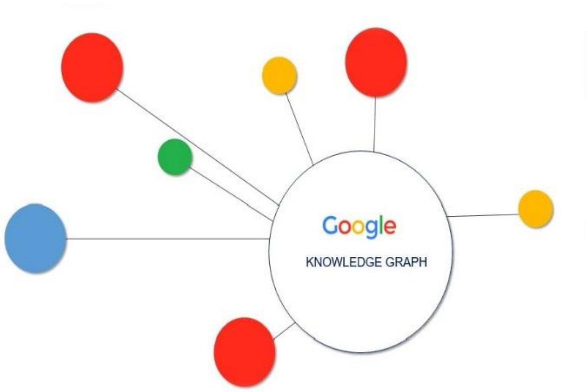
Knowledge Discovery and Management

Project Report



TechChamps

<u>Team 1</u>

1. Jakkepalli Rama Charan Pavan (8)
2. Puthana Sujitha (24)
3. Yalamanchili Sowmya (30)
4. Nandanamudi Sreelakshmi (17)

1.Project Motivation, Objectives, and Significance

1.1 Motivation:

Data Science is a system to extract knowledge of data in various forms, either structured or unstructured from various domains, similar to Knowledge Discovery in Databases(KDD). Natural language processing is used for processing the text which is machine understandable and which will help for fast retrieval of data.

1.2 Specific Objectives:

1.2.1 Easy search of information from huge amount of text.
1.2.2 Helps in precise answer for customized questions.
1.2.3 Increase the knowledge management process.

1.3 Specific Significance:

This application helps in fetching the answer to particular questions by using NLP Process, word2vec, TF-IDF, N-gram. NLP is useful step for text processing and then we are extracting the relevant data.

2. Domain and Q/A application

We are taking News as our domain for our project and applying NLP operations on it and further applying question answering system for the dataset. For this question answering system, we are considering two datasets from News domain.

3. Specific Datasets

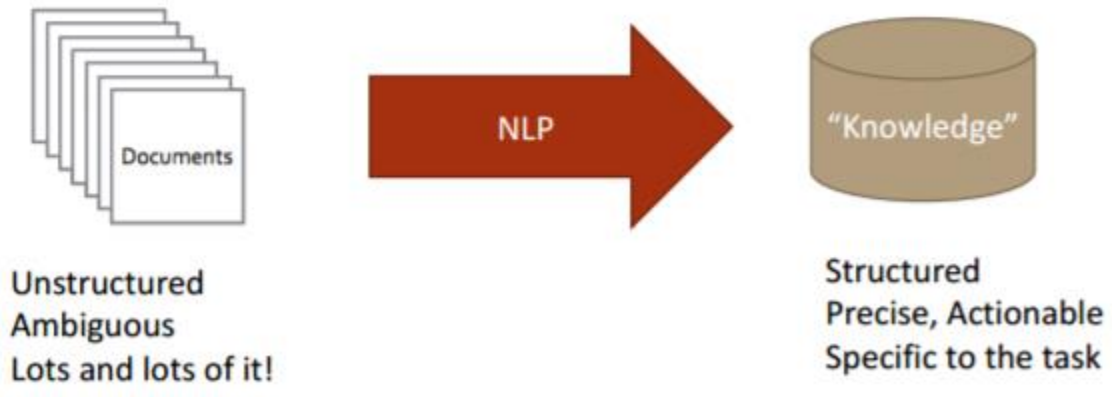
For our project implementation, we have considered two datasets as follows:

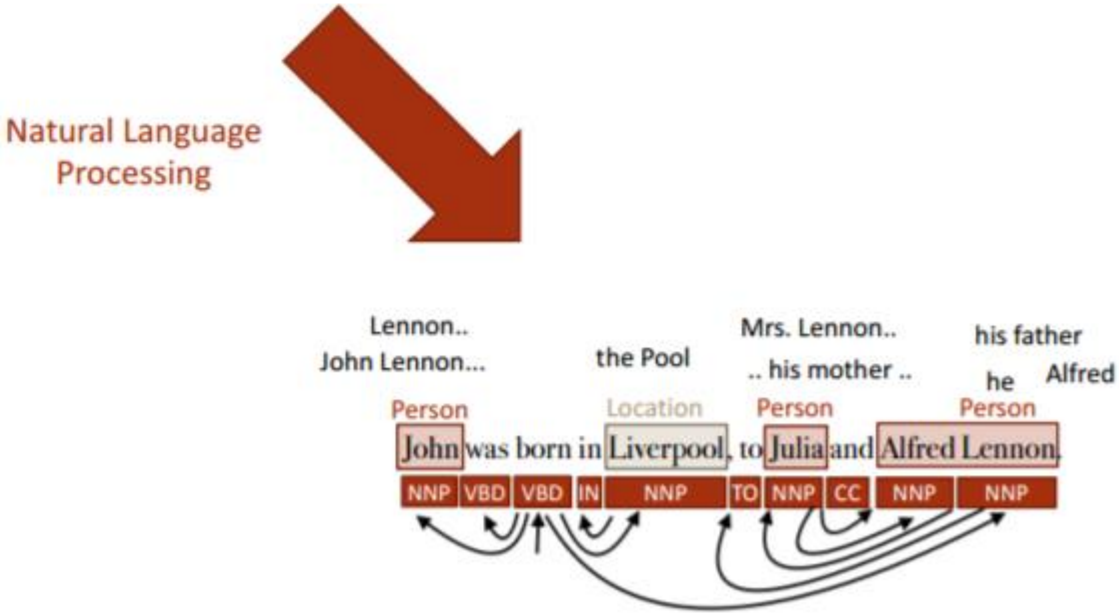
- WikiRef220
- BBC News-In this especially we have selected politics area and sports.

4. Design

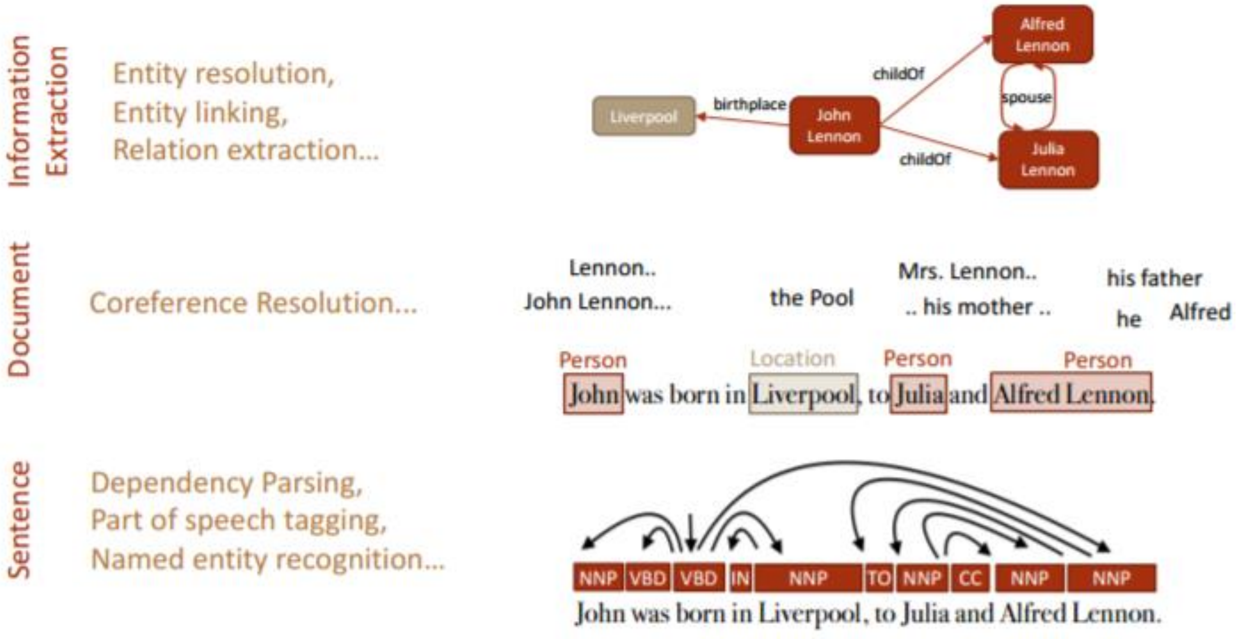
4.1 Workflow

Step 1: Natural language processing – This process includes the identification of token, lemmatization, named entity reference(NER), co-reference resolution.




Natural Language Processing
Lennon..
John Lennon...
the Pool
Mrs. Lennon..
.. his mother ..
his father
he
Alfred
Person
Location
Person
Person
John was born in Liverpool, to Julia and Alfred Lennon.
NNP VBD VBD IN NNP TO NNP CC NNP NNP

Step 2: Information Retrieval – Retrieving the information from the text. We are including the identification of the NER i.e., PERSON, LOCATION, ORGANIZATION.

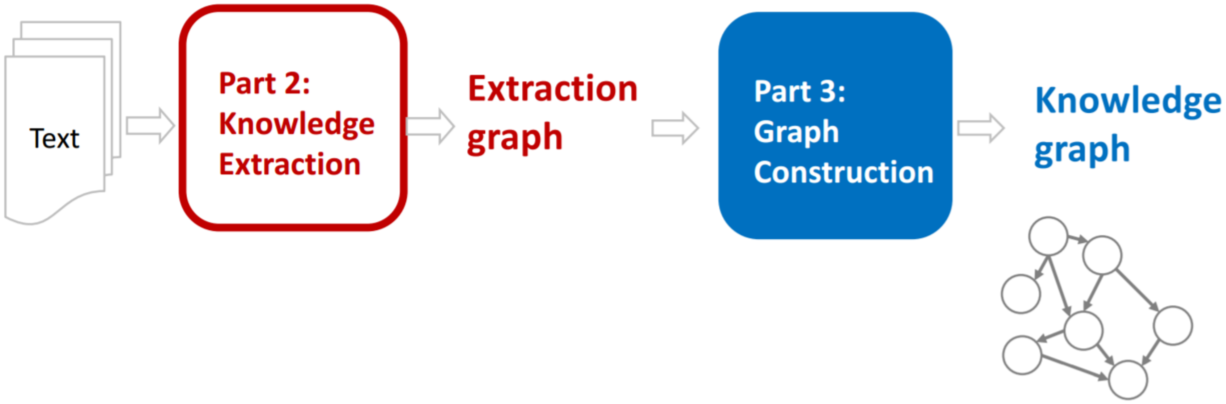


Step 3: Topic Discovery – Topic discovery helps identification of the topics from the context question.

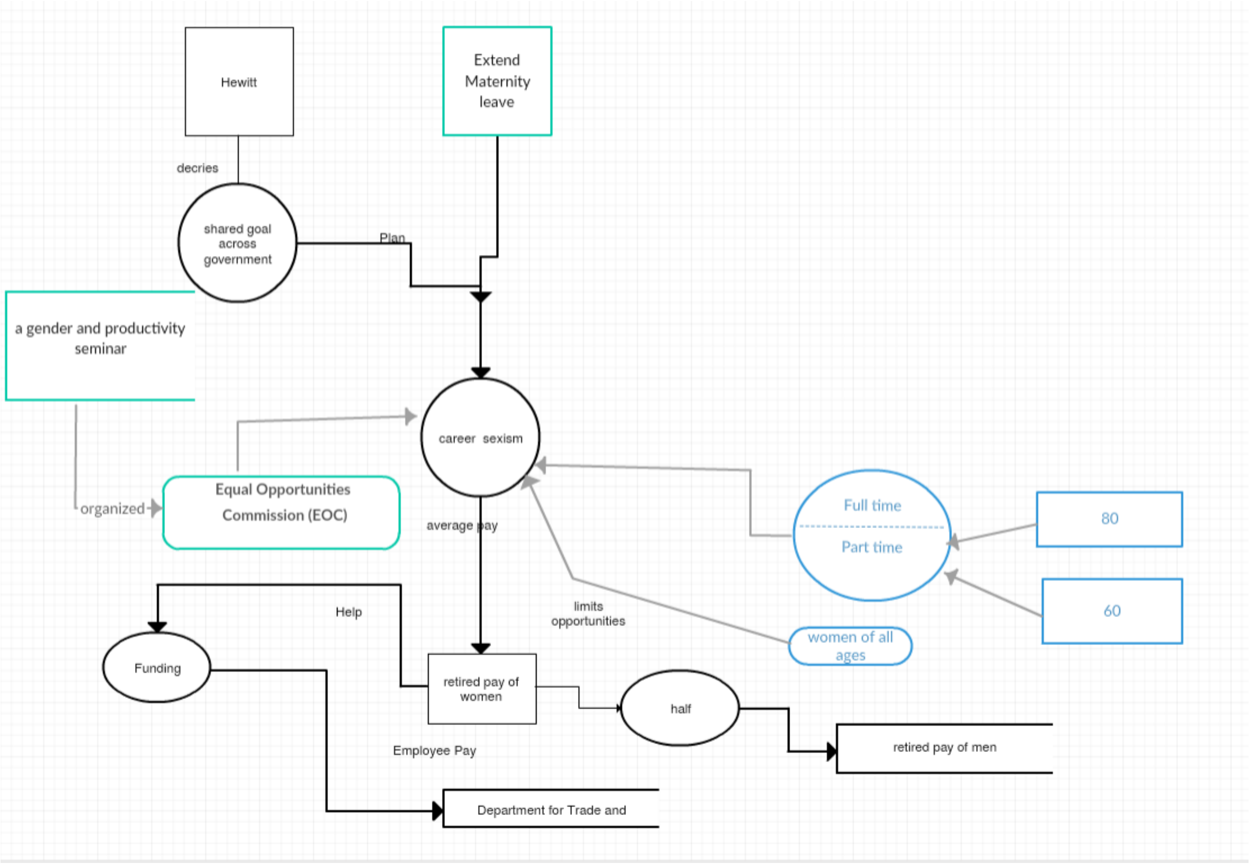
Step 4: Knowledge Graph construction – Construction of the knowledge graph from generated NER.

4.2 Knowledge Graph

4.2.1 Design workflow of knowledge Graph



4.2.2 Knowledge Graph for our dataset:



4.3 A Question-Answer Set for our Dataset.

We are designing the questions from datasets considering mainly the PERSON, LOCATION, ORGANIZATION, NUMBER entity.

1. When was Obama born?
   Born on Aug. 4, 1961.

2. Where did Obama did his schooling?
   Punahou School.

3. Who is father of Obama?
   Barack Hussein Obama.

4. Whom did Obama compete in primary race?
   Hillary Rodham Clinton.

5. What is the minimum duration for maternity leave?
   6 months.

6. What is the topic about?
   career sexism.

7. Who is the speaker?
   Ms. Hewitt.

8. What is the average pay for full-time women.
   80p

9. What is the average pay for part-time women.
   60p.

10. What is the average pay for retired women compared to men?
    Half.

5. Implementation

5.1 Output of NLP operations for our dataset

We have performed the NLP operations on the dataset which we have chosen and the result of each operation is shown in the below mentioned screenshots.

Tokenization:

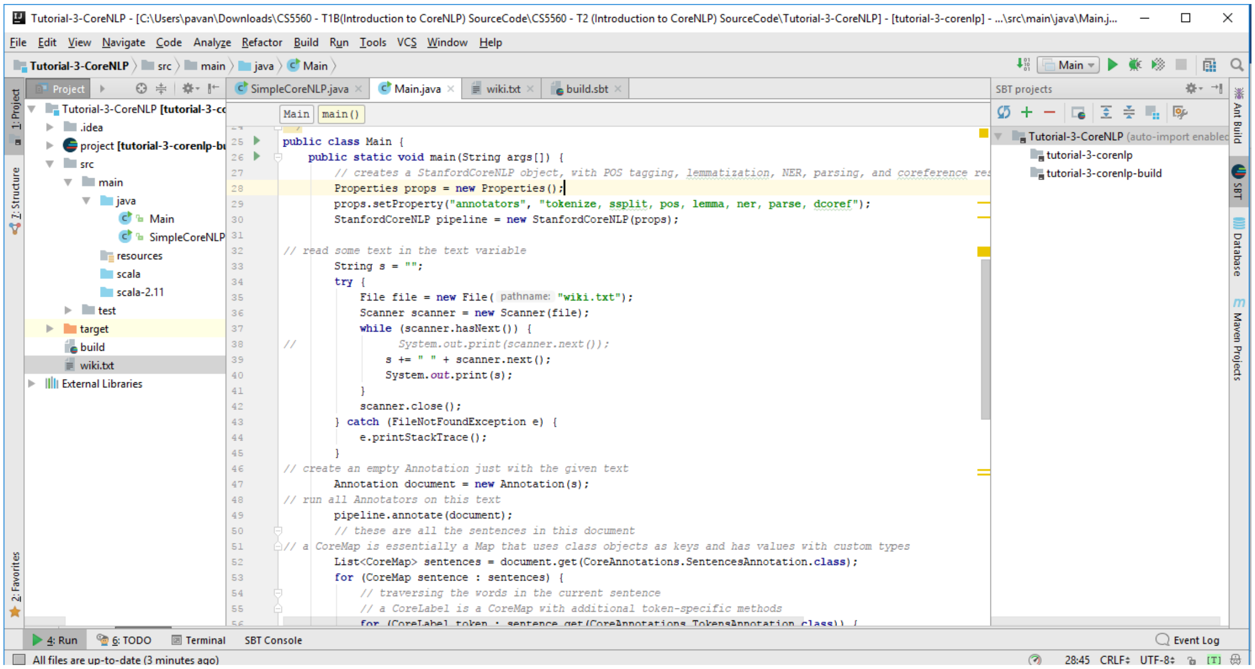
Lemmatization:

POS Tagging:

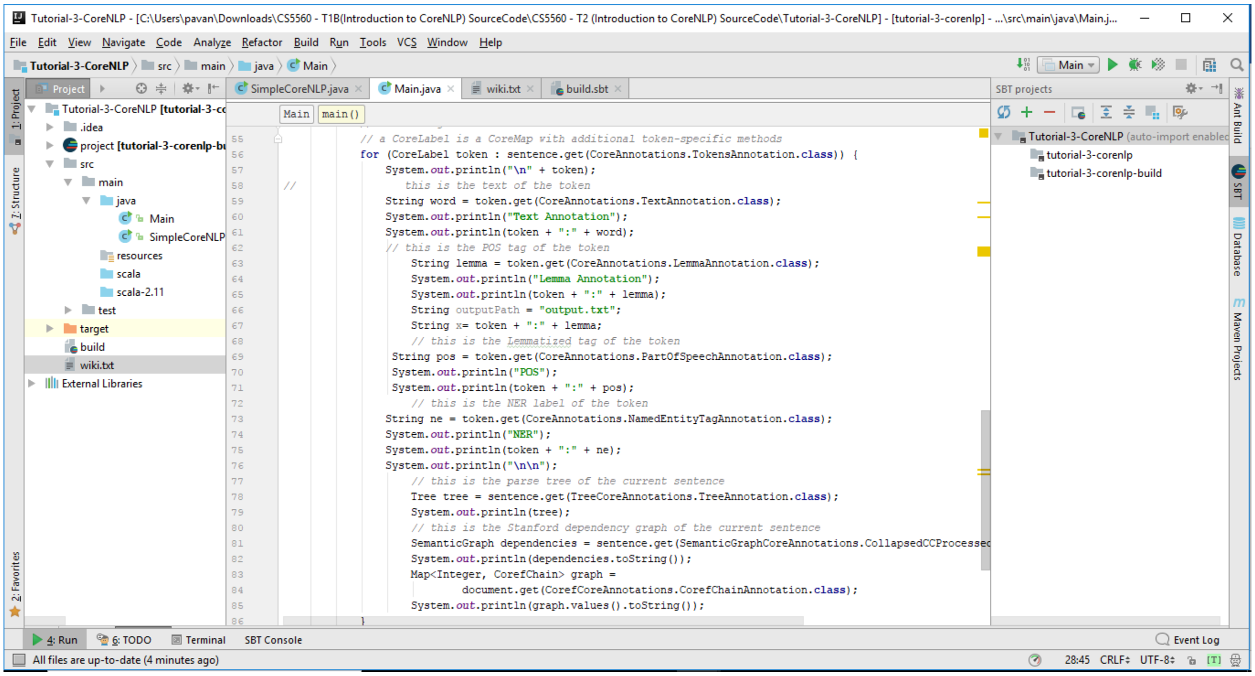
NER:

Coreference Resolution:

Below is the code for all the operations of the NLP.

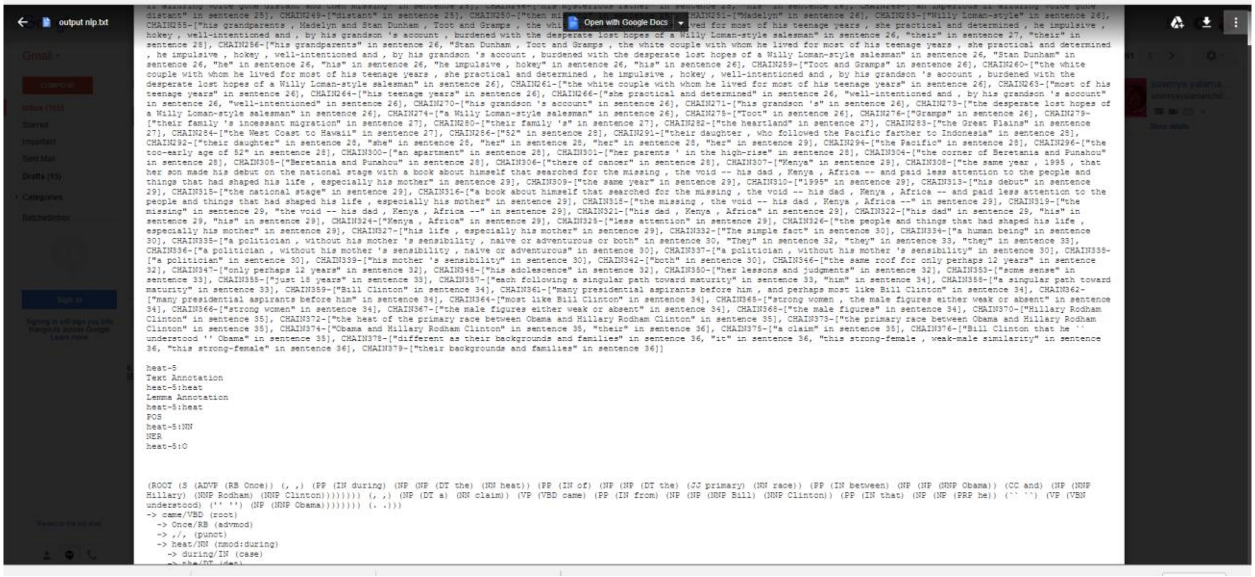

Tutorial-3-CoreNLP - [C:\Users\pavan\Downloads\CS5560 - T1B(Introduction to CoreNLP) SourceCode\CS5560 - T2 (Introduction to CoreNLP) SourceCode\Tutorial-3-CoreNLP] - [tutorial-3-corenlp] - ...\src\main\java\Main.j...
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Tutorial-3-CoreNLP src main java Main
Main
Project
Tutorial-3-CoreNLP [tutorial-3-co
.idea
project [tutorial-3-corenlp-bu
src
main
java
Main
SimpleCoreNLP
resources
scala
scala-2.11
test
target
build
wiki.txt
External Libraries
SimpleCoreNLP.java
Main.java
wiki.txt
build.sbt
Main
main()
SBT projects
Tutorial-3-CoreNLP (auto-import enabled
tutorial-3-corenlp
tutorial-3-corenlp-build
1: Project
7: Structure
2: Favorites
Ant Build
SBT
Database
Maven Projects
4: Run
6: TODO
Terminal
SBT Console
Event Log
All files are up-to-date (3 minutes ago)
28:45 CRLF÷ UTF-8÷

```java
public class Main {
    public static void main(String args[]) {
        // creates a StanfordCoreNLP object, with POS tagging, lemmatization, NER, parsing, and coreference res
        Properties props = new Properties();
        props.setProperty("annotators", "tokenize, ssplit, pos, lemma, ner, parse, dcoref");
        StanfordCoreNLP pipeline = new StanfordCoreNLP(props);

// read some text in the text variable
        String s = "";
        try {
            File file = new File( pathname: "wiki.txt");
            Scanner scanner = new Scanner(file);
            while (scanner.hasNext()) {
//                System.out.print(scanner.next());
                s += " " + scanner.next();
                System.out.print(s);
            }
            scanner.close();
        } catch (FileNotFoundException e) {
            e.printStackTrace();
        }
// create an empty Annotation just with the given text
        Annotation document = new Annotation(s);
// run all Annotators on this text
        pipeline.annotate(document);
        // these are all the sentences in this document
// a CoreMap is essentially a Map that uses class objects as keys and has values with custom types
        List<CoreMap> sentences = document.get(CoreAnnotations.SentencesAnnotation.class);
        for (CoreMap sentence : sentences) {
            // traversing the words in the current sentence
            // a CoreLabel is a CoreMap with additional token-specific methods
            for (CoreLabel token : sentence.get(CoreAnnotations.TokensAnnotation.class)) {
```

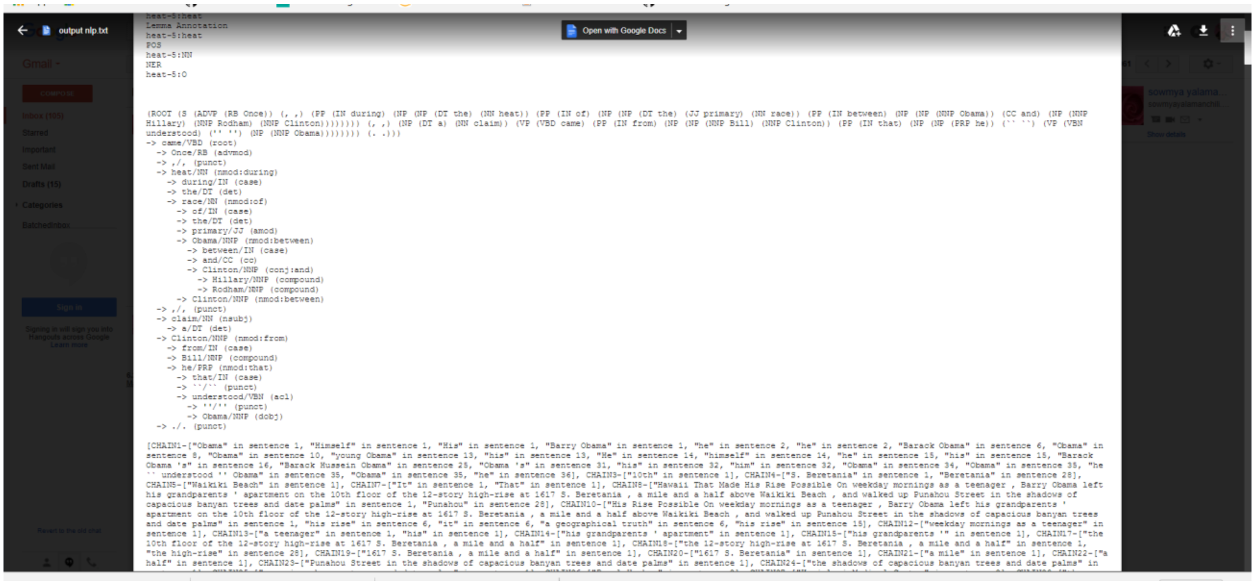

Tutorial-3-CoreNLP - [C:\Users\pavan\Downloads\CS5560 - T1B(Introduction to CoreNLP) SourceCode\CS5560 - T2 (Introduction to CoreNLP) SourceCode\Tutorial-3-CoreNLP] - [tutorial-3-corenlp] - ...\src\main\java\Main.j...
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Tutorial-3-CoreNLP src main java Main
Main
Project
SimpleCoreNLP.java
Main.java
wiki.txt
build.sbt
SBT projects
Tutorial-3-CoreNLP [tutorial-3-cc
.idea
project [tutorial-3-corenlp-bu
src
main
java
Main
SimpleCoreNLP
resources
scala
scala-2.11
test
target
build
wiki.txt
External Libraries
Main
main()
Tutorial-3-CoreNLP (auto-import enabled
tutorial-3-corenlp
tutorial-3-corenlp-build
1: Project
7: Structure
2: Favorites
Ant Build
SBT
Database
Maven Projects
4: Run
6: TODO
Terminal
SBT Console
Event Log
All files are up-to-date (4 minutes ago)
28:45 CRLF: UTF-8:

```
            // a CoreLabel is a CoreMap with additional token-specific methods
            for (CoreLabel token : sentence.get(CoreAnnotations.TokensAnnotation.class)) {
                System.out.println("\n" + token);
//                 this is the text of the token
                String word = token.get(CoreAnnotations.TextAnnotation.class);
                System.out.println("Text Annotation");
                System.out.println(token + ":" + word);
                // this is the POS tag of the token
                    String lemma = token.get(CoreAnnotations.LemmaAnnotation.class);
                    System.out.println("Lemma Annotation");
                    System.out.println(token + ":" + lemma);
                    String outputPath = "output.txt";
                    String x= token + ":" + lemma;
                    // this is the Lemmatized tag of the token
                 String pos = token.get(CoreAnnotations.PartOfSpeechAnnotation.class);
                 System.out.println("POS");
                 System.out.println(token + ":" + pos);
                    // this is the NER label of the token
                String ne = token.get(CoreAnnotations.NamedEntityTagAnnotation.class);
                System.out.println("NER");
                System.out.println(token + ":" + ne);
                System.out.println("\n\n");
                    // this is the parse tree of the current sentence
                    Tree tree = sentence.get(TreeCoreAnnotations.TreeAnnotation.class);
                    System.out.println(tree);
                    // this is the Stanford dependency graph of the current sentence
                    SemanticGraph dependencies = sentence.get(SemanticGraphCoreAnnotations.CollapsedCCProcesse
                    System.out.println(dependencies.toString());
                    Map<Integer, CorefChain> graph =
                            document.get(CorefCoreAnnotations.CorefChainAnnotation.class);
                    System.out.println(graph.values().toString());
            }
```

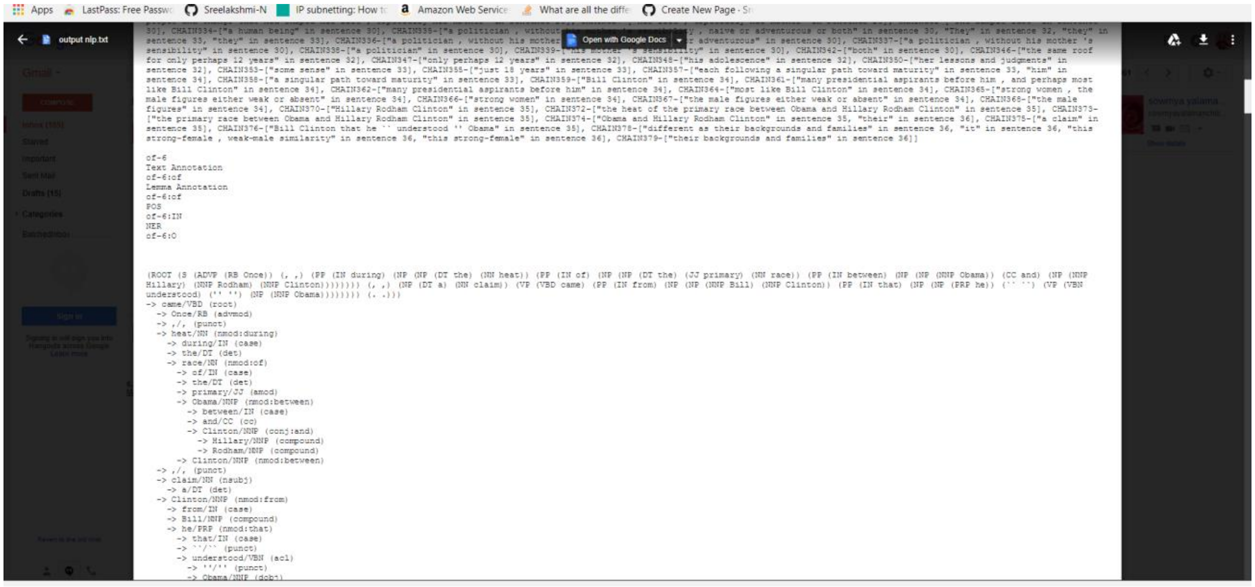
Below are the outputs of the operations.



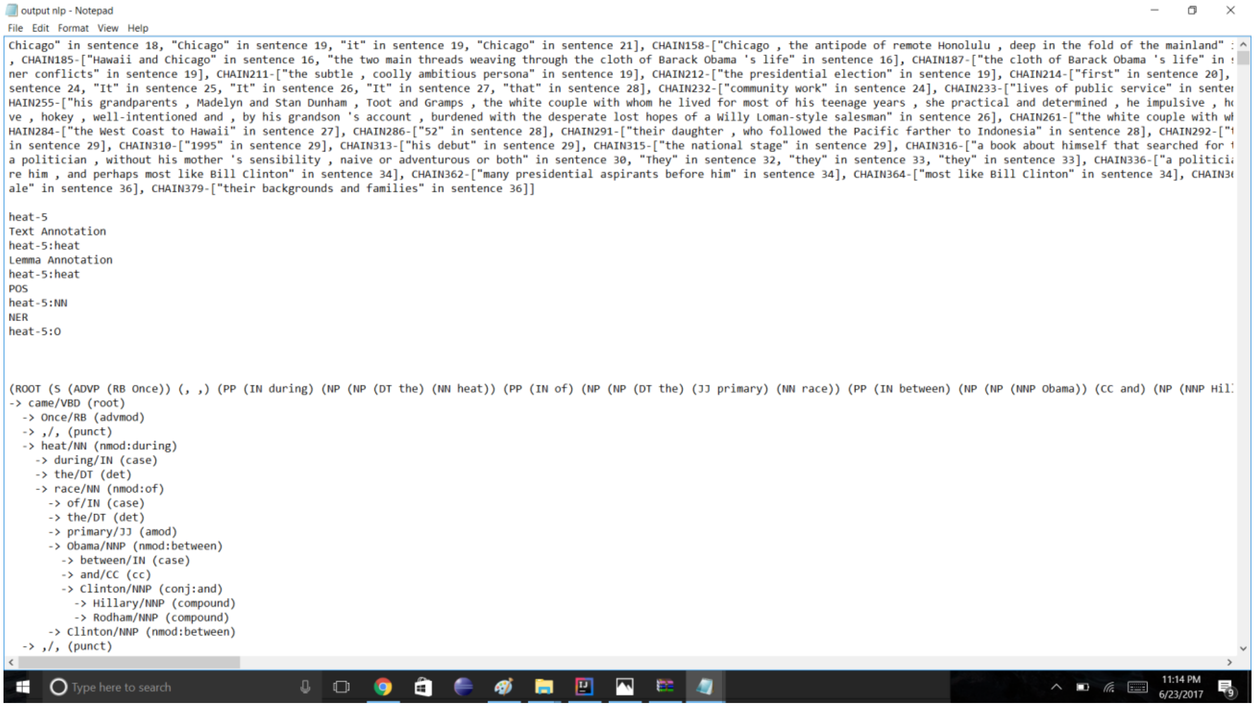
```
is alluring voice gone distant and then missing" in sentence 25], CHAIN248-["his mysterious father" in sentence 25, "his" in sentence 26], CHAIN248-["an imperious if alluring voice gone
distant" in sentence 25], CHAIN249-["distant" in sentence 25], CHAIN250-["then mi HAIN251-["Madelyn" in sentence 26], CHAIN253-["Willy Loman-style" in sentence 26],
CHAIN255-["his grandparents , Madelyn and Stan Dunham , Toot and Gramps , the whi ved for most of his teenage years , she practical and determined , he impulsive ,
hokey , well-intentioned and , by his grandson 's account , burdened with the desperate lost hopes of a Willy Loman-style salesman" in sentence 26, "their" in sentence 27, "their" in
sentence 28], CHAIN256-["his grandparents" in sentence 26, "Stan Dunham , Toot and Gramps , the white couple with whom he lived for most of his teenage years , she practical and determined
, he impulsive , hokey , well-intentioned and , by his grandson 's account , burdened with the desperate lost hopes of a Willy Loman-style salesman" in sentence 26, "Stan Dunham" in
sentence 26, "he" in sentence 26, "his" in sentence 26, "he impulsive , hokey" in sentence 26, "his" in sentence 26], CHAIN259-["Toot and Gramps" in sentence 26], CHAIN260-["the white
couple with whom he lived for most of his teenage years , she practical and determined , he impulsive , hokey , well-intentioned and , by his grandson 's account , burdened with the
desperate lost hopes of a Willy Loman-style salesman" in sentence 26], CHAIN261-["the white couple with whom he lived for most of his teenage years" in sentence 26], CHAIN263-["most of his
teenage years" in sentence 26], CHAIN264-["his teenage years" in sentence 26], CHAIN266-["she practical and determined" in sentence 26, "well-intentioned and , by his grandson 's account"
in sentence 26, "well-intentioned" in sentence 26], CHAIN270-["his grandson 's account" in sentence 26], CHAIN271-["his grandson 's" in sentence 26], CHAIN273-["the desperate lost hopes of
a Willy Loman-style salesman" in sentence 26], CHAIN274-["a Willy Loman-style salesman" in sentence 26], CHAIN275-["Toot" in sentence 26], CHAIN276-["Gramps" in sentence 26], CHAIN279-
["their family 's incessant migration" in sentence 27], CHAIN280-["their family 's" in sentence 27], CHAIN282-["the heartland" in sentence 27], CHAIN283-["the Great Plains" in sentence
27], CHAIN284-["the West Coast to Hawaii" in sentence 27], CHAIN286-["52" in sentence 28], CHAIN291-["their daughter , who followed the Pacific farther to Indonesia" in sentence 28],
CHAIN292-["their daughter" in sentence 28, "she" in sentence 28, "her" in sentence 28, "her" in sentence 28, "her" in sentence 29], CHAIN294-["the Pacific" in sentence 28], CHAIN296-["the
too-early age of 52" in sentence 28], CHAIN300-["an apartment" in sentence 28], CHAIN301-["her parents ' in the high-rise" in sentence 28], CHAIN304-["the corner of Beretania and Punahou"
in sentence 28], CHAIN305-["Beretania and Punahou" in sentence 28], CHAIN306-["there of cancer" in sentence 28], CHAIN307-["Kenya" in sentence 29], CHAIN308-["the same year , 1995 , that
her son made his debut on the national stage with a book about himself that searched for the missing , the void -- his dad , Kenya , Africa -- and paid less attention to the people and
things that had shaped his life , especially his mother" in sentence 29], CHAIN309-["the same year" in sentence 29], CHAIN310-["1995" in sentence 29], CHAIN313-["his debut" in sentence
29], CHAIN315-["the national stage" in sentence 29], CHAIN316-["a book about himself that searched for the missing , the void -- his dad , Kenya , Africa -- and paid less attention to the
people and things that had shaped his life , especially his mother" in sentence 29], CHAIN318-["the missing , the void -- his dad , Kenya , Africa --" in sentence 29], CHAIN319-["the
missing" in sentence 29, "the void -- his dad , Kenya , Africa --" in sentence 29], CHAIN321-["his dad , Kenya , Africa" in sentence 29], CHAIN322-["his dad" in sentence 29, "his" in
sentence 29, "his" in sentence 29], CHAIN324-["Kenya , Africa" in sentence 29], CHAIN325-["less attention" in sentence 29], CHAIN326-["the people and things that had shaped his life ,
especially his mother" in sentence 29], CHAIN327-["his life , especially his mother" in sentence 29], CHAIN332-["The simple fact" in sentence 30], CHAIN334-["a human being" in sentence
30], CHAIN335-["a politician , without his mother 's sensibility , naive or adventurous or both" in sentence 30, "They" in sentence 32, "they" in sentence 33, "they" in sentence 33],
CHAIN336-["a politician , without his mother 's sensibility , naive or adventurous" in sentence 30], CHAIN337-["a politician , without his mother 's sensibility" in sentence 30], CHAIN338-
["a politician" in sentence 30], CHAIN339-["his mother 's sensibility" in sentence 30], CHAIN342-["both" in sentence 30], CHAIN346-["the same roof for only perhaps 12 years" in sentence
32], CHAIN347-["only perhaps 12 years" in sentence 32], CHAIN348-["his adolescence" in sentence 32], CHAIN350-["her lessons and judgments" in sentence 32], CHAIN353-["some sense" in
sentence 33], CHAIN355-["just 18 years" in sentence 33], CHAIN357-["each following a singular path toward maturity" in sentence 33, "him" in sentence 34], CHAIN358-["a singular path toward
maturity" in sentence 33], CHAIN359-["Bill Clinton" in sentence 34], CHAIN361-["many presidential aspirants before him , and perhaps most like Bill Clinton" in sentence 34], CHAIN362-
["many presidential aspirants before him" in sentence 34], CHAIN364-["most like Bill Clinton" in sentence 34], CHAIN365-["strong women , the male figures either weak or absent" in sentence
34], CHAIN366-["strong women" in sentence 34], CHAIN367-["the male figures either weak or absent" in sentence 34], CHAIN368-["the male figures" in sentence 34], CHAIN370-["Hillary Rodham
Clinton" in sentence 35], CHAIN372-["the heat of the primary race between Obama and Hillary Rodham Clinton" in sentence 35], CHAIN373-["the primary race between Obama and Hillary Rodham
Clinton" in sentence 35], CHAIN374-["Obama and Hillary Rodham Clinton" in sentence 35, "their" in sentence 36], CHAIN375-["a claim" in sentence 35], CHAIN376-["Bill Clinton that he ``
understood '' Obama" in sentence 35], CHAIN378-["different as their backgrounds and families" in sentence 36, "it" in sentence 36, "this strong-female , weak-male similarity" in sentence
36, "this strong-female" in sentence 36], CHAIN379-["their backgrounds and families" in sentence 36]]

heat-5
Text Annotation
heat-5:heat
Lemma Annotation
heat-5:heat
POS
heat-5:NN
NER
heat-5:O



(ROOT (S (ADVP (RB Once)) (, ,) (PP (IN during) (NP (NP (DT the) (NN heat)) (PP (IN of) (NP (NP (DT the) (JJ primary) (NN race)) (PP (IN between) (NP (NP (NNP Obama)) (CC and) (NP (NNP
Hillary) (NNP Rodham) (NNP Clinton)))))))) (, ,) (NP (DT a) (NN claim)) (VP (VBD came) (PP (IN from) (NP (NP (NNP Bill) (NNP Clinton)) (PP (IN that) (NP (NP (PRP he)) (`` ``) (VP (VBN
understood) ('' '') (NP (NNP Obama)))))))) (. .)))
-> came/VBD (root)
  -> Once/RB (advmod)
  -> ,/, (punct)
  -> heat/NN (nmod:during)
    -> during/IN (case)
    -> the/DT (det)
```



```
heat-5:heat
Lemma Annotation
heat-5:heat
POS
heat-5:NN
NER
heat-5:O


(ROOT (S (ADVP (RB Once)) (, ,) (PP (IN during) (NP (NP (DT the) (NN heat)) (PP (IN of) (NP (NP (DT the) (JJ primary) (NN race)) (PP (IN between) (NP (NP (NNP Obama)) (CC and) (NP (NNP
Hillary) (NNP Rodham) (NNP Clinton)))))))) (, ,) (NP (DT a) (NN claim)) (VP (VBD came) (PP (IN from) (NP (NP (NNP Bill) (NNP Clinton)) (PP (IN that) (NP (NP (PRP he)) (`` ``) (VP (VBN
understood) ('' '') (NP (NNP Obama)))))))) (. .)))
-> came/VBD (root)
  -> Once/RB (advmod)
  -> ,/, (punct)
  -> heat/NN (nmod:during)
    -> during/IN (case)
    -> the/DT (det)
    -> race/NN (nmod:of)
      -> of/IN (case)
      -> the/DT (det)
      -> primary/JJ (amod)
      -> Obama/NNP (nmod:between)
        -> between/IN (case)
        -> and/CC (cc)
        -> Clinton/NNP (conj:and)
          -> Hillary/NNP (compound)
          -> Rodham/NNP (compound)
      -> Clinton/NNP (nmod:between)
  -> ,/, (punct)
  -> claim/NN (nsubj)
    -> a/DT (det)
  -> Clinton/NNP (nmod:from)
    -> from/IN (case)
    -> Bill/NNP (compound)
    -> he/PRP (nmod:that)
      -> that/IN (case)
      -> ``/`` (punct)
      -> understood/VBN (acl)
        -> ''/'' (punct)
        -> Obama/NNP (dobj)
  -> ./. (punct)

[CHAIN1-["Obama" in sentence 1, "Himself" in sentence 1, "His" in sentence 1, "Barry Obama" in sentence 1, "he" in sentence 2, "he" in sentence 2, "Barack Obama" in sentence 6, "Obama" in
sentence 8, "Obama" in sentence 10, "young Obama" in sentence 13, "his" in sentence 13, "He" in sentence 14, "himself" in sentence 14, "he" in sentence 15, "his" in sentence 15, "Barack
Obama 's" in sentence 16, "Barack Hussein Obama" in sentence 25, "Obama 's" in sentence 31, "his" in sentence 32, "him" in sentence 32, "Obama" in sentence 34, "Obama" in sentence 35, "he
`` understood '' Obama" in sentence 35, "Obama" in sentence 35, "he" in sentence 36], CHAIN3-["10th" in sentence 1], CHAIN4-["S. Beretania" in sentence 1, "Beretania" in sentence 28],
CHAIN5-["Waikiki Beach" in sentence 1], CHAIN7-["It" in sentence 1, "That" in sentence 1], CHAIN8-["Hawaii That Made His Rise Possible On weekday mornings as a teenager , Barry Obama left
his grandparents ' apartment on the 10th floor of the 12-story high-rise at 1617 S. Beretania , a mile and a half above Waikiki Beach , and walked up Punahou Street in the shadows of
capacious banyan trees and date palms" in sentence 1, "Punahou" in sentence 28], CHAIN10-["His Rise Possible On weekday mornings as a teenager , Barry Obama left his grandparents '
apartment on the 10th floor of the 12-story high-rise at 1617 S. Beretania , a mile and a half above Waikiki Beach , and walked up Punahou Street in the shadows of capacious banyan trees
and date palms" in sentence 1, "his rise" in sentence 6, "it" in sentence 6, "a geographical truth" in sentence 6, "his rise" in sentence 15], CHAIN12-["weekday mornings as a teenager" in
sentence 1], CHAIN13-["a teenager" in sentence 1, "his" in sentence 1], CHAIN14-["his grandparents ' apartment" in sentence 1], CHAIN15-["his grandparents '" in sentence 1], CHAIN17-["the
10th floor of the 12-story high-rise at 1617 S. Beretania , a mile and a half" in sentence 1], CHAIN18-["the 12-story high-rise at 1617 S. Beretania , a mile and a half" in sentence 1,
"the high-rise" in sentence 28], CHAIN19-["1617 S. Beretania , a mile and a half" in sentence 1], CHAIN20-["1617 S. Beretania" in sentence 1], CHAIN21-["a mile" in sentence 1], CHAIN22-["a
half" in sentence 1], CHAIN23-["Punahou Street in the shadows of capacious banyan trees and date palms" in sentence 1], CHAIN24-["the shadows of capacious banyan trees and date palms" in
```



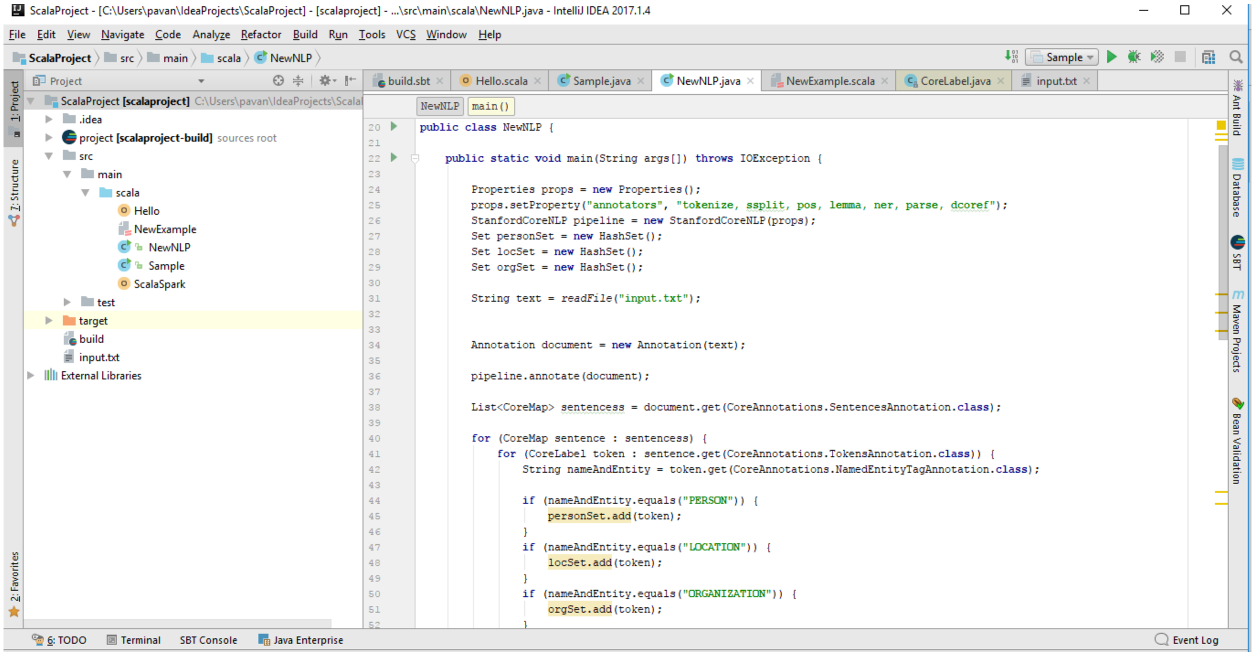
```
30], CHAIN334-["a human being" in sentence 30], CHAIN335-["a politician , without [illegible] , naive or adventurous or both" in sentence 30, "They" in sentence 32, "they" in
sentence 33, "they" in sentence 33], CHAIN336-["a politician , without his mother [illegible] adventurous" in sentence 30], CHAIN337-["a politician , without his mother 's
sensibility" in sentence 30], CHAIN338-["a politician" in sentence 30], CHAIN339-["his mother 's sensibility" in sentence 30], CHAIN342-["both" in sentence 30], CHAIN346-["the same roof
for only perhaps 12 years" in sentence 32], CHAIN347-["only perhaps 12 years" in sentence 32], CHAIN348-["his adolescence" in sentence 32], CHAIN350-["her lessons and judgments" in
sentence 32], CHAIN353-["some sense" in sentence 33], CHAIN355-["just 18 years" in sentence 33], CHAIN357-["each following a singular path toward maturity" in sentence 33, "him" in
sentence 34], CHAIN358-["a singular path toward maturity" in sentence 33], CHAIN359-["Bill Clinton" in sentence 34], CHAIN361-["many presidential aspirants before him , and perhaps most
like Bill Clinton" in sentence 34], CHAIN362-["many presidential aspirants before him" in sentence 34], CHAIN364-["most like Bill Clinton" in sentence 34], CHAIN365-["strong women , the
male figures either weak or absent" in sentence 34], CHAIN366-["strong women" in sentence 34], CHAIN367-["the male figures either weak or absent" in sentence 34], CHAIN368-["the male
figures" in sentence 34], CHAIN370-["Hillary Rodham Clinton" in sentence 35], CHAIN372-["the heat of the primary race between Obama and Hillary Rodham Clinton" in sentence 35], CHAIN373-
["the primary race between Obama and Hillary Rodham Clinton" in sentence 35], CHAIN374-["Obama and Hillary Rodham Clinton" in sentence 35, "their" in sentence 36], CHAIN375-["a claim" in
sentence 35], CHAIN376-["Bill Clinton that he `` understood '' Obama" in sentence 35], CHAIN378-["different as their backgrounds and families" in sentence 36, "it" in sentence 36, "this
strong-female , weak-male similarity" in sentence 36, "this strong-female" in sentence 36], CHAIN379-["their backgrounds and families" in sentence 36]]

of-6
Text Annotation
of-6:of
Lemma Annotation
of-6:of
POS
of-6:IN
NER
of-6:O


(ROOT (S (ADVP (RB Once)) (, ,) (PP (IN during) (NP (NP (DT the) (NN heat)) (PP (IN of) (NP (NP (DT the) (JJ primary) (NN race)) (PP (IN between) (NP (NP (NNP Obama)) (CC and) (NP (NNP
Hillary) (NNP Rodham) (NNP Clinton)))))))) (, ,) (NP (DT a) (NN claim)) (VP (VBD came) (PP (IN from) (NP (NP (NNP Bill) (NNP Clinton)) (PP (IN that) (NP (NP (PRP he)) (`` ``) (VP (VBN
understood) ('' '') (NP (NNP Obama)))))))) (. .)))
-> came/VBD (root)
  -> Once/RB (advmod)
  -> ,/, (punct)
  -> heat/NN (nmod:during)
    -> during/IN (case)
    -> the/DT (det)
    -> race/NN (nmod:of)
      -> of/IN (case)
      -> the/DT (det)
      -> primary/JJ (amod)
      -> Obama/NNP (nmod:between)
        -> between/IN (case)
        -> and/CC (cc)
        -> Clinton/NNP (conj:and)
          -> Hillary/NNP (compound)
          -> Rodham/NNP (compound)
      -> Clinton/NNP (nmod:between)
  -> ,/, (punct)
  -> claim/NN (nsubj)
    -> a/DT (det)
  -> Clinton/NNP (nmod:from)
    -> from/IN (case)
    -> Bill/NNP (compound)
    -> he/PRP (nmod:that)
      -> that/IN (case)
      -> ``/`` (punct)
      -> understood/VBN (acl)
        -> ''/'' (punct)
        -> Obama/NNP (dobj)
```



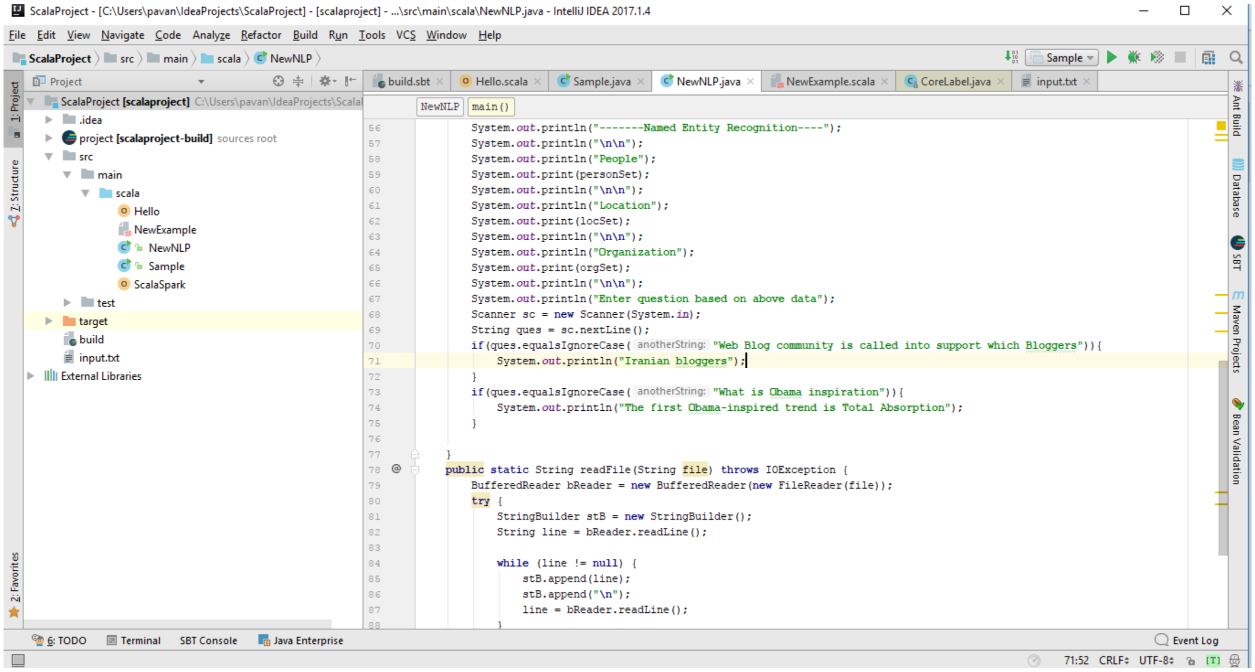
```
Chicago" in sentence 18, "Chicago" in sentence 19, "it" in sentence 19, "Chicago" in sentence 21], CHAIN158-["Chicago , the antipode of remote Honolulu , deep in the fold of the mainland"
, CHAIN185-["Hawaii and Chicago" in sentence 16, "the two main threads weaving through the cloth of Barack Obama 's life" in sentence 16], CHAIN187-["the cloth of Barack Obama 's life" in
ner conflicts" in sentence 19], CHAIN211-["the subtle , coolly ambitious persona" in sentence 19], CHAIN212-["the presidential election" in sentence 19], CHAIN214-["first" in sentence 20],
sentence 24, "It" in sentence 25, "It" in sentence 26, "It" in sentence 27, "that" in sentence 28], CHAIN232-["community work" in sentence 24], CHAIN233-["lives of public service" in senten
HAIN255-["his grandparents , Madelyn and Stan Dunham , Toot and Gramps , the white couple with whom he lived for most of his teenage years , she practical and determined , he impulsive , h
ve , hokey , well-intentioned and , by his grandson 's account , burdened with the desperate lost hopes of a Willy Loman-style salesman" in sentence 26], CHAIN261-["the white couple with wh
HAIN284-["the West Coast to Hawaii" in sentence 27], CHAIN286-["52" in sentence 28], CHAIN291-["their daughter , who followed the Pacific farther to Indonesia" in sentence 28], CHAIN292-["
in sentence 29], CHAIN310-["1995" in sentence 29], CHAIN313-["his debut" in sentence 29], CHAIN315-["the national stage" in sentence 29], CHAIN316-["a book about himself that searched for
a politician , without his mother 's sensibility , naive or adventurous or both" in sentence 30, "They" in sentence 32, "they" in sentence 33, "they" in sentence 33], CHAIN336-["a politici
re him , and perhaps most like Bill Clinton" in sentence 34], CHAIN362-["many presidential aspirants before him" in sentence 34], CHAIN364-["most like Bill Clinton" in sentence 34], CHAIN3
ale" in sentence 36], CHAIN379-["their backgrounds and families" in sentence 36]]

heat-5
Text Annotation
heat-5:heat
Lemma Annotation
heat-5:heat
POS
heat-5:NN
NER
heat-5:O



(ROOT (S (ADVP (RB Once)) (, ,) (PP (IN during) (NP (NP (DT the) (NN heat)) (PP (IN of) (NP (NP (DT the) (JJ primary) (NN race)) (PP (IN between) (NP (NP (NNP Obama)) (CC and) (NP (NNP Hil
-> came/VBD (root)
  -> Once/RB (advmod)
  -> ,/, (punct)
  -> heat/NN (nmod:during)
    -> during/IN (case)
    -> the/DT (det)
    -> race/NN (nmod:of)
      -> of/IN (case)
      -> the/DT (det)
      -> primary/JJ (amod)
      -> Obama/NNP (nmod:between)
        -> between/IN (case)
        -> and/CC (cc)
        -> Clinton/NNP (conj:and)
          -> Hillary/NNP (compound)
          -> Rodham/NNP (compound)
      -> Clinton/NNP (nmod:between)
  -> ,/, (punct)
```

5.2 Question Answering for our dataset

After performing the NLP operations, we have taken the post processed dataset and we have separately stored the result of NER output like from the NER result we have the all person related entities to one file and similarly we have done for every group and based on that we have generated answers for the questions we choose. The below screenshots will depict the same.

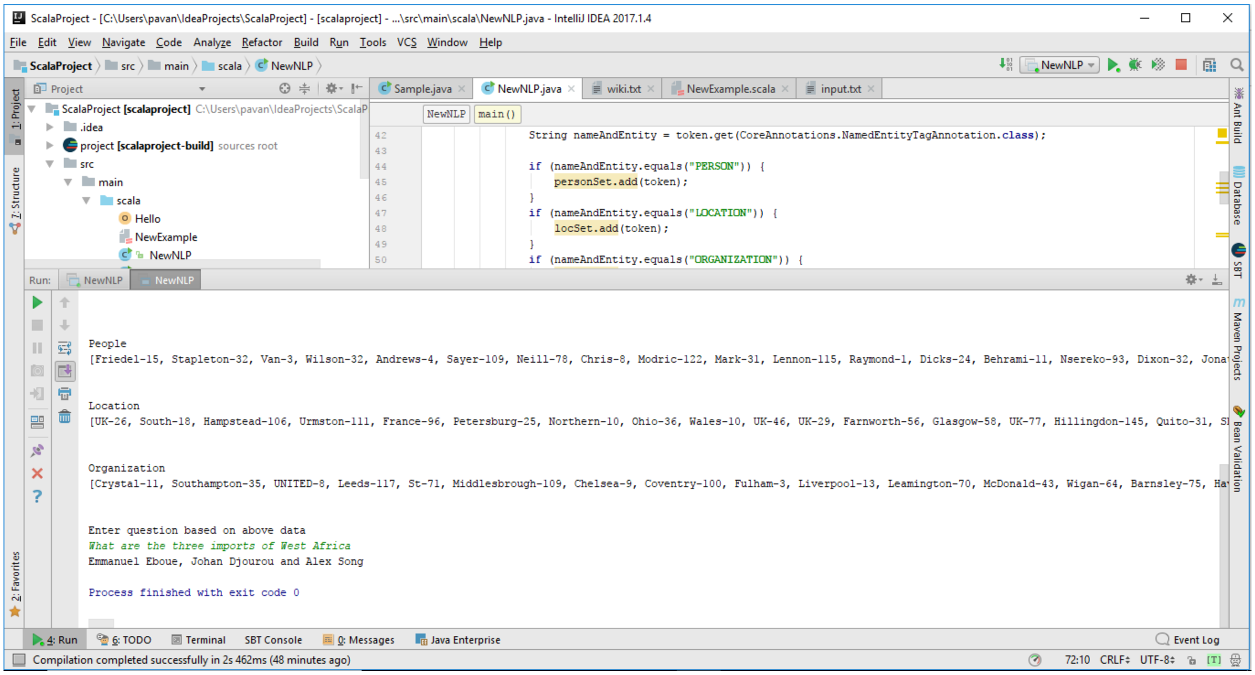

ScalaProject - [C:\Users\pavan\IdeaProjects\ScalaProject] - [scalaproject] - ...\src\main\scala\NewNLP.java - IntelliJ IDEA 2017.1.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalaProject src main scala NewNLP
Sample
Project
ScalaProject [scalaproject] C:\Users\pavan\IdeaProjects\Scala
.idea
project [scalaproject-build] sources root
src
main
scala
Hello
NewExample
NewNLP
Sample
ScalaSpark
test
target
build
input.txt
External Libraries
build.sbt
Hello.scala
Sample.java
NewNLP.java
NewExample.scala
CoreLabel.java
input.txt
NewNLP
main()
1: Project
7: Structure
2: Favorites
Ant Build
Database
SBT
Maven Projects
Bean Validation
6: TODO
Terminal
SBT Console
Java Enterprise
Event Log

```java
public class NewNLP {

    public static void main(String args[]) throws IOException {

        Properties props = new Properties();
        props.setProperty("annotators", "tokenize, ssplit, pos, lemma, ner, parse, dcoref");
        StanfordCoreNLP pipeline = new StanfordCoreNLP(props);
        Set personSet = new HashSet();
        Set locSet = new HashSet();
        Set orgSet = new HashSet();

        String text = readFile("input.txt");


        Annotation document = new Annotation(text);

        pipeline.annotate(document);

        List<CoreMap> sentencess = document.get(CoreAnnotations.SentencesAnnotation.class);

        for (CoreMap sentence : sentencess) {
            for (CoreLabel token : sentence.get(CoreAnnotations.TokensAnnotation.class)) {
                String nameAndEntity = token.get(CoreAnnotations.NamedEntityTagAnnotation.class);

                if (nameAndEntity.equals("PERSON")) {
                    personSet.add(token);
                }
                if (nameAndEntity.equals("LOCATION")) {
                    locSet.add(token);
                }
                if (nameAndEntity.equals("ORGANIZATION")) {
                    orgSet.add(token);
                }
```


ScalaProject - [C:\Users\pavan\IdeaProjects\ScalaProject] - [scalaproject] - ...\src\main\scala\NewNLP.java - IntelliJ IDEA 2017.1.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalaProject src main scala NewNLP
Sample
Project
build.sbt Hello.scala Sample.java NewNLP.java NewExample.scala CoreLabel.java input.txt
ScalaProject [scalaproject] C:\Users\pavan\IdeaProjects\Scala
.idea
project [scalaproject-build] sources root
src
main
scala
Hello
NewExample
NewNLP
Sample
ScalaSpark
test
target
build
input.txt
External Libraries
NewNLP main()
6: TODO Terminal SBT Console Java Enterprise
Event Log
71:52 CRLF UTF-8

```
        System.out.println("-------Named Entity Recognition----");
        System.out.println("\n\n");
        System.out.println("People");
        System.out.print(personSet);
        System.out.println("\n\n");
        System.out.println("Location");
        System.out.print(locSet);
        System.out.println("\n\n");
        System.out.println("Organization");
        System.out.print(orgSet);
        System.out.println("\n\n");
        System.out.println("Enter question based on above data");
        Scanner sc = new Scanner(System.in);
        String ques = sc.nextLine();
        if(ques.equalsIgnoreCase( anotherString: "Web Blog community is called into support which Bloggers")){
            System.out.println("Iranian bloggers");
        }
        if(ques.equalsIgnoreCase( anotherString: "What is Obama inspiration")){
            System.out.println("The first Obama-inspired trend is Total Absorption");
        }

    }
    public static String readFile(String file) throws IOException {
        BufferedReader bReader = new BufferedReader(new FileReader(file));
        try {
            StringBuilder stB = new StringBuilder();
            String line = bReader.readLine();

            while (line != null) {
                stB.append(line);
                stB.append("\n");
                line = bReader.readLine();
            }
```

The below screenshot shows the output of our dataset which was categorized into different NER tags. When we compose a question, the corresponding answer will be taken based on the list of the NER entities.


ScalaProject - [C:\Users\pavan\IdeaProjects\ScalaProject] - [scalaproject] - ...\src\main\scala\NewNLP.java - IntelliJ IDEA 2017.1.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalaProject src main scala NewNLP
NewNLP
Project
Sample.java NewNLP.java wiki.txt NewExample.scala input.txt
ScalaProject [scalaproject] C:\Users\pavan\IdeaProjects\ScalaP
.idea
project [scalaproject-build] sources root
src
main
scala
Hello
NewExample
NewNLP
NewNLP main()
Run: NewNLP NewNLP
4: Run 6: TODO Terminal SBT Console 0: Messages Java Enterprise
Event Log
72:10 CRLF UTF-8

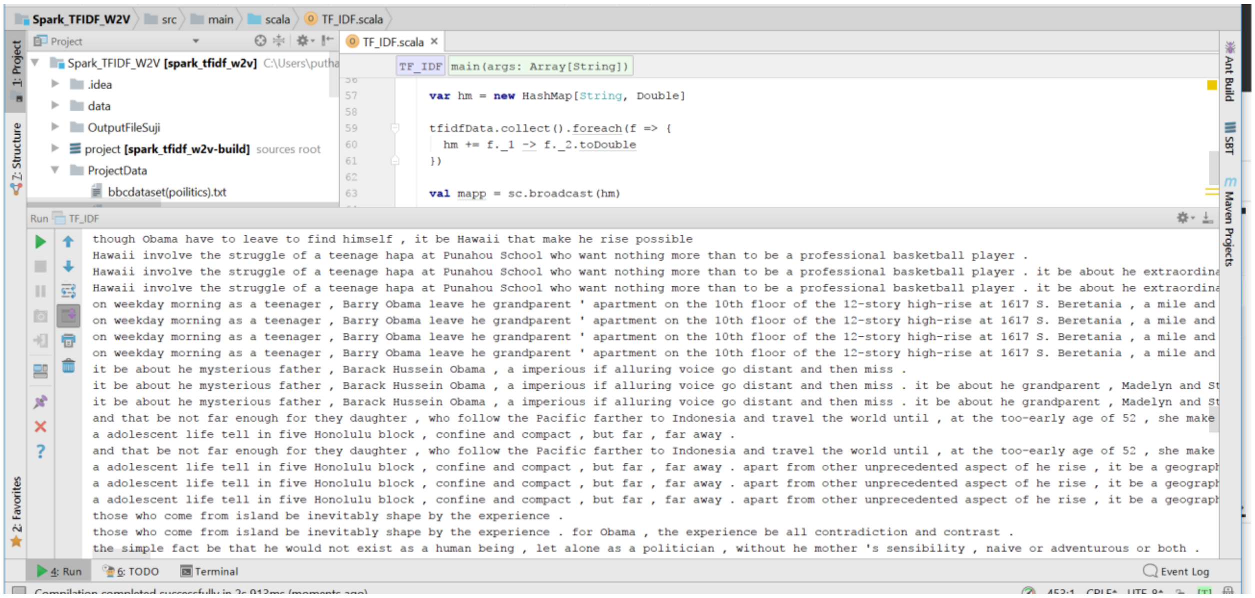
```
                String nameAndEntity = token.get(CoreAnnotations.NamedEntityTagAnnotation.class);

                if (nameAndEntity.equals("PERSON")) {
                    personSet.add(token);
                }
                if (nameAndEntity.equals("LOCATION")) {
                    locSet.add(token);
                }
                if (nameAndEntity.equals("ORGANIZATION")) {
```

```
People
[Friedel-15, Stapleton-32, Van-3, Wilson-32, Andrews-4, Sayer-109, Neill-78, Chris-8, Modric-122, Mark-31, Lennon-115, Raymond-1, Dicks-24, Behrami-11, Nsereko-93, Dixon-32, Jona


Location
[UK-26, South-18, Hampstead-106, Urmston-111, France-96, Petersburg-25, Northern-10, Ohio-36, Wales-10, UK-46, UK-29, Farnworth-56, Glasgow-58, UK-77, Hillingdon-145, Quito-31, S


Organization
[Crystal-11, Southampton-35, UNITED-8, Leeds-117, St-71, Middlesbrough-109, Chelsea-9, Coventry-100, Fulham-3, Liverpool-13, Leamington-70, McDonald-43, Wigan-64, Barnsley-75, Ha


Enter question based on above data
What are the three imports of West Africa
Emmanuel Eboue, Johan Djourou and Alex Song

Process finished with exit code 0
```

Compilation completed successfully in 2s 462ms (48 minutes ago)

5.3 TF-IDF for our dataset

Generating the term frequency for the words in the dataset.

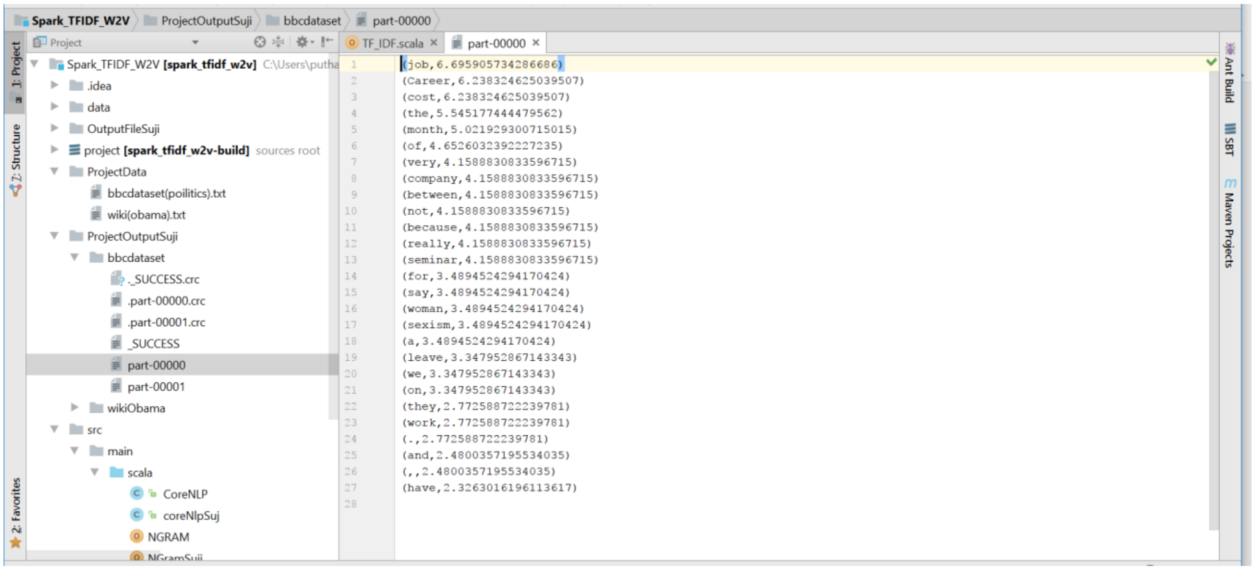


```
    var hm = new HashMap[String, Double]

    tfidfData.collect().foreach(f => {
      hm += f._1 -> f._2.toDouble
    })

    val mapp = sc.broadcast(hm)
```

```
though Obama have to leave to find himself , it be Hawaii that make he rise possible
Hawaii involve the struggle of a teenage hapa at Punahou School who want nothing more than to be a professional basketball player .
Hawaii involve the struggle of a teenage hapa at Punahou School who want nothing more than to be a professional basketball player . it be about he extraordina
Hawaii involve the struggle of a teenage hapa at Punahou School who want nothing more than to be a professional basketball player . it be about he extraordina
on weekday morning as a teenager , Barry Obama leave he grandparent ' apartment on the 10th floor of the 12-story high-rise at 1617 S. Beretania , a mile and
on weekday morning as a teenager , Barry Obama leave he grandparent ' apartment on the 10th floor of the 12-story high-rise at 1617 S. Beretania , a mile and
on weekday morning as a teenager , Barry Obama leave he grandparent ' apartment on the 10th floor of the 12-story high-rise at 1617 S. Beretania , a mile and
on weekday morning as a teenager , Barry Obama leave he grandparent ' apartment on the 10th floor of the 12-story high-rise at 1617 S. Beretania , a mile and
it be about he mysterious father , Barack Hussein Obama , a imperious if alluring voice go distant and then miss .
it be about he mysterious father , Barack Hussein Obama , a imperious if alluring voice go distant and then miss . it be about he grandparent , Madelyn and St
it be about he mysterious father , Barack Hussein Obama , a imperious if alluring voice go distant and then miss . it be about he grandparent , Madelyn and St
and that be not far enough for they daughter , who follow the Pacific farther to Indonesia and travel the world until , at the too-early age of 52 , she make
a adolescent life tell in five Honolulu block , confine and compact , but far , far away .
and that be not far enough for they daughter , who follow the Pacific farther to Indonesia and travel the world until , at the too-early age of 52 , she make
a adolescent life tell in five Honolulu block , confine and compact , but far , far away . apart from other unprecedented aspect of he rise , it be a geograph
a adolescent life tell in five Honolulu block , confine and compact , but far , far away . apart from other unprecedented aspect of he rise , it be a geograph
a adolescent life tell in five Honolulu block , confine and compact , but far , far away . apart from other unprecedented aspect of he rise , it be a geograph
those who come from island be inevitably shape by the experience .
those who come from island be inevitably shape by the experience . for Obama , the experience be all contradiction and contrast .
the simple fact be that he would not exist as a human being , let alone as a politician , without he mother 's sensibility , naive or adventurous or both .
```

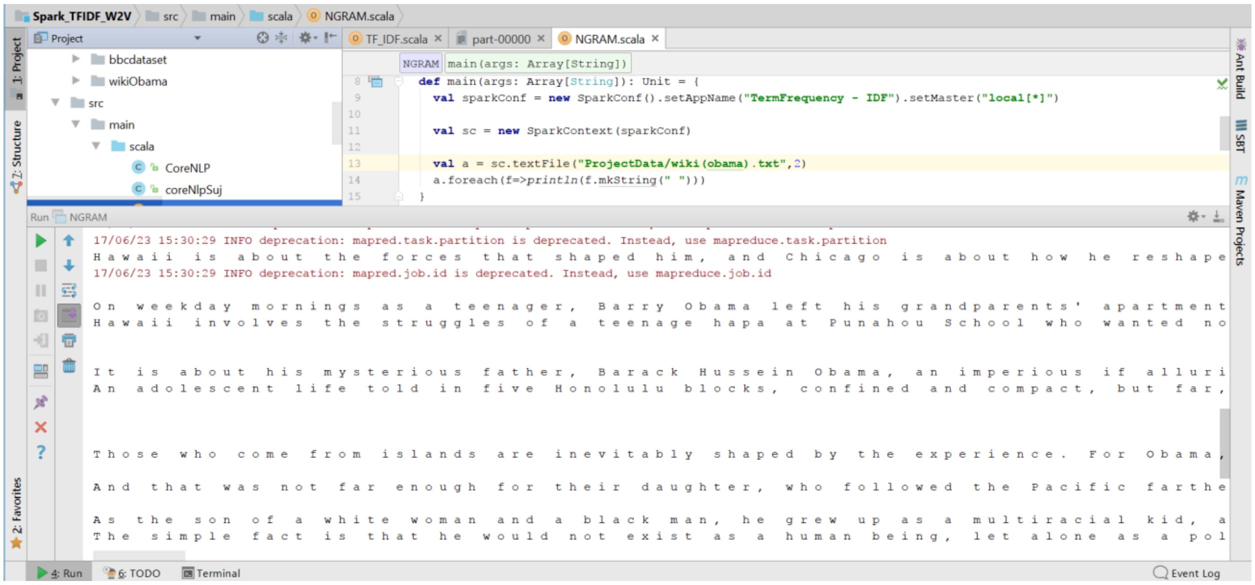
Compilation completed successfully in 2s 913ms (moments ago)
453:1 CRLF UTF-8

Generated Output.



```
(job,6.695905734286686)
(Career,6.238324625039507)
(cost,6.238324625039507)
(the,5.545177444479562)
(month,5.021929300715015)
(of,4.6526032392227235)
(very,4.1588830833596715)
(company,4.1588830833596715)
(between,4.1588830833596715)
(not,4.1588830833596715)
(because,4.1588830833596715)
(really,4.1588830833596715)
(seminar,4.1588830833596715)
(for,3.4894524294170424)
(say,3.4894524294170424)
(woman,3.4894524294170424)
(sexism,3.4894524294170424)
(a,3.4894524294170424)
(leave,3.347952867143343)
(we,3.347952867143343)
(on,3.347952867143343)
(they,2.772588722239781)
(work,2.772588722239781)
(.,2.772588722239781)
(and,2.4800357195534035)
(,,2.4800357195534035)
(have,2.3263016196113617)
```

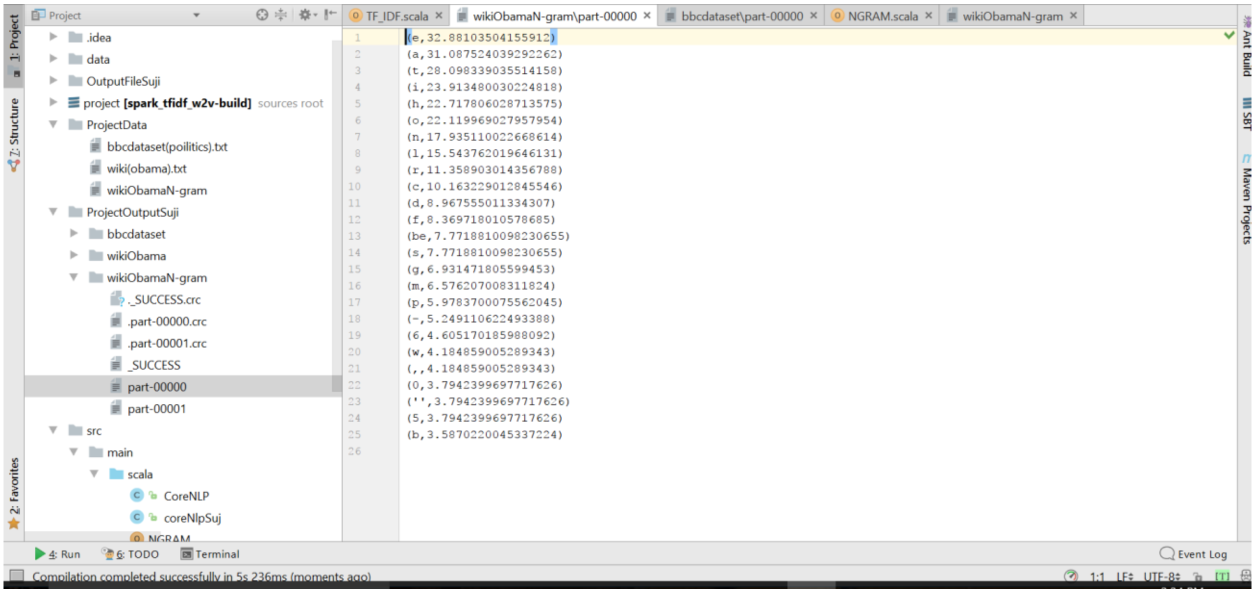
Generation N-gram for the dataset



```
def main(args: Array[String]): Unit = {
  val sparkConf = new SparkConf().setAppName("TermFrequency - IDF").setMaster("local[*]")

  val sc = new SparkContext(sparkConf)

  val a = sc.textFile("ProjectData/wiki(obama).txt",2)
  a.foreach(f=>println(f.mkString(" ")))
}
```

```
17/06/23 15:30:29 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
H a w a i i   i s   a b o u t   t h e   f o r c e s   t h a t   s h a p e d   h i m ,   a n d   C h i c a g o   i s   a b o u t   h o w   h e   r e s h a p e
17/06/23 15:30:29 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id

O n   w e e k d a y   m o r n i n g s   a s   a   t e e n a g e r ,   B a r r y   O b a m a   l e f t   h i s   g r a n d p a r e n t s '   a p a r t m e n t
H a w a i i   i n v o l v e s   t h e   s t r u g g l e s   o f   a   t e e n a g e   h a p a   a t   P u n a h o u   S c h o o l   w h o   w a n t e d   n o


I t   i s   a b o u t   h i s   m y s t e r i o u s   f a t h e r ,   B a r a c k   H u s s e i n   O b a m a ,   a n   i m p e r i o u s   i f   a l l u r i
A n   a d o l e s c e n t   l i f e   t o l d   i n   f i v e   H o n o l u l u   b l o c k s ,   c o n f i n e d   a n d   c o m p a c t ,   b u t   f a r ,



T h o s e   w h o   c o m e   f r o m   i s l a n d s   a r e   i n e v i t a b l y   s h a p e d   b y   t h e   e x p e r i e n c e .   F o r   O b a m a ,

A n d   t h a t   w a s   n o t   f a r   e n o u g h   f o r   t h e i r   d a u g h t e r ,   w h o   f o l l o w e d   t h e   P a c i f i c   f a r t h e

A s   t h e   s o n   o f   a   w h i t e   w o m a n   a n d   a   b l a c k   m a n ,   h e   g r e w   u p   a s   a   m u l t i r a c i a l   k i d ,   a
T h e   s i m p l e   f a c t   i s   t h a t   h e   w o u l d   n o t   e x i s t   a s   a   h u m a n   b e i n g ,   l e t   a l o n e   a s   a   p o l
```

Generating the TF_IDF for N-gram output.



```
(e,32.88103504155912)
(a,31.087524039292262)
(t,28.098339035514158)
(i,23.913480030224818)
(h,22.717806028713575)
(o,22.119969027957954)
(n,17.935110022668614)
(l,15.543762019646131)
(r,11.358903014356788)
(c,10.163229012845546)
(d,8.967555011334307)
(f,8.369718010578685)
(be,7.7718810098230655)
(s,7.7718810098230655)
(g,6.931471805599453)
(m,6.576207008311824)
(p,5.9783700075562045)
(-,5.249110622493388)
(6,4.605170185988092)
(w,4.184859005289343)
(,,4.184859005289343)
(0,3.794239969771626)
('',3.794239969771626)
(5,3.794239969771626)
(b,3.5870220045337224)
```

Compilation completed successfully in 5s 236ms (moments ago)

6. Project Management

Programming Language Used:

We have collaborated various languages in the development of the project and in building the application. Some of them are,

- Java
- Scala
- Spark

IDE Used:

- IntelliJ
- PyCharm

6.1 Contributors

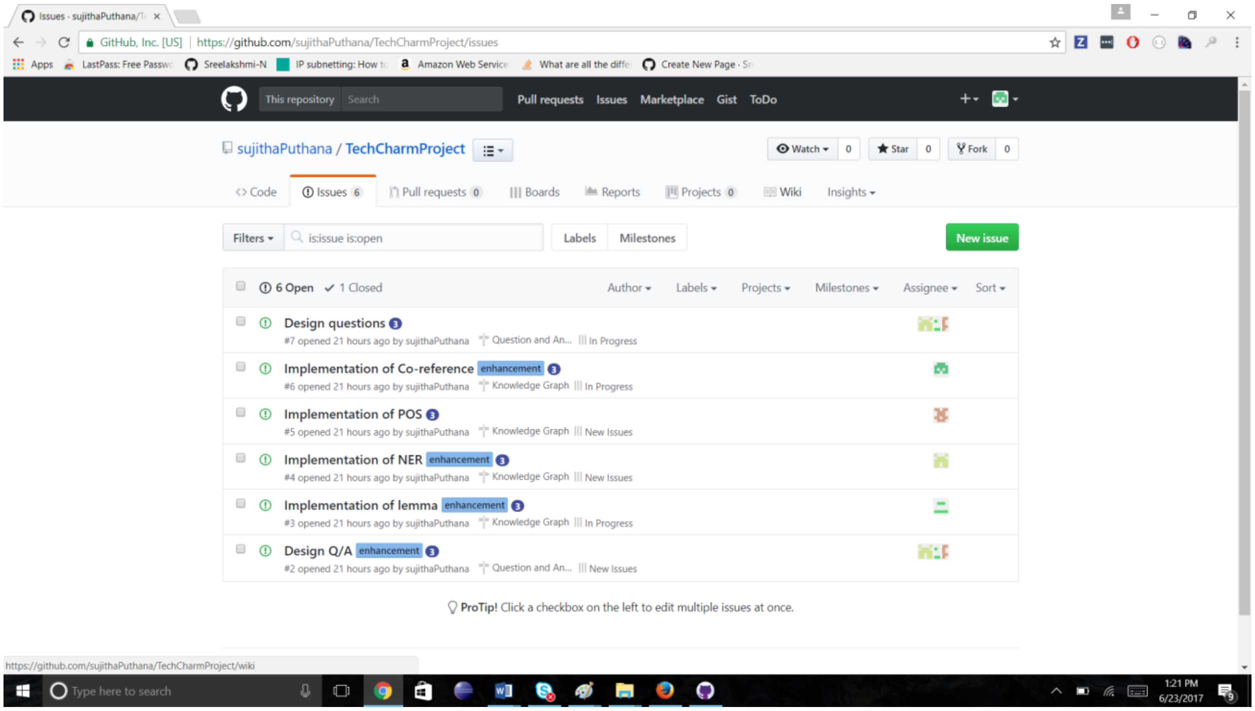
- Jakkepalli Rama Charan Pavan - **25%**
- Puthana Sujitha - **25%**
- Yalamanchili Sowmya **- 25%**
- Nandanamudi Sreelakshmi **- 25%**

Below is the bar graph that represents contribution of each person in the team towards project,

BAR CHART
25%
Pavan
25%
Sujitha
25%
Sowmya
25%
Sreelakshmi

6.2 Zen-Hub Screenshots

For the first increment, we had issues regarding the working of the questions and answers section and generating the NLP output for the dataset we have chosen as the size of the dataset is larger.



Design questions 3
#7 opened 21 hours ago by sujithaPuthana Question and An... In Progress

Implementation of Co-reference enhancement 3
#6 opened 21 hours ago by sujithaPuthana Knowledge Graph In Progress

Implementation of POS 3
#5 opened 21 hours ago by sujithaPuthana Knowledge Graph New Issues

Implementation of NER enhancement 3
#4 opened 21 hours ago by sujithaPuthana Knowledge Graph New Issues

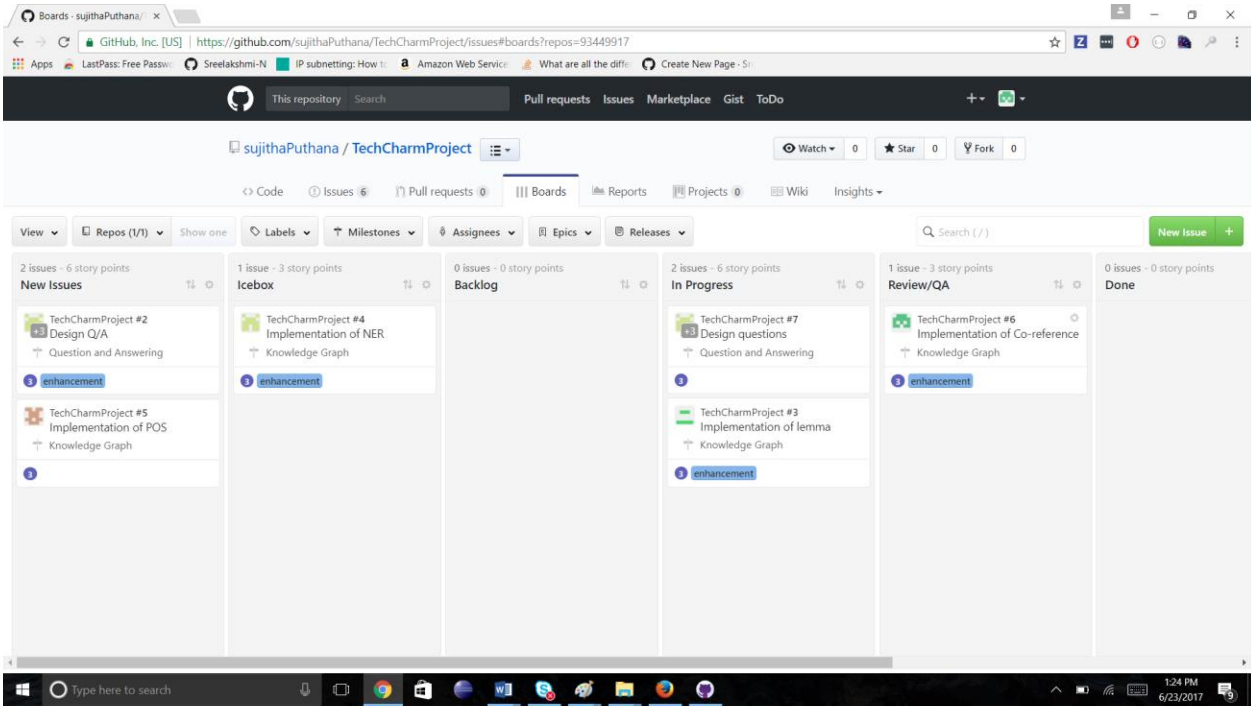
Implementation of lemma enhancement 3
#3 opened 21 hours ago by sujithaPuthana Knowledge Graph In Progress

Design Q/A enhancement 3
#2 opened 21 hours ago by sujithaPuthana Question and An... New Issues

ProTip! Click a checkbox on the left to edit multiple issues at once.

Project Timeline, Members, and Task Responsibility

The issues that are registered and current one’s which we are working are updated and can be viewed in GitHub repository. The below screenshot will show you the issues and their respective categorization’s i.e. New issues, Icebox, Backlog, In Progress.



New Issues — 2 issues - 6 story points

- TechCharmProject #2 Design Q/A — Question and Answering — 3 enhancement
- TechCharmProject #5 Implementation of POS — Knowledge Graph — 3

Icebox — 1 issue - 3 story points

- TechCharmProject #4 Implementation of NER — Knowledge Graph — 3 enhancement

Backlog — 0 issues - 0 story points

In Progress — 2 issues - 6 story points

- TechCharmProject #7 Design questions — Question and Answering — 3
- TechCharmProject #3 Implementation of lemma — Knowledge Graph — 3 enhancement

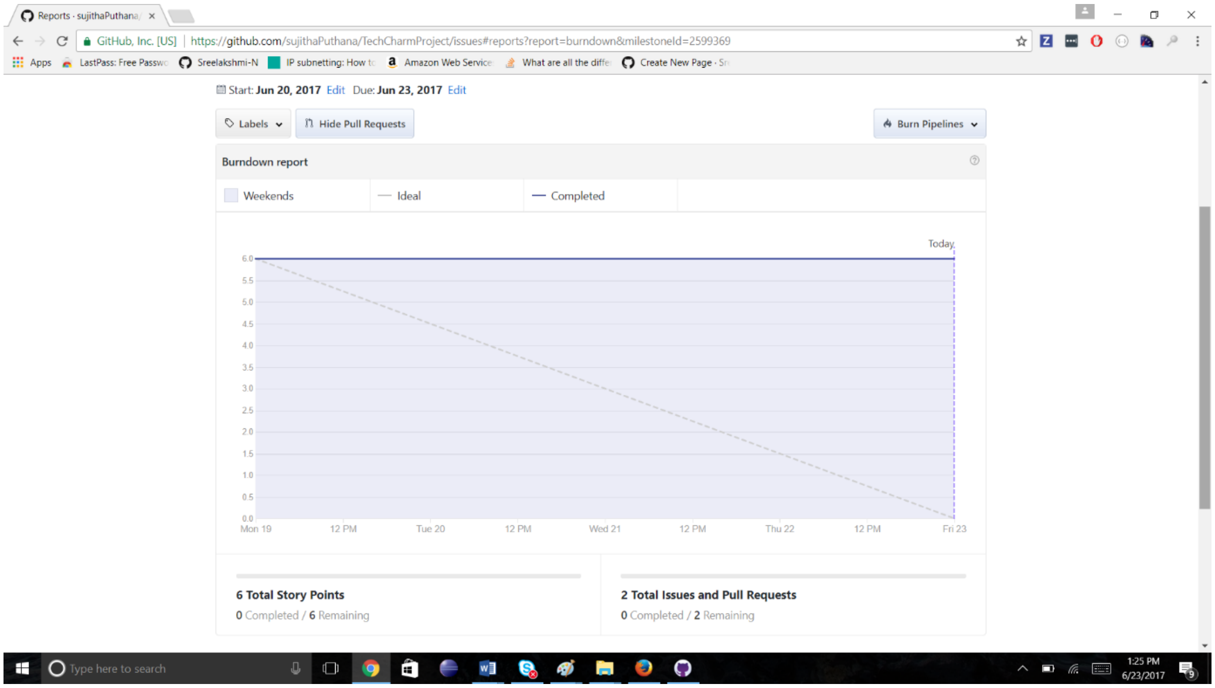
Review/QA — 1 issue - 3 story points

- TechCharmProject #6 Implementation of Co-reference — Knowledge Graph — 3 enhancement

Done — 0 issues - 0 story points

Burn-Down Chart:

Burn-Down chart is created for the above issues via Milestones in GitHub. Below is the screenshot for more information,

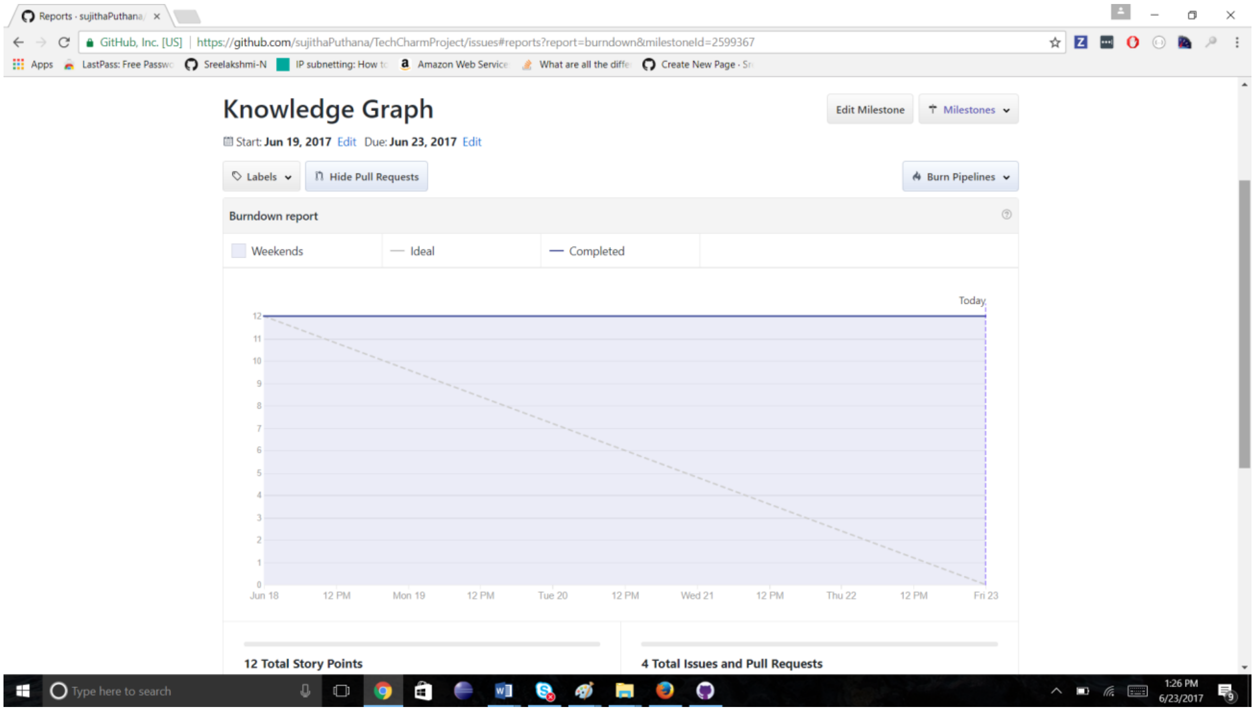


Start: Jun 20, 2017 Edit Due: Jun 23, 2017 Edit

Burndown report

6 Total Story Points
0 Completed / 6 Remaining

2 Total Issues and Pull Requests
0 Completed / 2 Remaining



Knowledge Graph

Start: Jun 19, 2017 Edit Due: Jun 23, 2017 Edit

Burndown report

12 Total Story Points

4 Total Issues and Pull Requests

GitHub Wiki Page

The GitHub wiki page URL for the screenshots and the process flow is updated in the following link

- https://github.com/sujithaPuthana/TechcharmProject

6.3 Work Completed

The completed tasks in this increment are,

- Performed the NLP operations on the dataset.
- Designing the question and answers for the NLP output.
- TF-IDE and N-gram analysis

6.3 Future Work

- We need to implement the question and answer approach using the TF-IDF integrated with the NLP operations.
- Need to integrate the TF-IDF approach with the N-GRAM and Word2Vec for better performance.

Bibliography

1. https://blog.algorithmia.com/introduction-natural-language-processing-nlp/
2. https://en.wikipedia.org/wiki/Question_answering
3. https://nlp.stanford.edu/