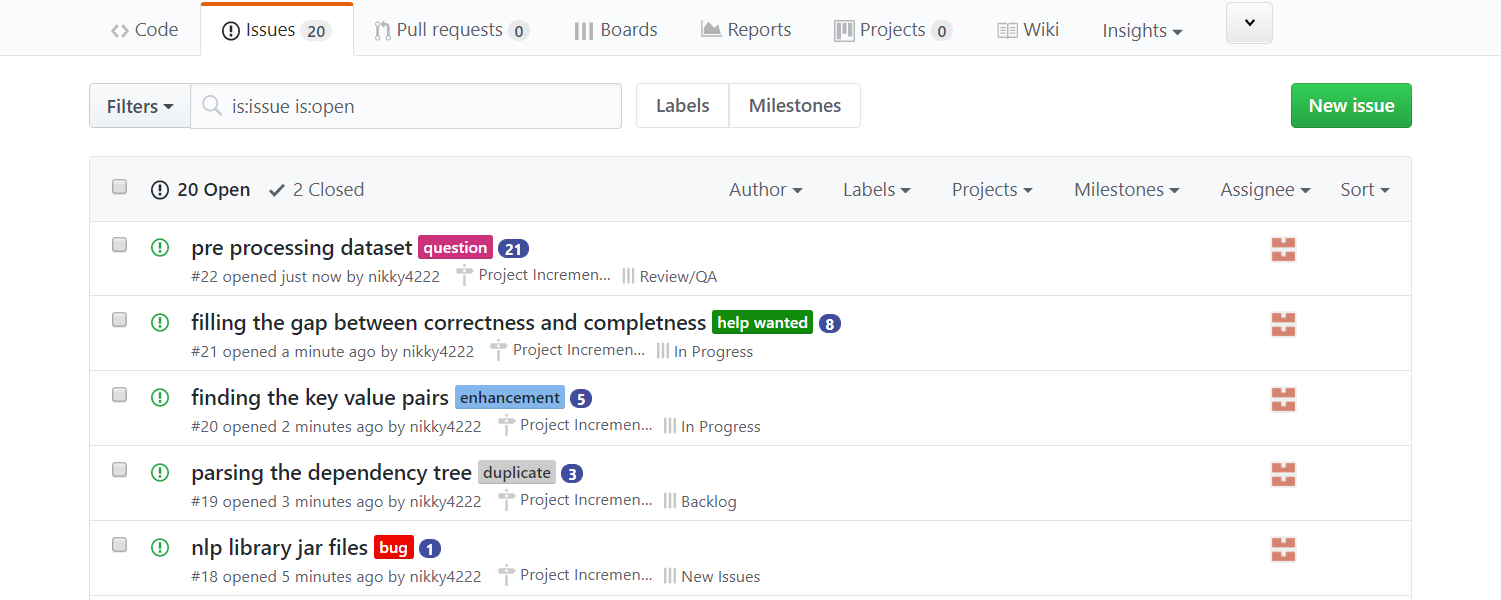


# **6.Project Management**

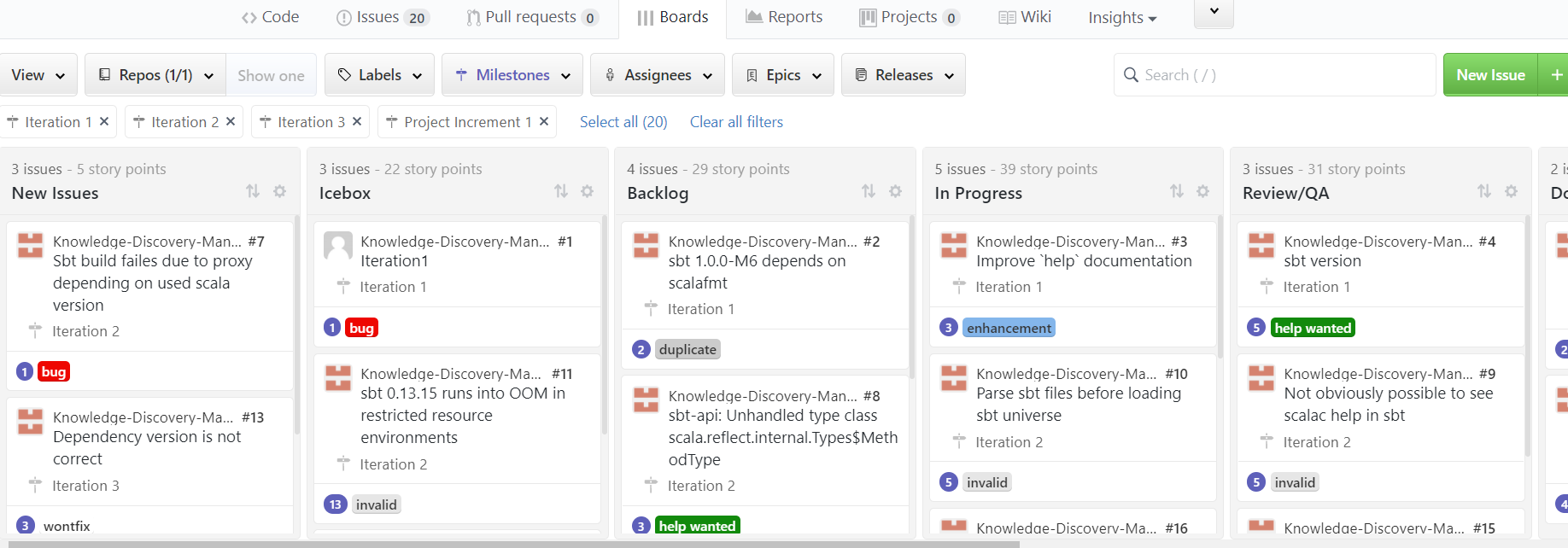
## **6.1 Contribution**

Prudhvi:25%||Sadanand:25%||Nikitha:25%||Harsha:25%

## **6.2 Issues**



## **6.3 Board**



## **6.3 Burndown Chart**



## **6.4 Future Works**

We would like to implement our existing question answering system to various sectors like medicine and healthcare, education, entertainment, also we would like to use the machine learning to enrich the ontology so that our QA system becomes rich.