ReDefine - Research Definitions Summarized
First Increment Report

Team ID: 7
Team name: Cognitos
Team members:
Nikhita Sharma - 37
Dig Vijay Kumar Yarlagadda - 47
Harsha Komalla - 14
Sai Madhavi Sunkari - 39

1. INTRODUCTION:

Motivation:

Our motivation for this project stemmed from having to read dozens of research papers just to understand all the terms in a single research paper. There is no simple way of understanding a research article, the information from the internet sources like Wikipedia doesn't provide a clear insight into many technical concepts. Having experienced this first-hand, we realized the importance of a system which can provide a concise summary of a research article.

We aspire to extend this system as a knowledge graph, which can be used as a handy search tool by students to find articles, summaries of articles, authors, and domains. This system has the potential to be useful in many different ways: a beginner in a field can explore research papers in a particular domain, a search engine can crawl through the knowledge graph for more relevant search results, etc.,

Objectives:

The objectives of our project are:

- Build a scalable system for summarization of research articles
- Create an interactive knowledge graph of research articles

Features:

- User can provide a corpus of research papers related to a field as input.
- Users are presented with an interactive graph UI, which can be explored to view the interactions between authors and articles.
- The summary generated can be read by clicking on a particular paper title.

Expected Outcomes:

1. Knowledge Graph:

We expect this project to generate a Knowledge graph representing all the publications in a specific field of study selected based on the requirement of the user. This graph would provide important results about the different papers written in the field, the paper titles, information about author of the paper, co-authors' information, other related papers an author or co-author has written and important results about each paper itself like focus of a paper, key terms in the paper and other important information.

2. Publication Summary:

This project will also generate a text summary of each research paper/publication. A graphical representation of a research article would show the most important keywords in the article and relation between them to help understand the focus of research of the author. It would also show the major fields and topics of discussion. The summary will give an overview of contents of each paper to the end user, without the user having to read the complete research paper.

2. DOMAIN OF THE PROJECT:

Our project focuses on building a knowledge graph model using NLP and ML techniques to automatically generate summaries from a dataset of research articles. Due to the difficulty of finding sources for obtaining research articles, we are limiting our dataset to a few thousand articles. However, our system is based on Apache Spark and we are confident that it can scale to large volumes. Further we are concentrated on improving the quality of summaries generated, we are not particular about generating output in different formats like text summaries or graphical user interfaces; the output will be generated as a knowledge graph.

3. DATA COLLECTION:

Finding an API which provides a dataset of entire research articles proved to be a difficult task, metadata of articles is available from many sources, but the actual full content of an article is not open sourced. We opted for using IEEE Xplore XML API. Initially we plan to perform summarization on 100 research articles and then extend to a larger dataset.

4. TASKS AND FEATURES IMPLEMENTED:

Cognitive services:

Using IBM Watson's document Conversion for converting research documents in image or pdf formats into readable text files.

NLP Processing:

StanfordCoreNLP for NLP processing which includes Tokenization, Lemmatization, Stop words removal, Name-Entity Recognition and POS tagging.

Information Extraction and retrieval:

Used SparkMLlib classes to calculate the TF values and IDF values. Tried to sort keywords with TFIDF values and getting top keywords.

5. IMPLEMENTATION SPECIFICATION:

A. Software Architecture & UML Model:

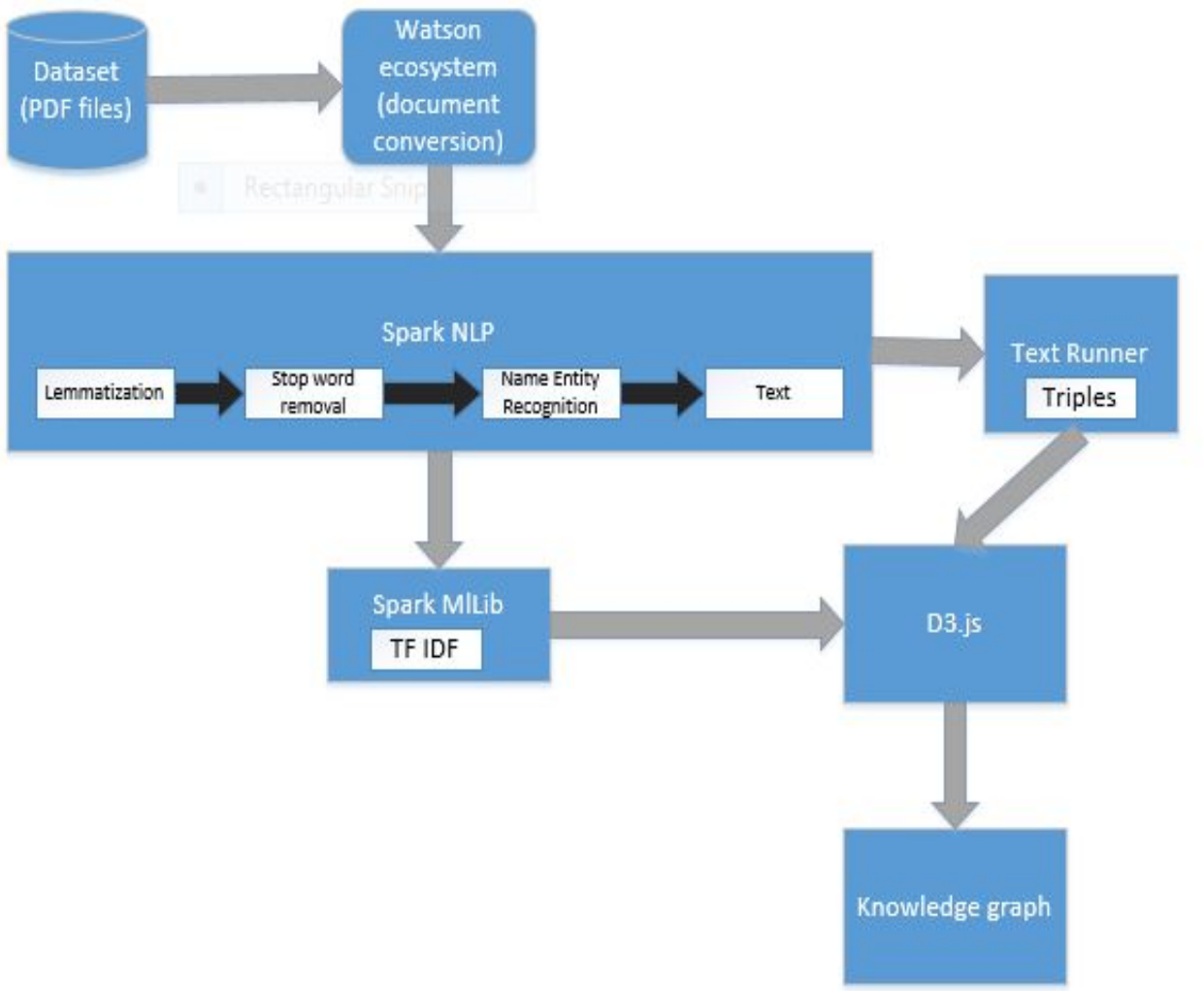


Fig 1: Architecture Diagram

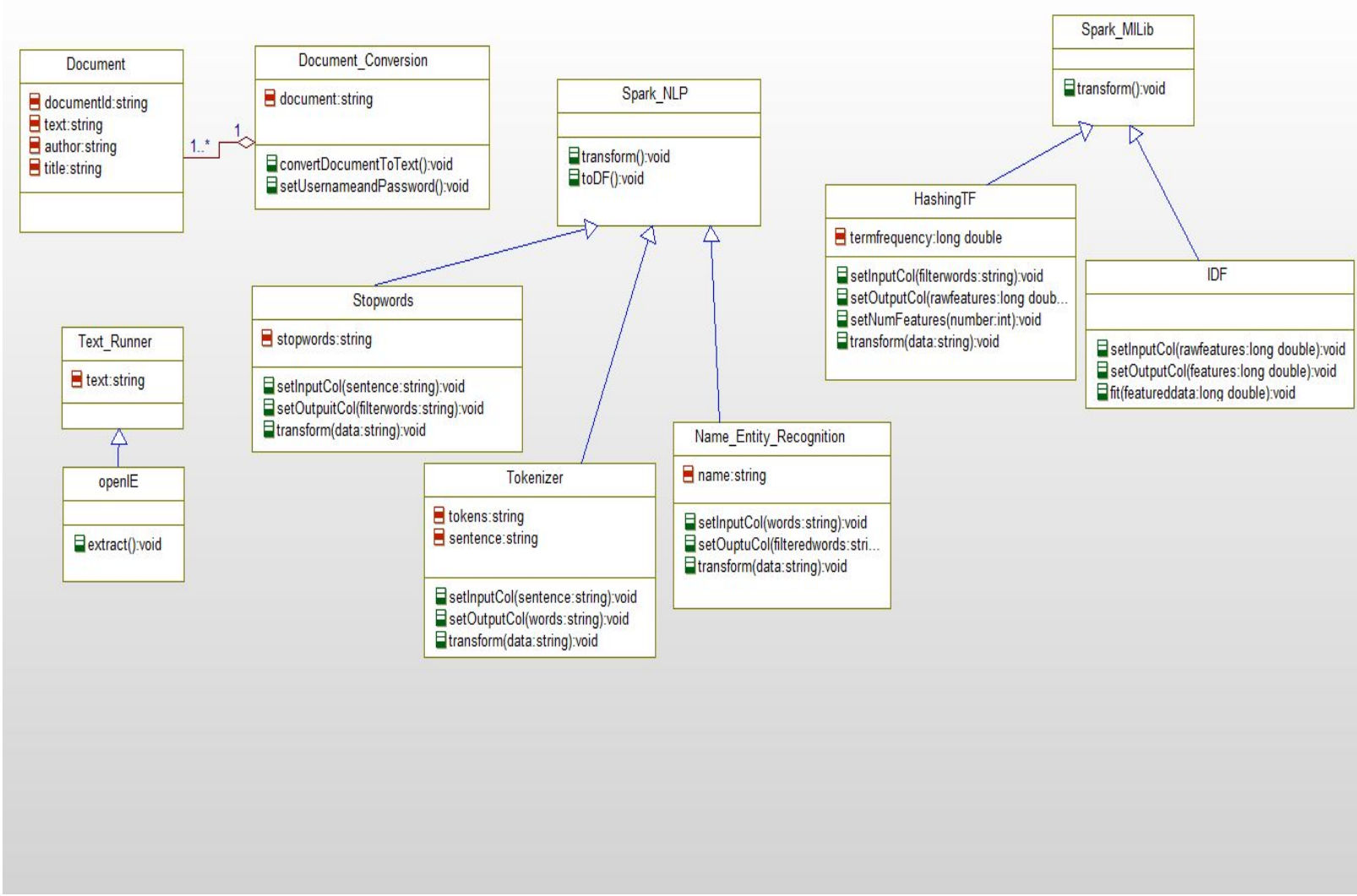


Fig 2: Class diagram

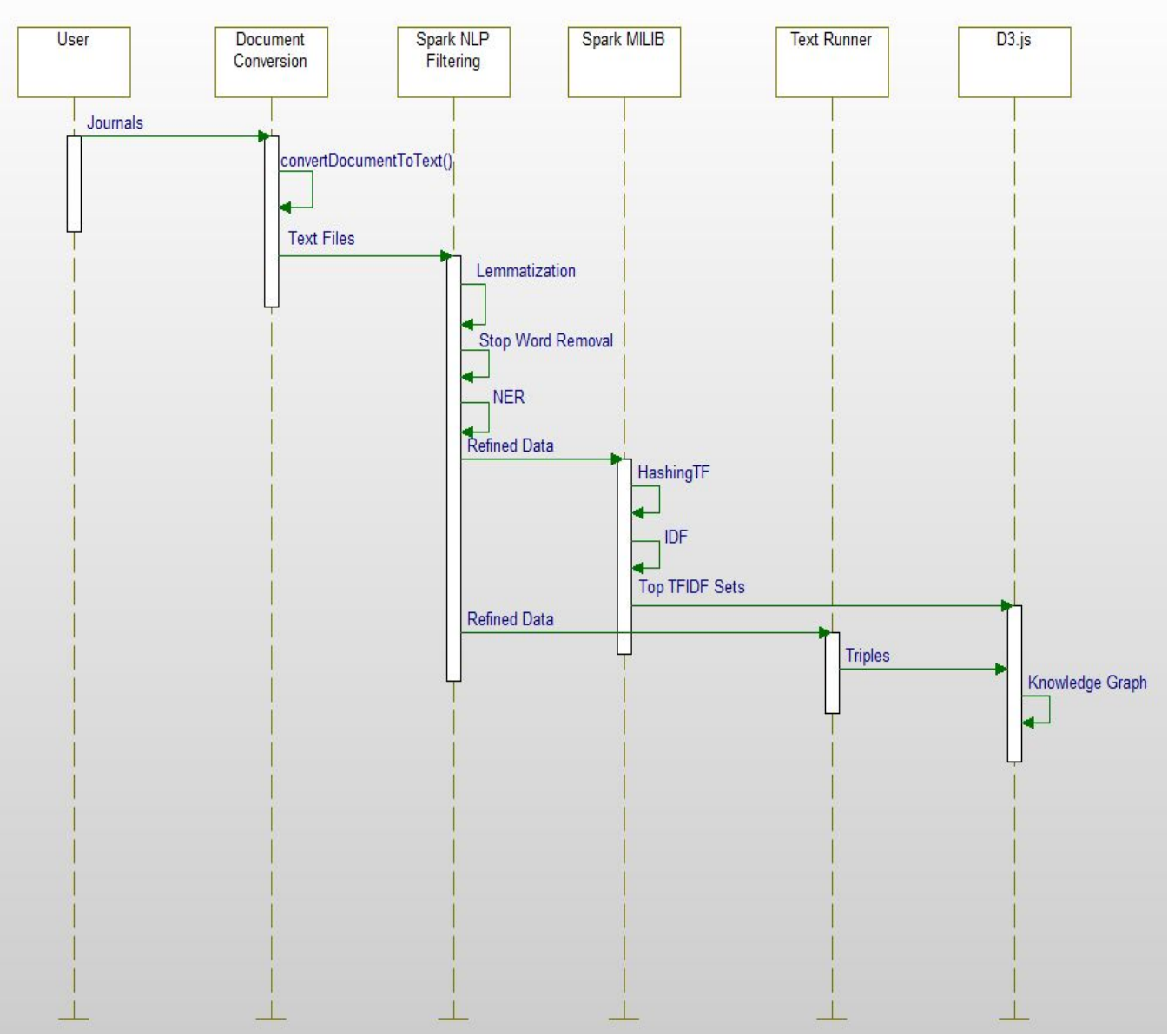


Fig 3: Sequence Diagram

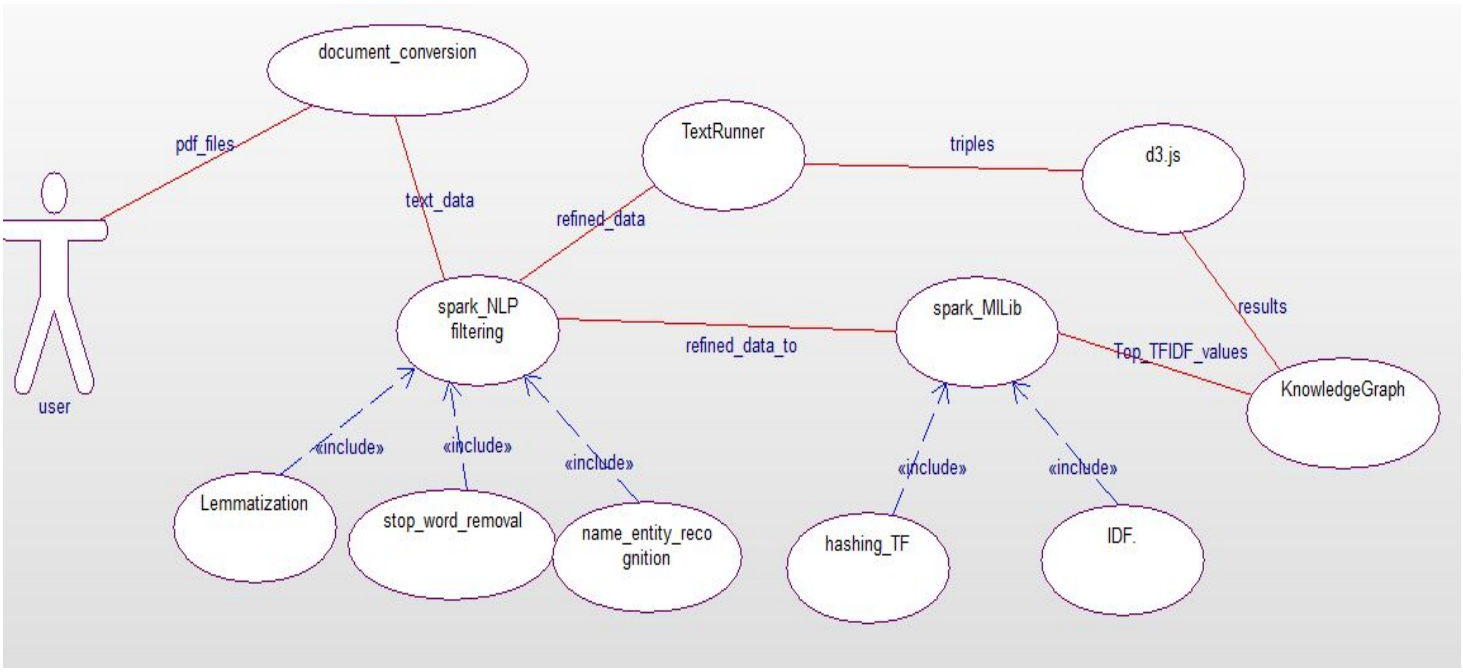


Fig 4 : Use case diagram

B. Workflow:

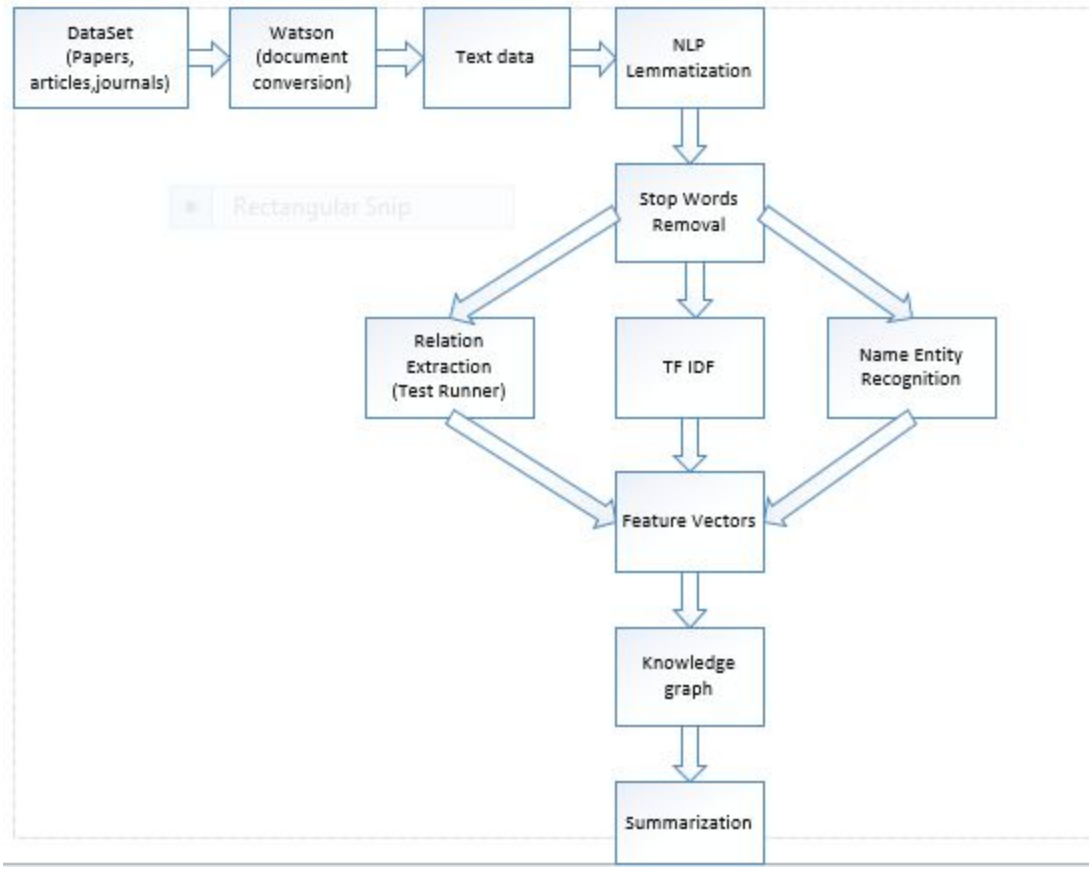


Fig 5: Workflow Diagram

C. Existing Services/APIs:

ReDefine uses different services and APIs for NLP and Information Extraction:

- Document Conversion API in Watson Developer Cloud Java SDK using IBM Bluemix, for converting research documents in image or pdf formats into readable text files.
- Text Runner - Open IE, for Relation Extraction
- D3.js for visualizing knowledge graph

D. New services implemented:

- Created module to pick a text research paper file as input and perform NLP and Information Extraction on the data

6. PROJECT MANAGEMENT:

A. Team members contribution: (overall)

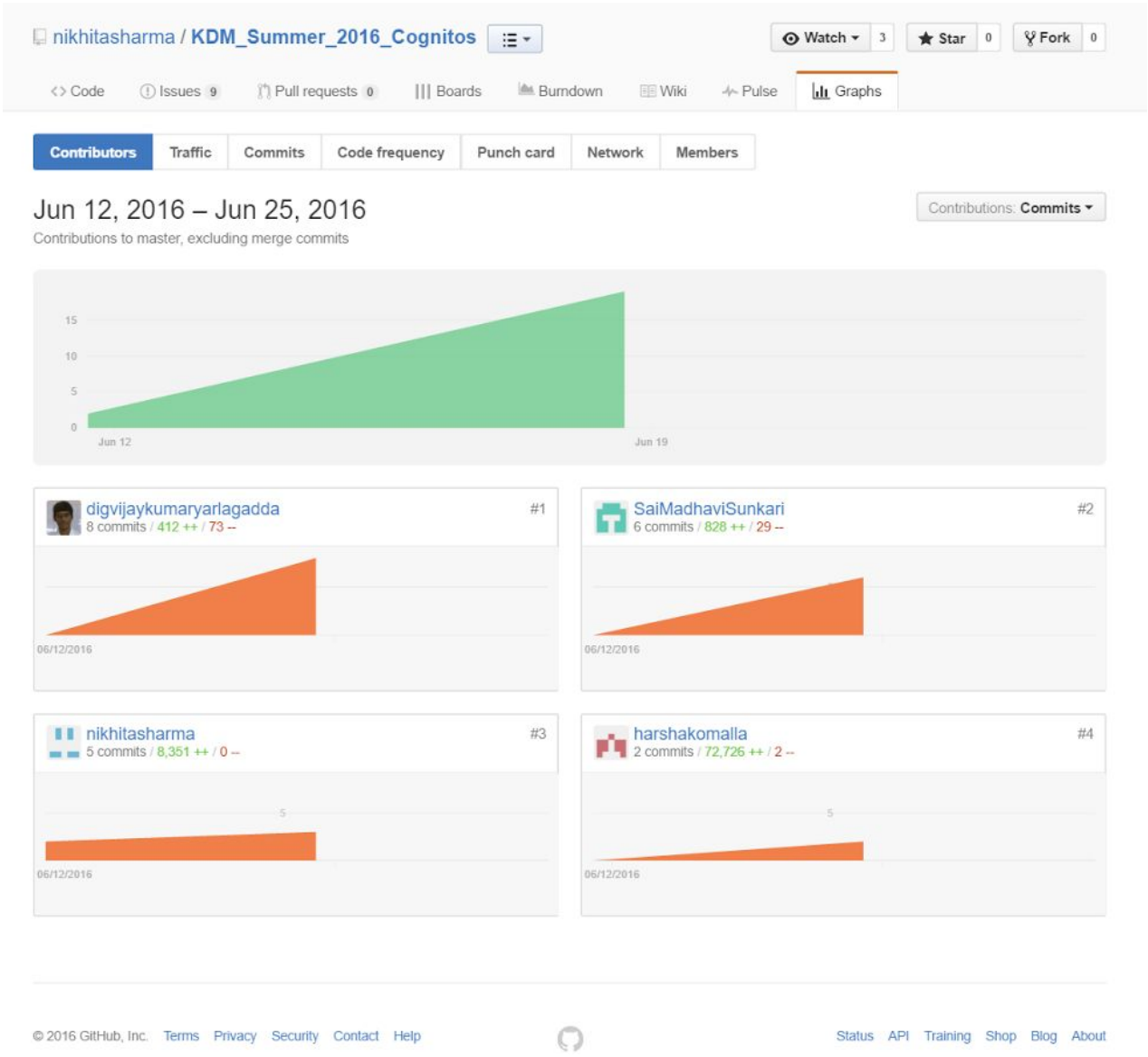
Nikhita Sharma - 25%
Dig Vijay Kumar Yarlagadda - 25%
Harsha Komalla - 25%
Sai Madhavi Sunkari - 25%

B. Zenhub and Github URL/Statistics

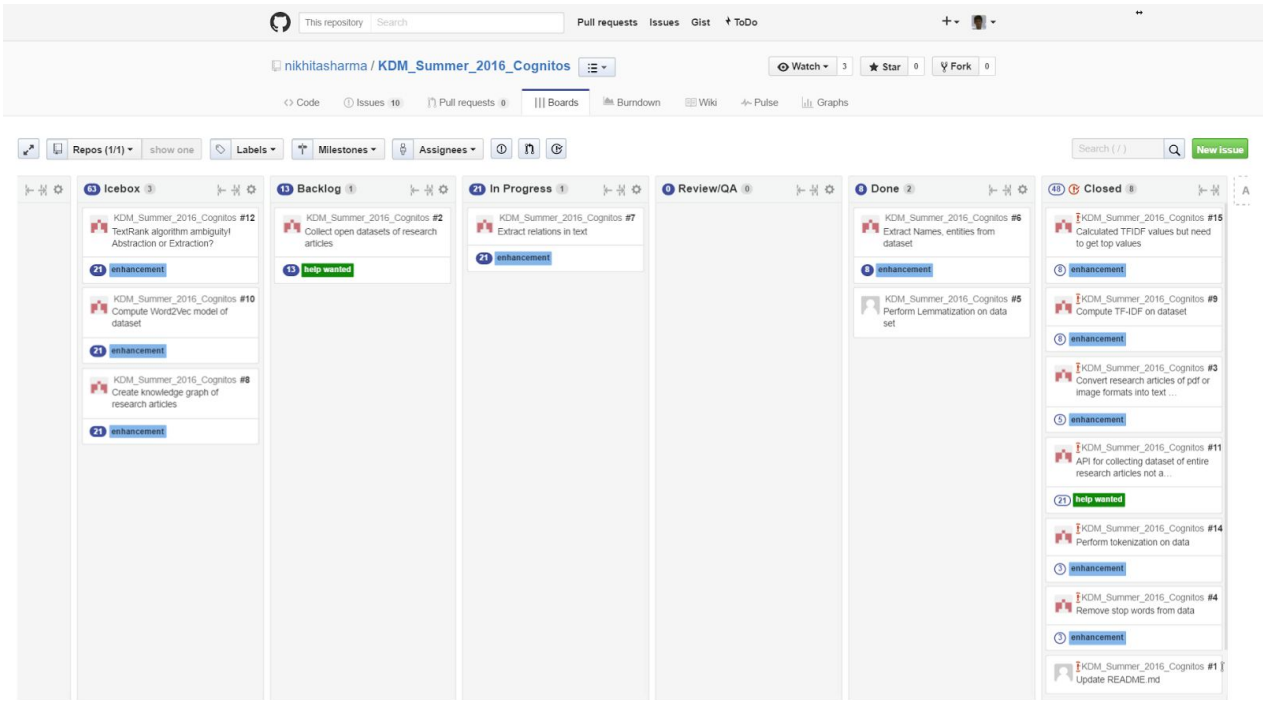
Project GitHub repository:
https://github.com/nikhitasharma/KDM_Summer_2016_Cognitos

Zenhub statistics:

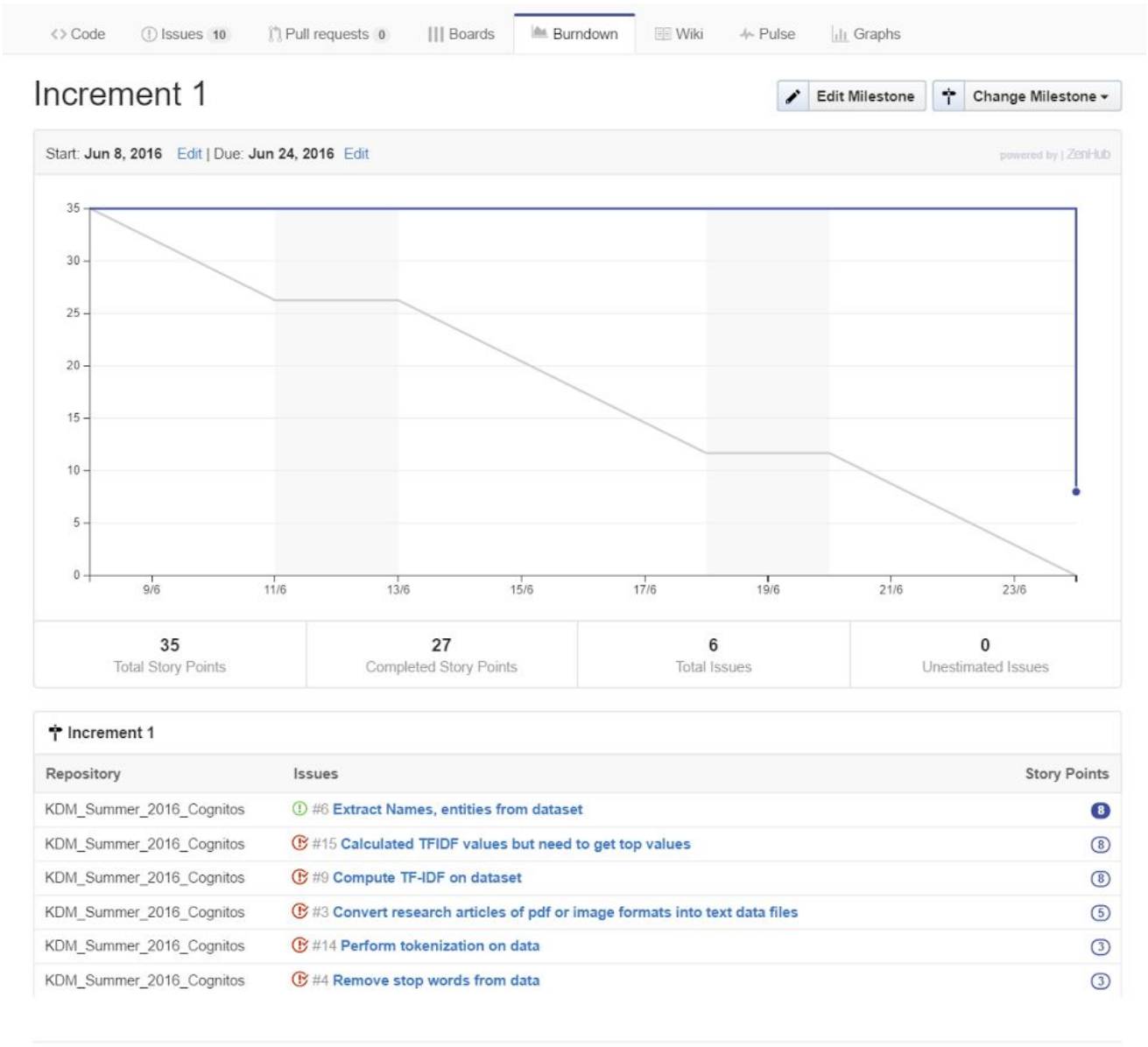
Team members contribution:



Zenhub Boards:



Burndown Chart:



C. Concerns or Issues:

- The main issue we faced in developing the system was that we are unable to access entire research article datasets. Metadata of research articles is available for access, but the actual content of a research article is not accessible. Microsoft Academic Search API used to provide a good interface for such data, but it is deprecated.

- Another issue we faced is while performing optical character recognition. Tesseract provides a good interface for performing OCR, but it uses up most of our computing resources. So, we chose Document Conversion API in Watson Developer Cloud Java SDK, running OCR on IBM Bluemix, thereby transferring the computational load to the cloud.

- In a knowledge graph, relationships between entities play a key role in understanding the gist of the concept discussed in a paper. In our case, after removing stop words from the data, we are performing lemmatization. We observed that this resulted in losing some important relationships between key terms. Since TextRunner values verb forms, it will not deliver sensible results on lemmatized dataset. We are mulling over this issue and trying with different articles to assess the impact of lemmatization on results generated by TextRunner.

D. Future Work:

- The system is designed to be scalable, but never been tested for large datasets due to lack of data and computing resources.
- Also, our knowledge graph can be further extended to represent different sub-fields in a field of research.

REFERENCES:

[1]. http://apache.org/
[2]. http://spark.apache.org/docs/latest/mllib-guide.html
[3]. https://spark.apache.org/docs/1.3.1/api/java/org/apache/spark/rdd/RDD.html
[4]. https://spark.apache.org/docs/1.5.1/api/java/org/apache/spark/sql/DataFrame.html
[5]. https://spark.apache.org/docs/1.1.0/mllib-feature-extraction.html
[6]. https://github.com/watson-developer-cloud
[7]. https://github.com/stanfordnlp/CoreNLP
[8]. Open Information Extraction from the Web Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, Department of Computer Science and Engineering University of Washington.