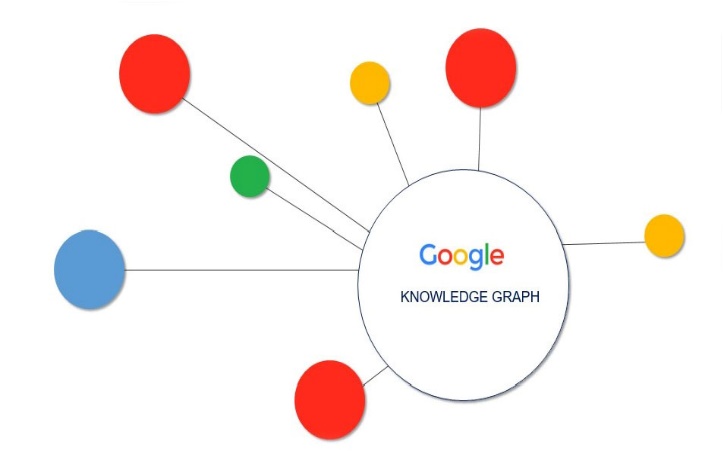
**Knowledge Discovery and Management**

**Project Report**



TechChamps

**Team 1**

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. **Specific Objectives:** 
     1. Easy search of information from huge amount of text.
     2. Helps in precise answer for customized questions.
     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**

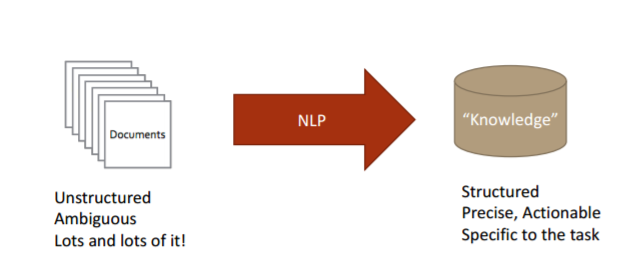
For our project implementation, we have considered two datasets as follows:

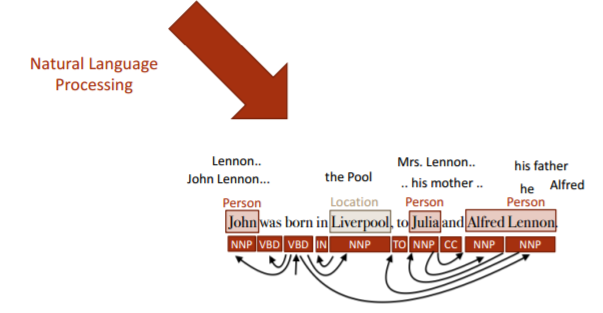
* WikiRef220
* BBC News-In this especially we have selected politics area and sports.

**4. Design**

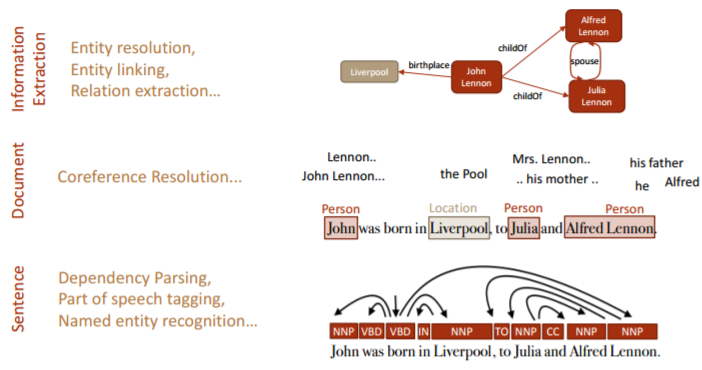
**4.1 Workflow**

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





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

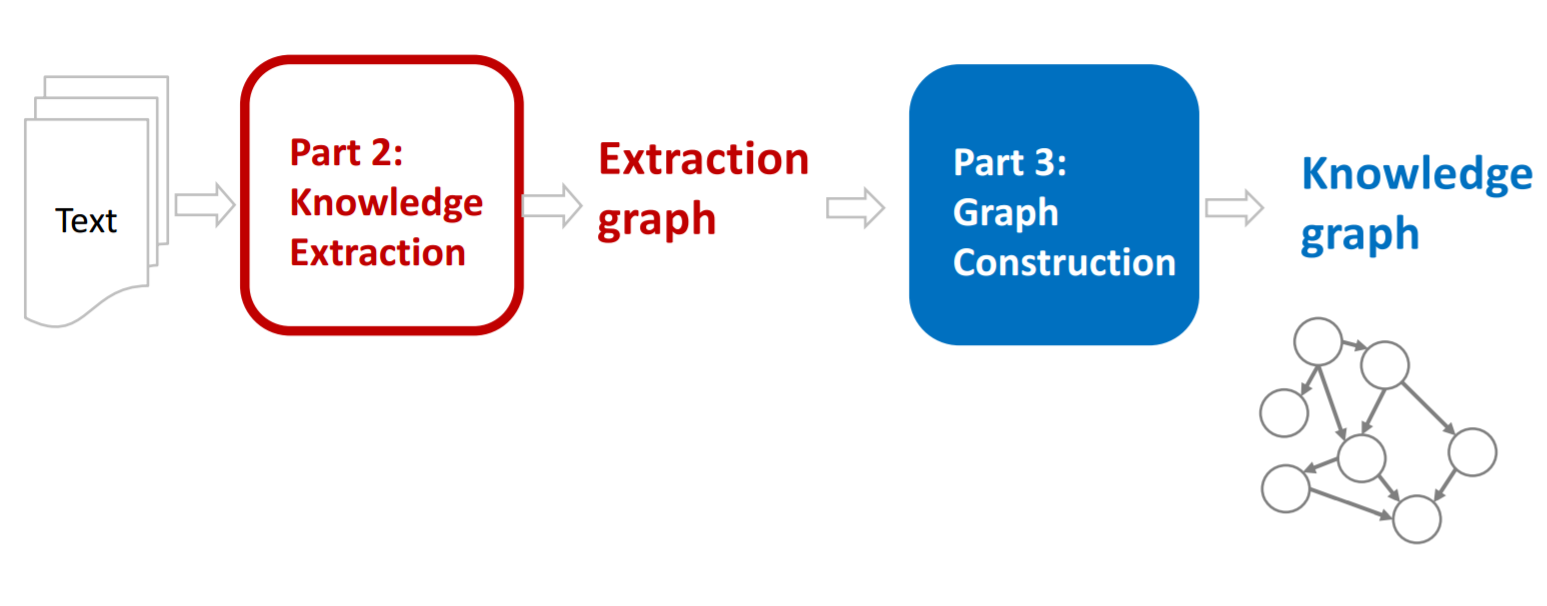


Step 3: Topic Discovery – Topic discovery helps identification of the topics from the context question.

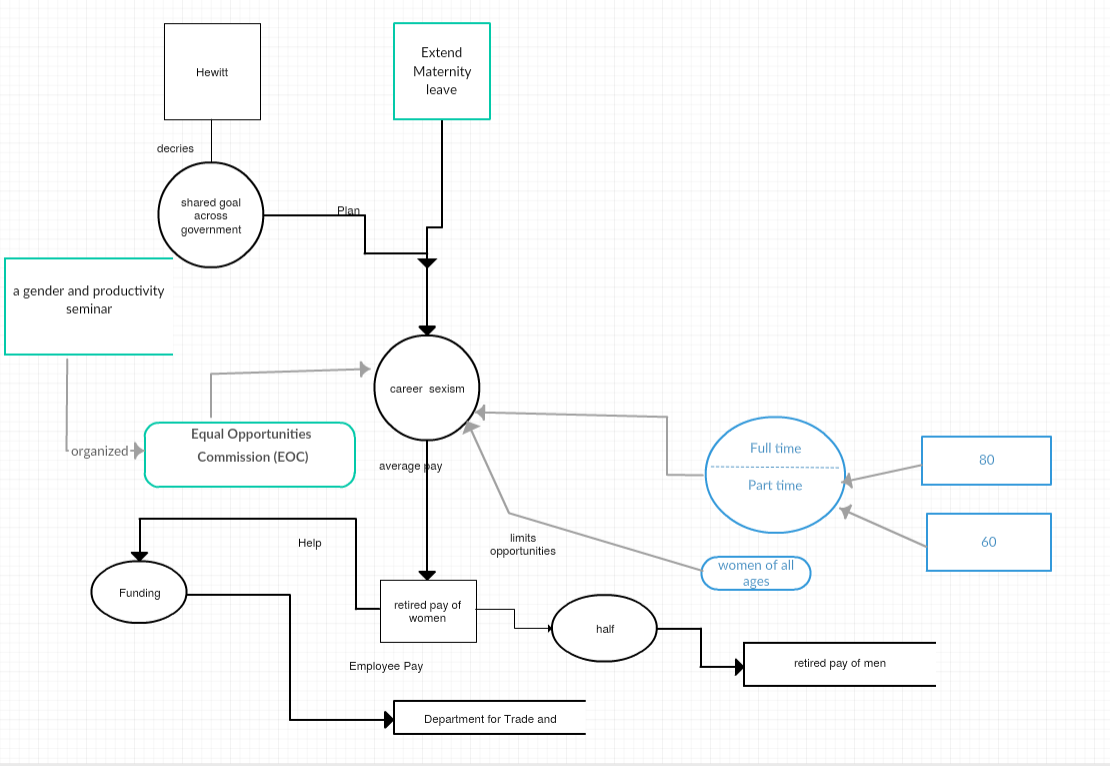
Step 4: Knowledge Graph construction – Construction of the knowledge graph from generated NER.

**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.

1. Where did Obama did his schooling?

Punahou School.

1. Who is father of Obama?

Barack Hussein Obama.

1. Whom did Obama compete in primary race?

Hillary Rodham Clinton.

1. What is the minimum duration for maternity leave?

6 months.

1. What is the topic about?

career sexism.

1. Who is the speaker?

Ms. Hewitt.

1. What is the average pay for full-time women.

80p

1. What is the average pay for part-time women.

60p.

1. 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:

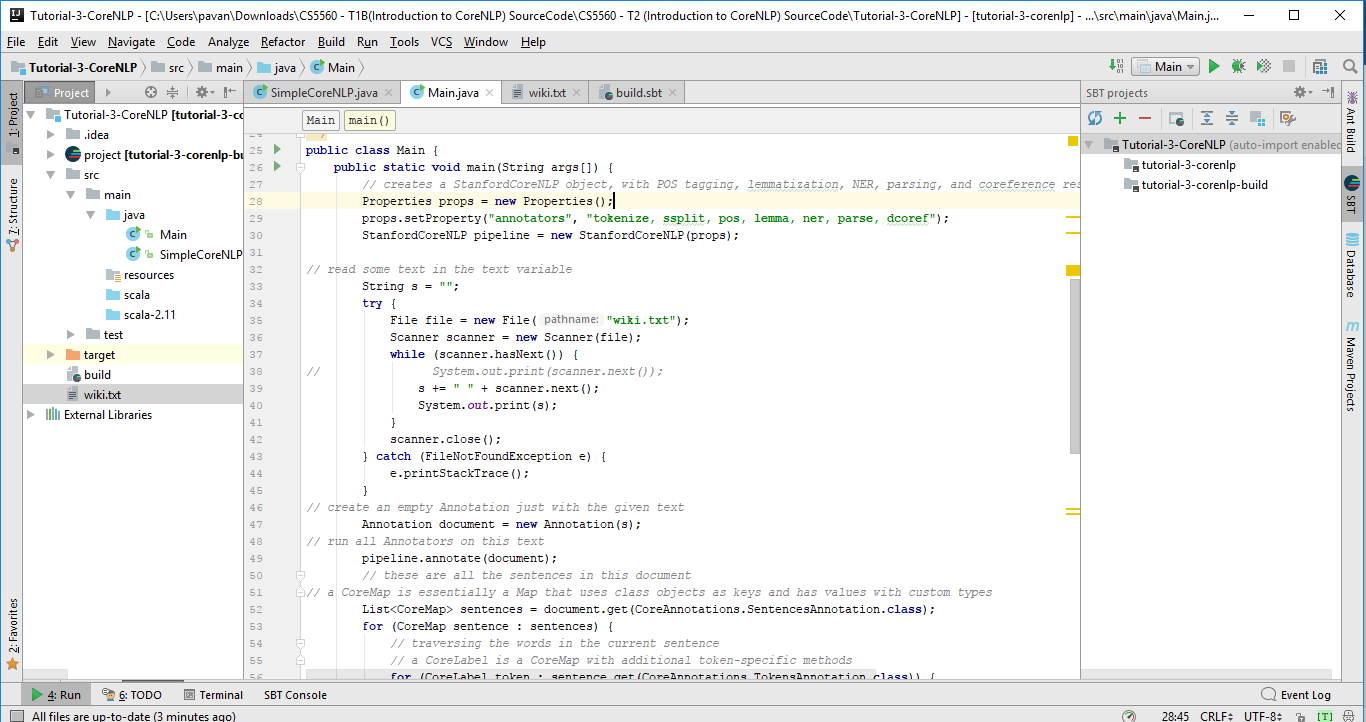
Lemmatization:

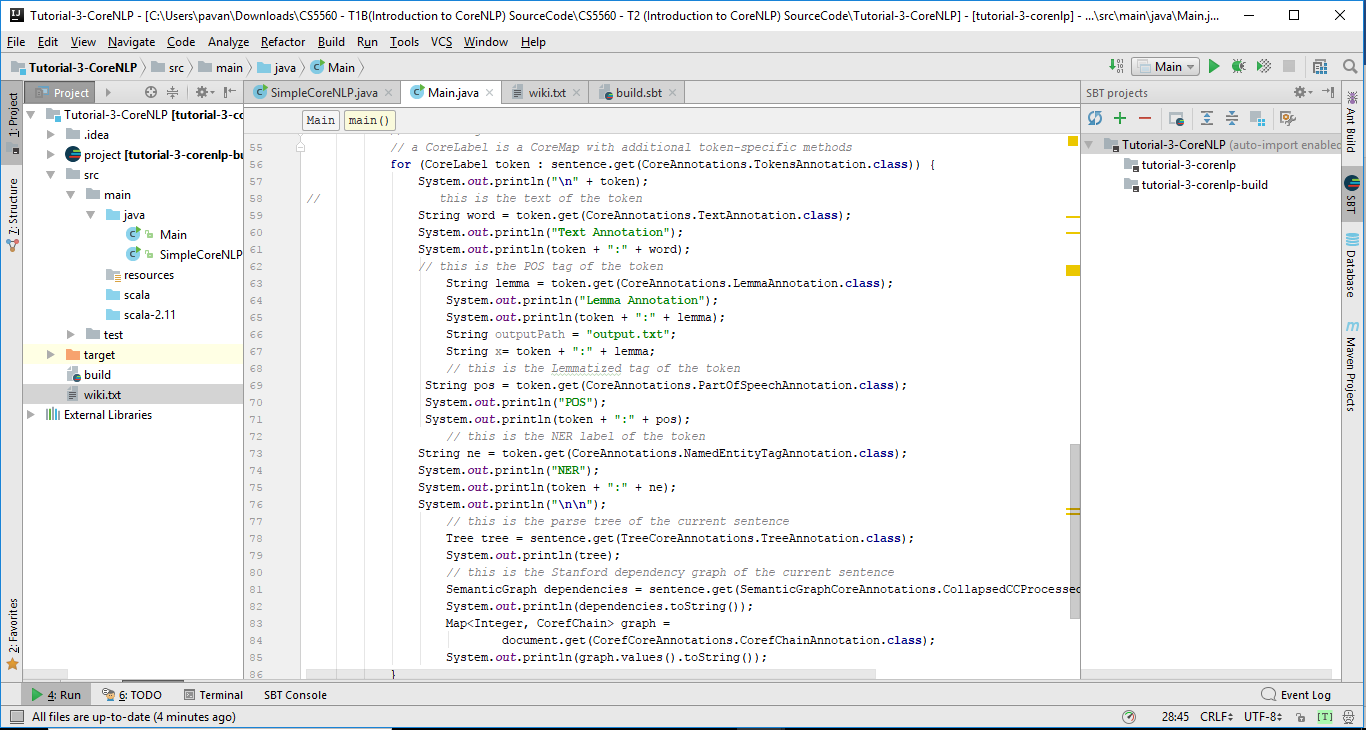
POS Tagging:

NER:

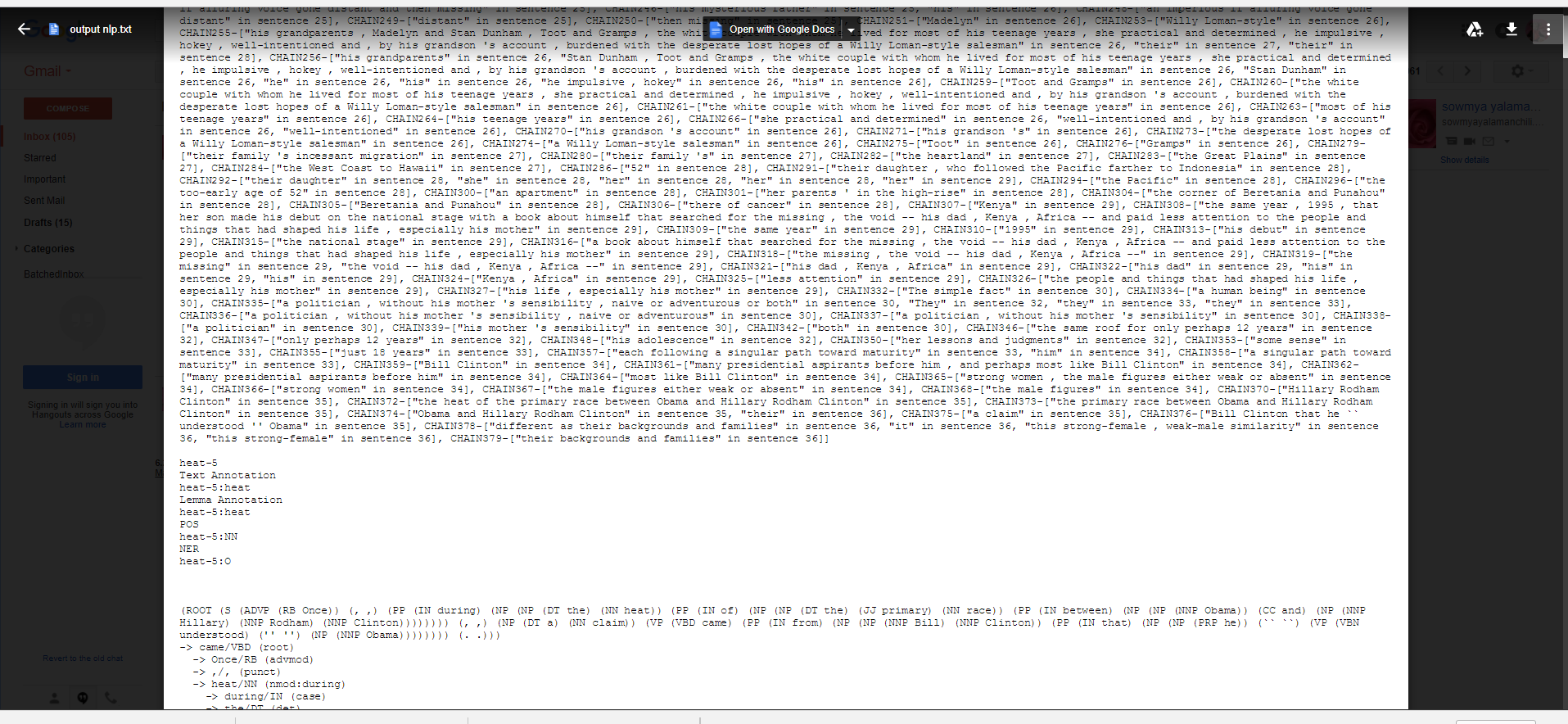
Coreference Resolution:

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



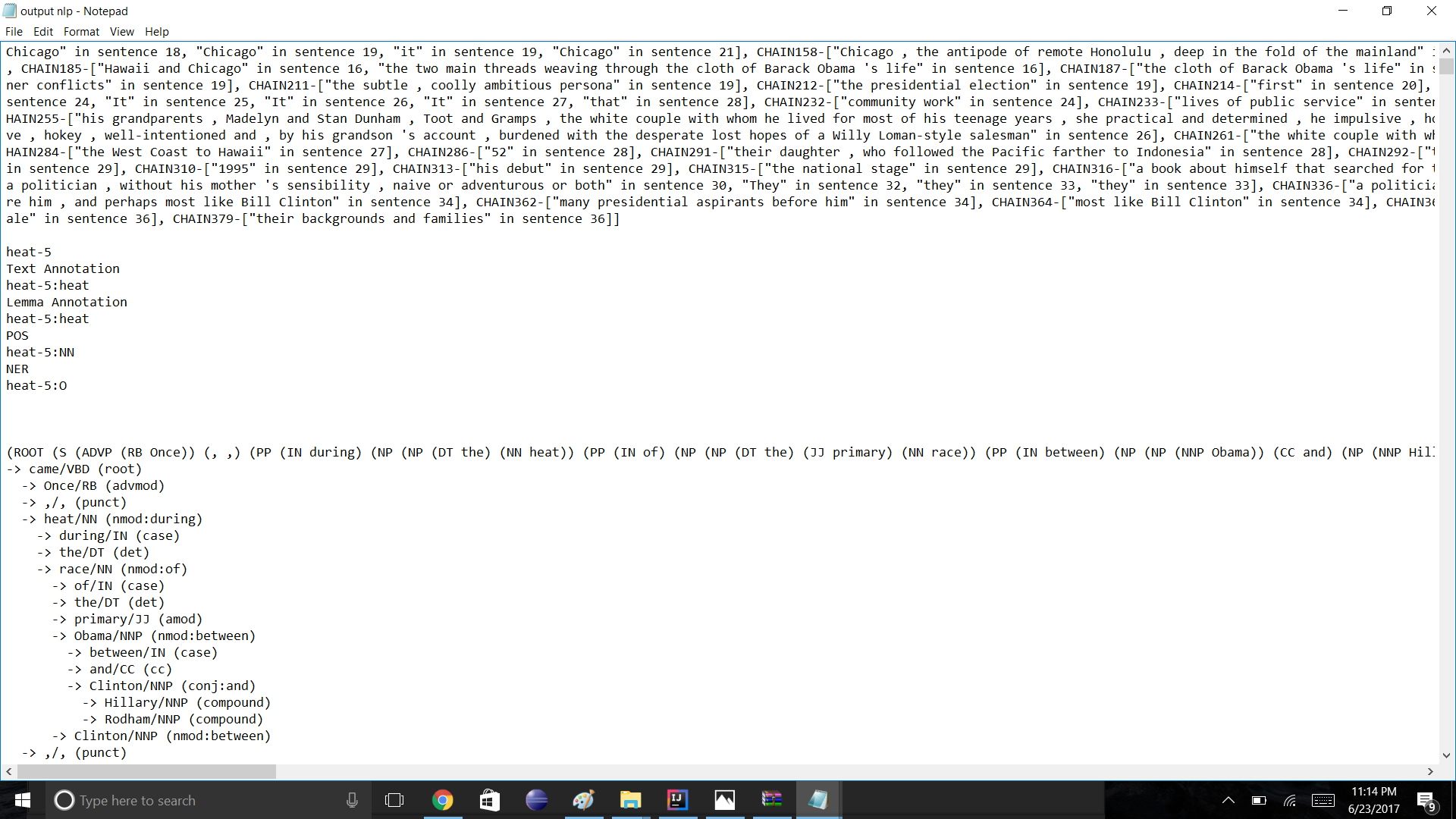


Below are the outputs of the operations.



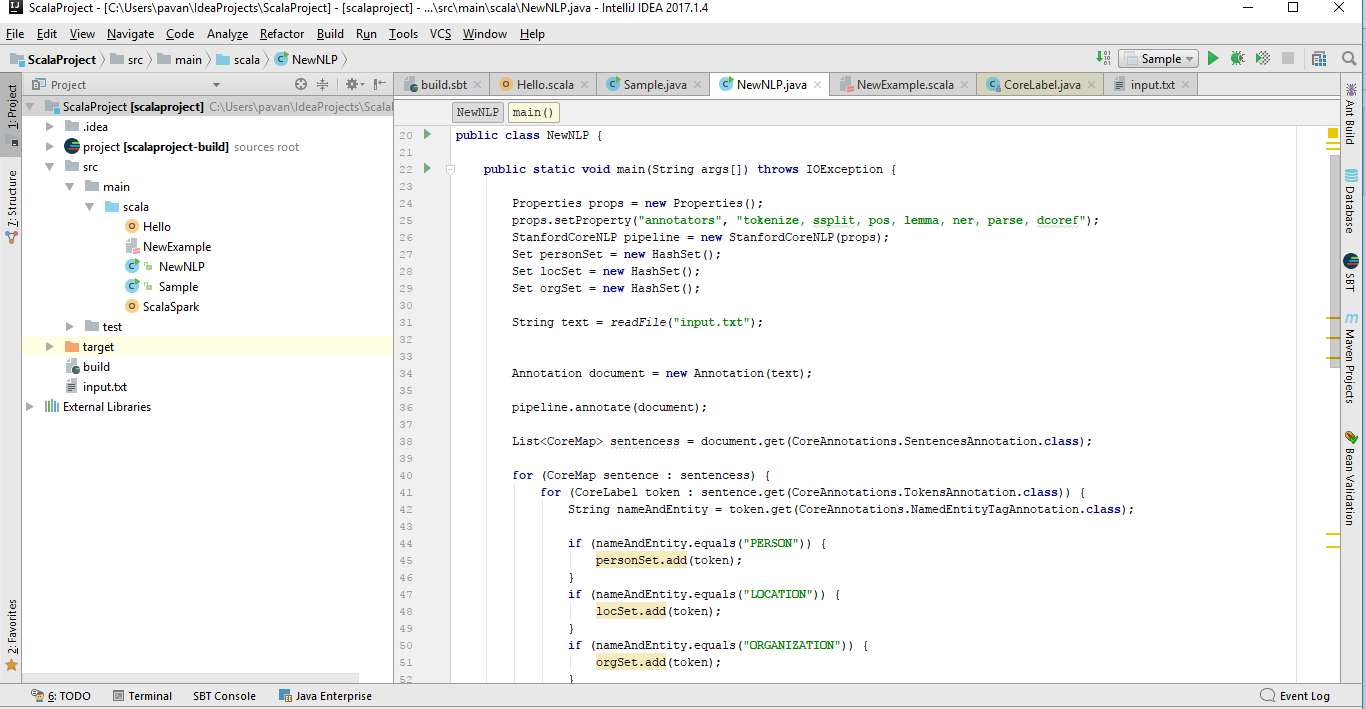


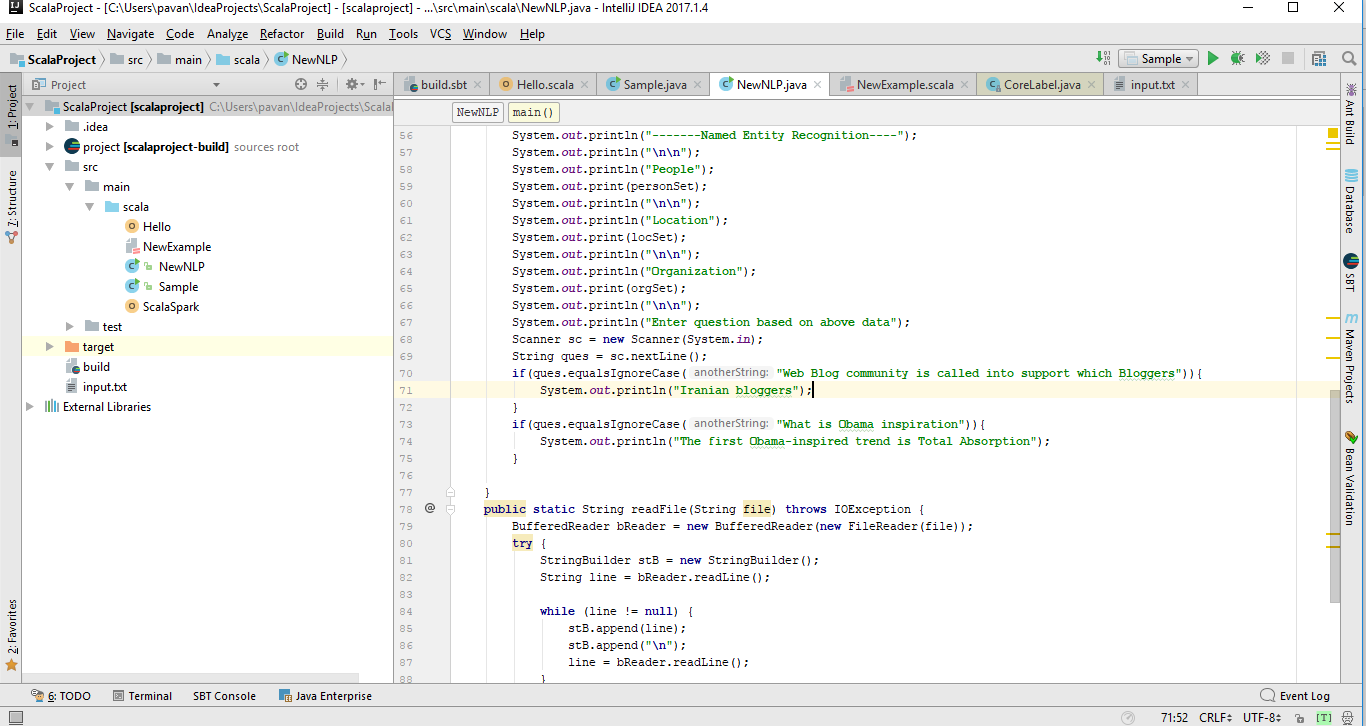




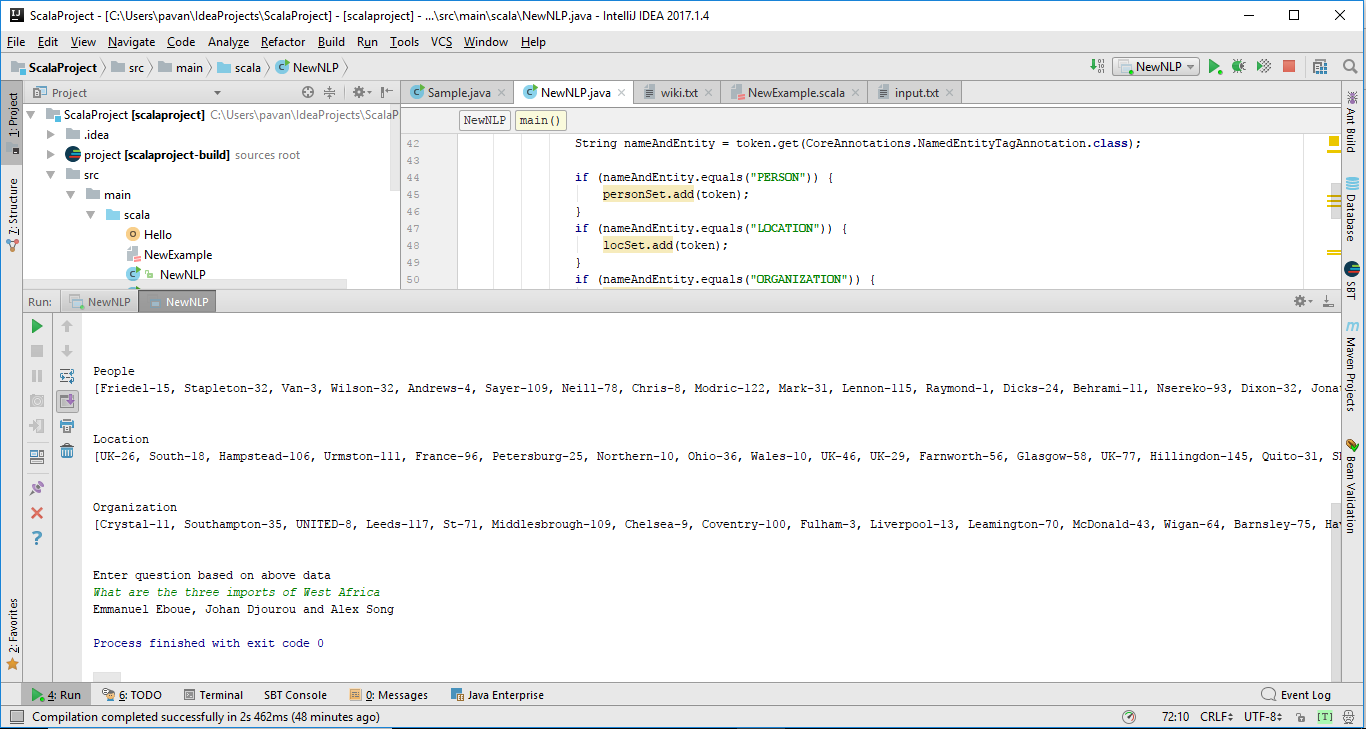
**5.2 Question Answering for our dataset**

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.



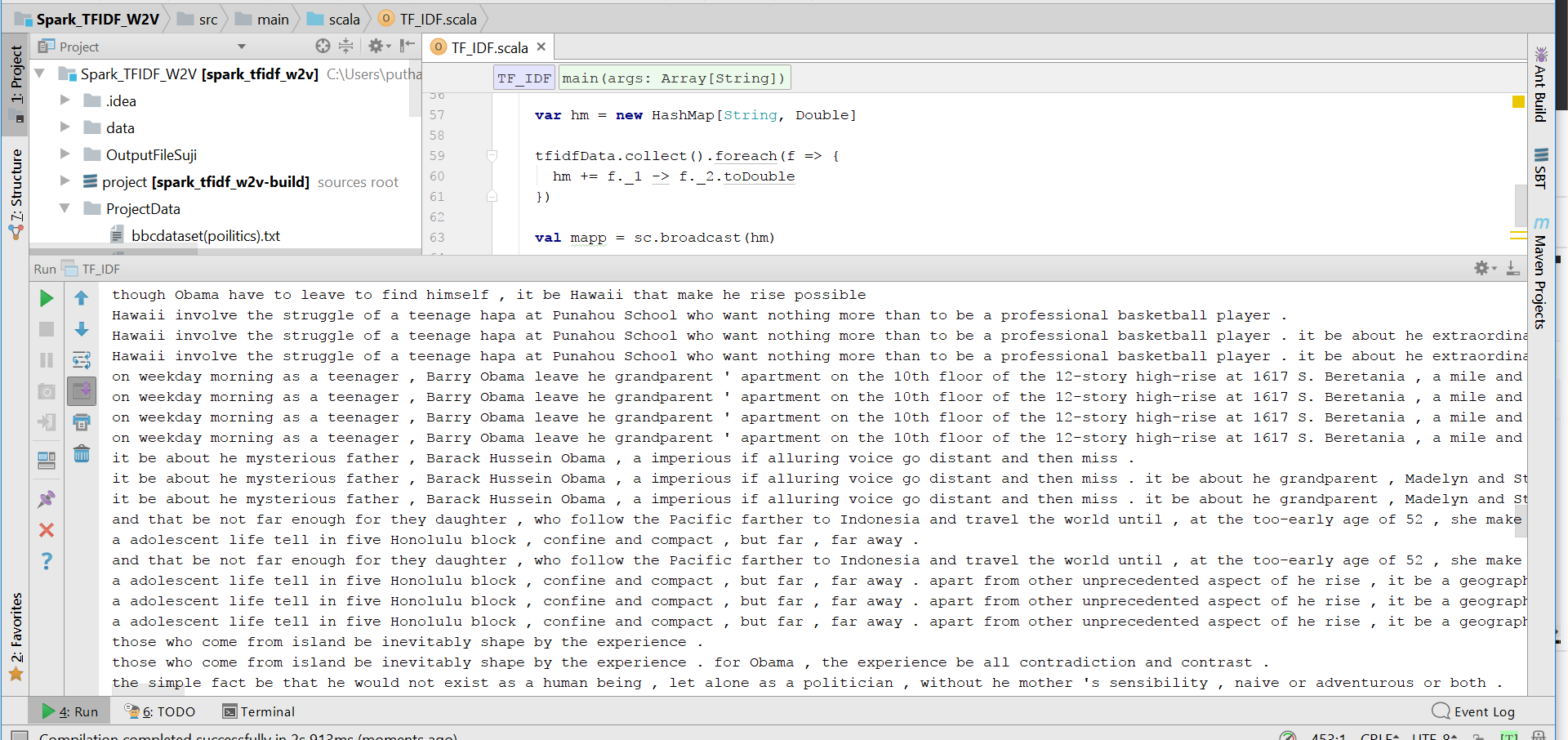


The below screenshot shows the output of our dataset which was categorized into different NER tags. When we compose a question, the corresponding answer will be taken based on the list of the NER entities.

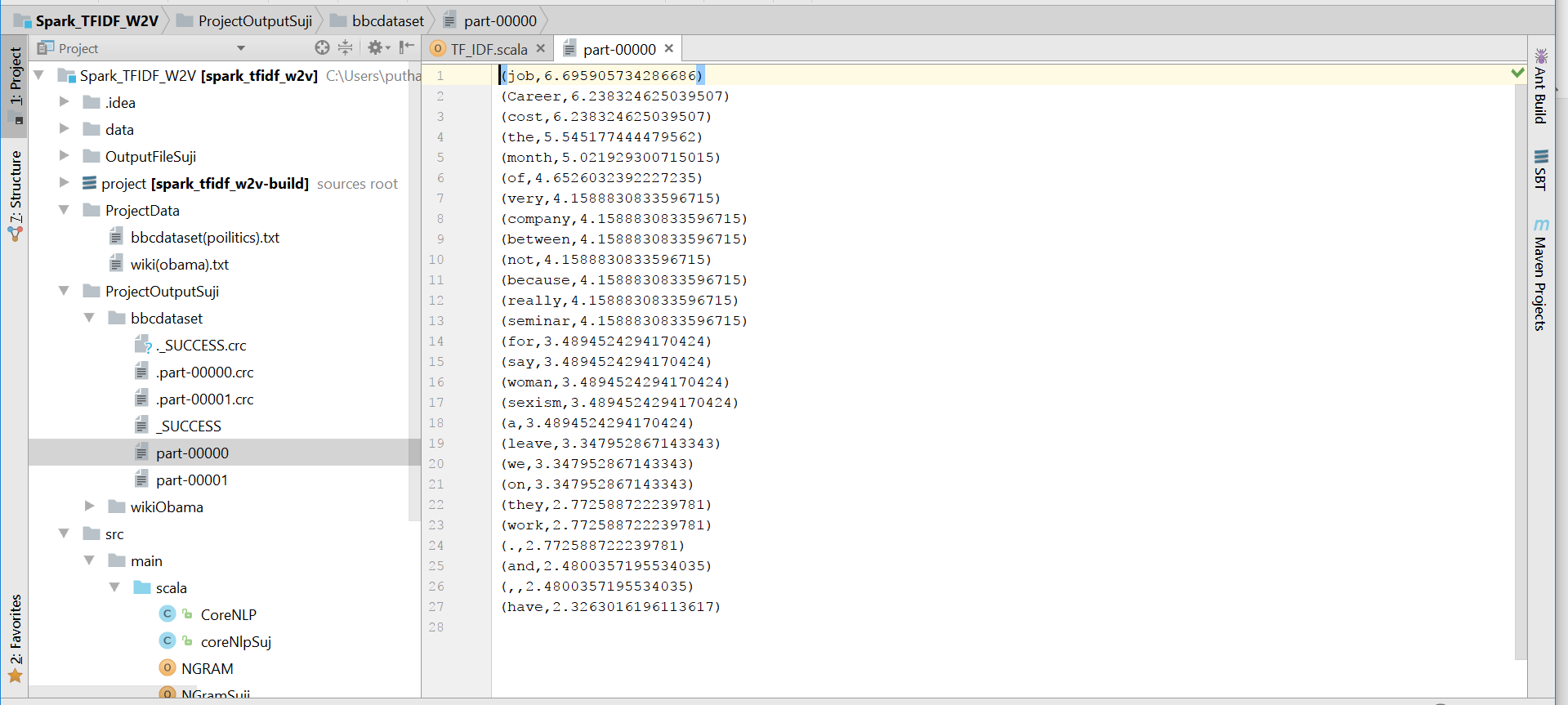


**5.3 TF-IDF for our dataset**

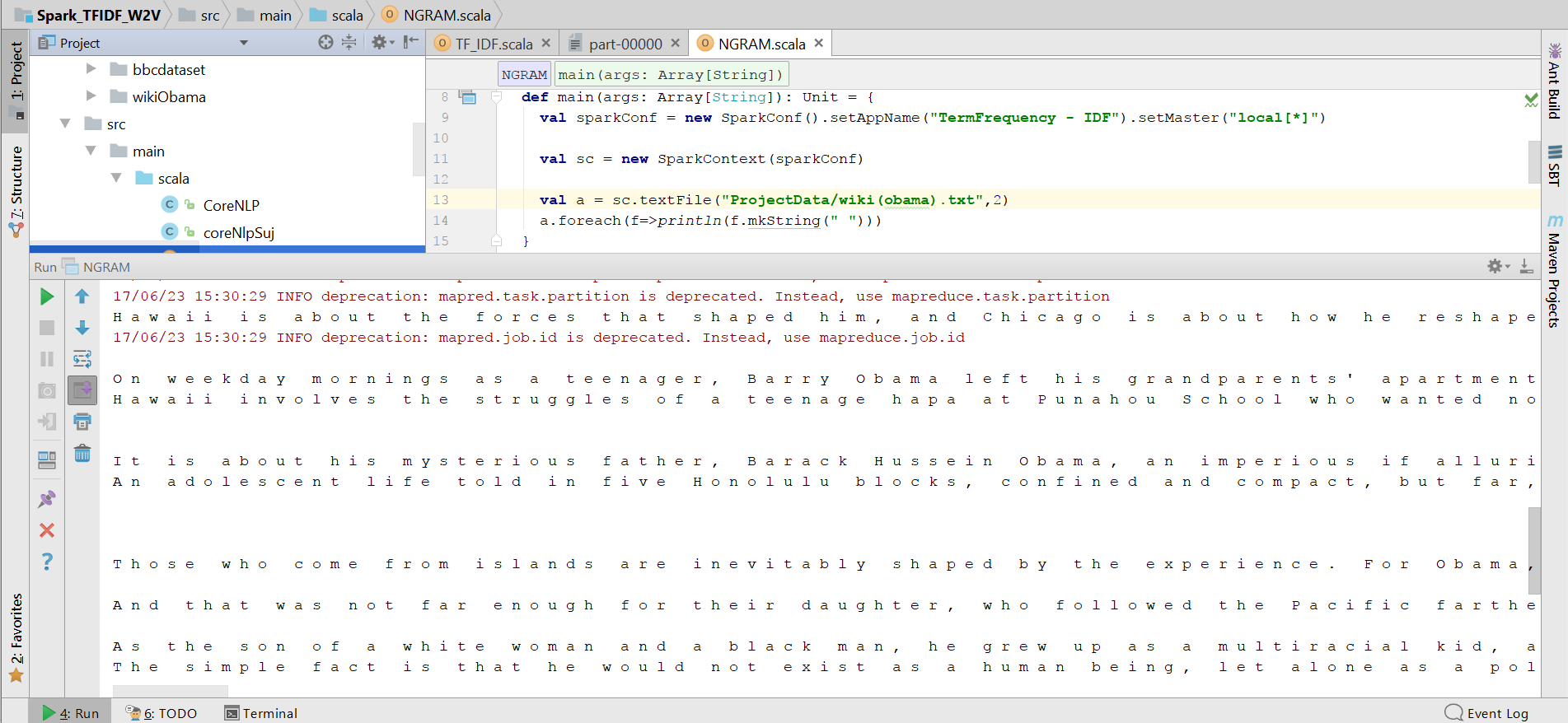
Generating the term frequency for the words in the dataset.



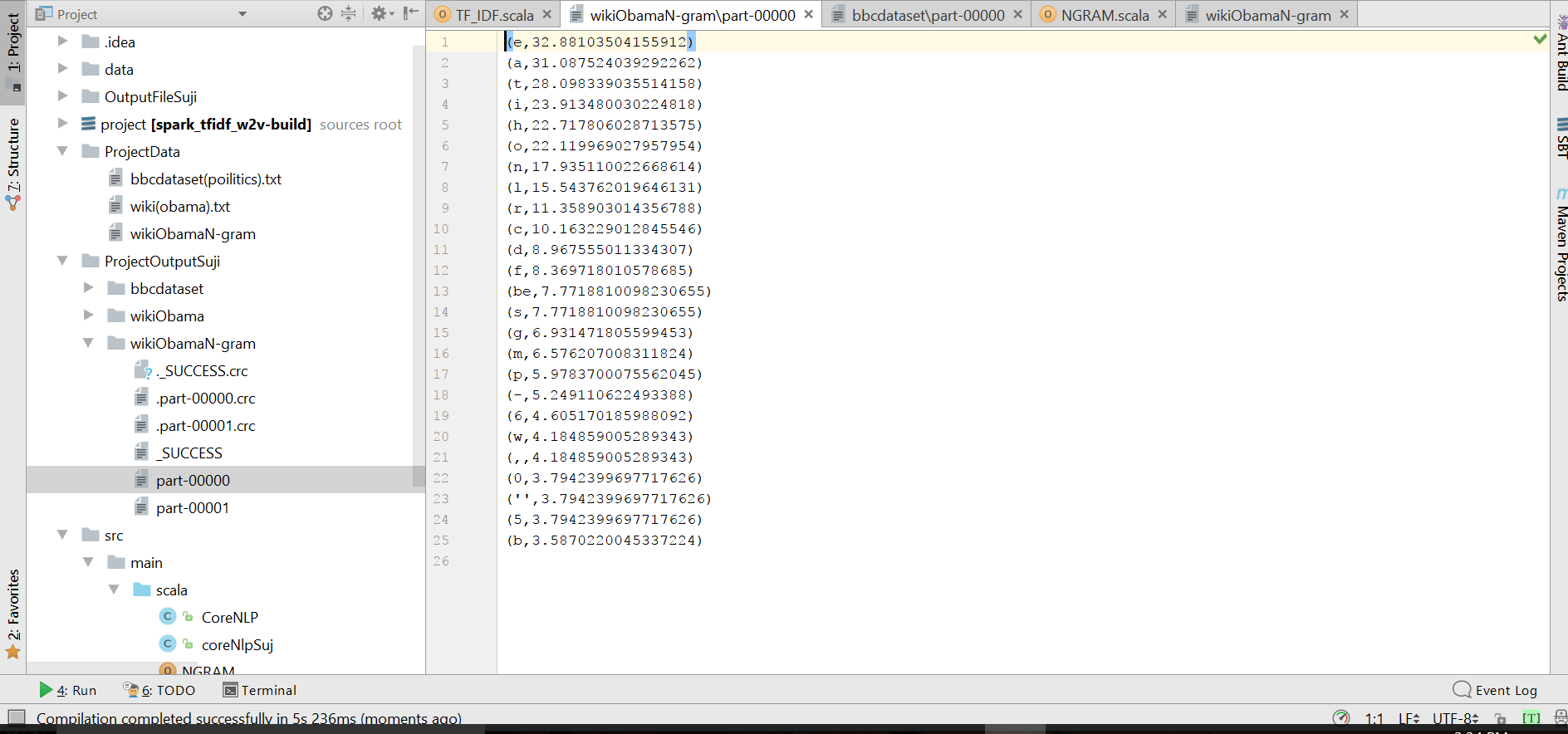
Generated Output.



Generation N-gram for the dataset

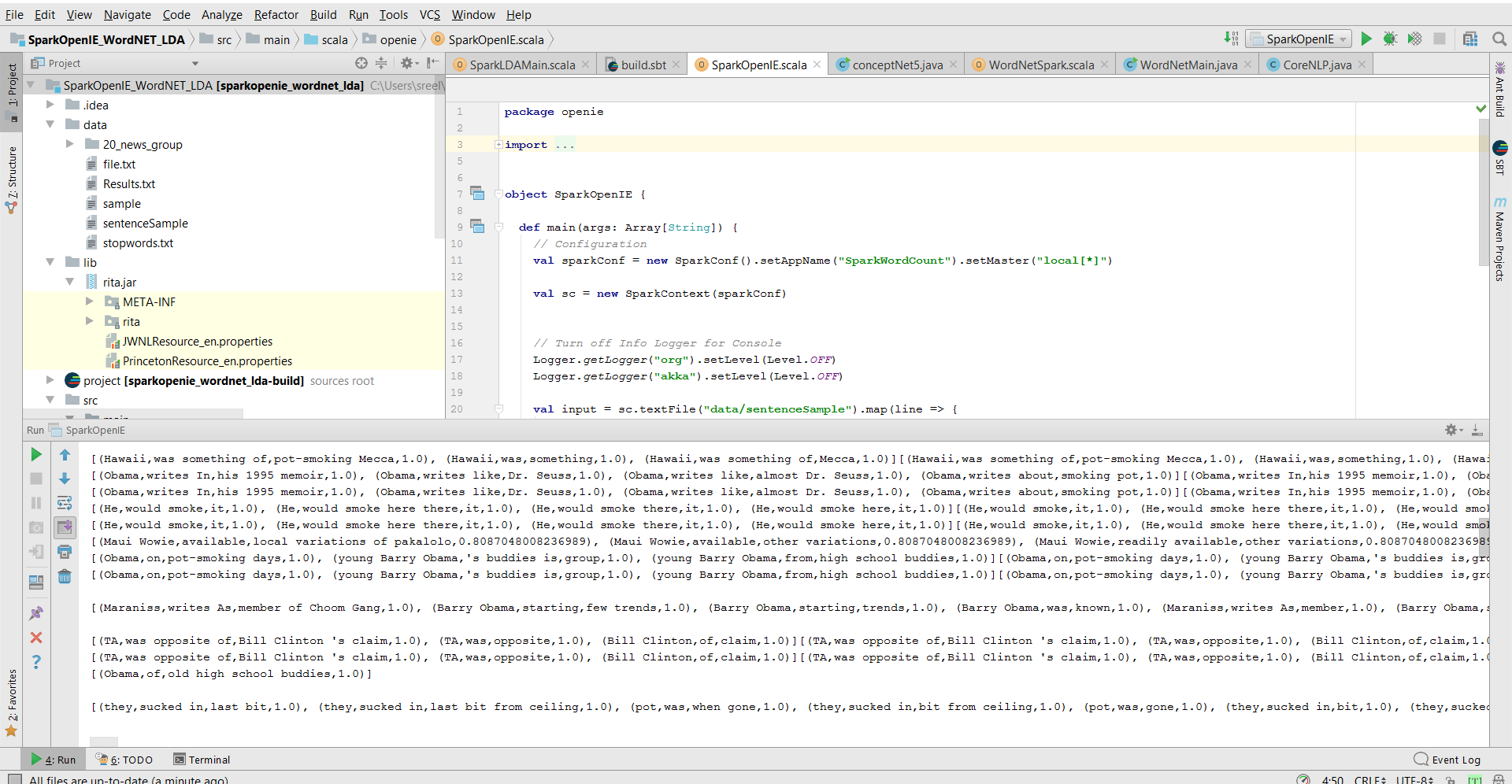


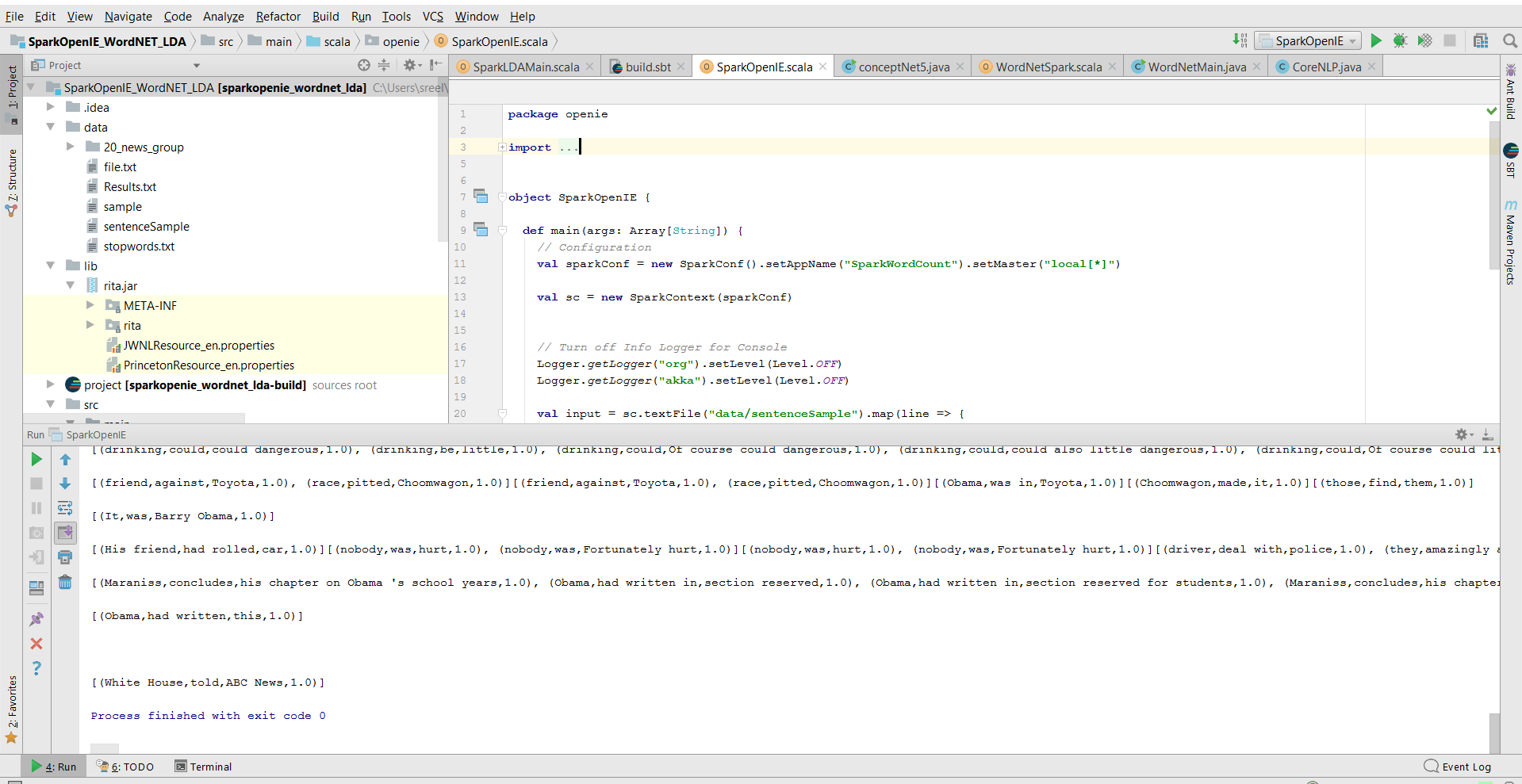
Generating the TF\_IDF for N-gram output.

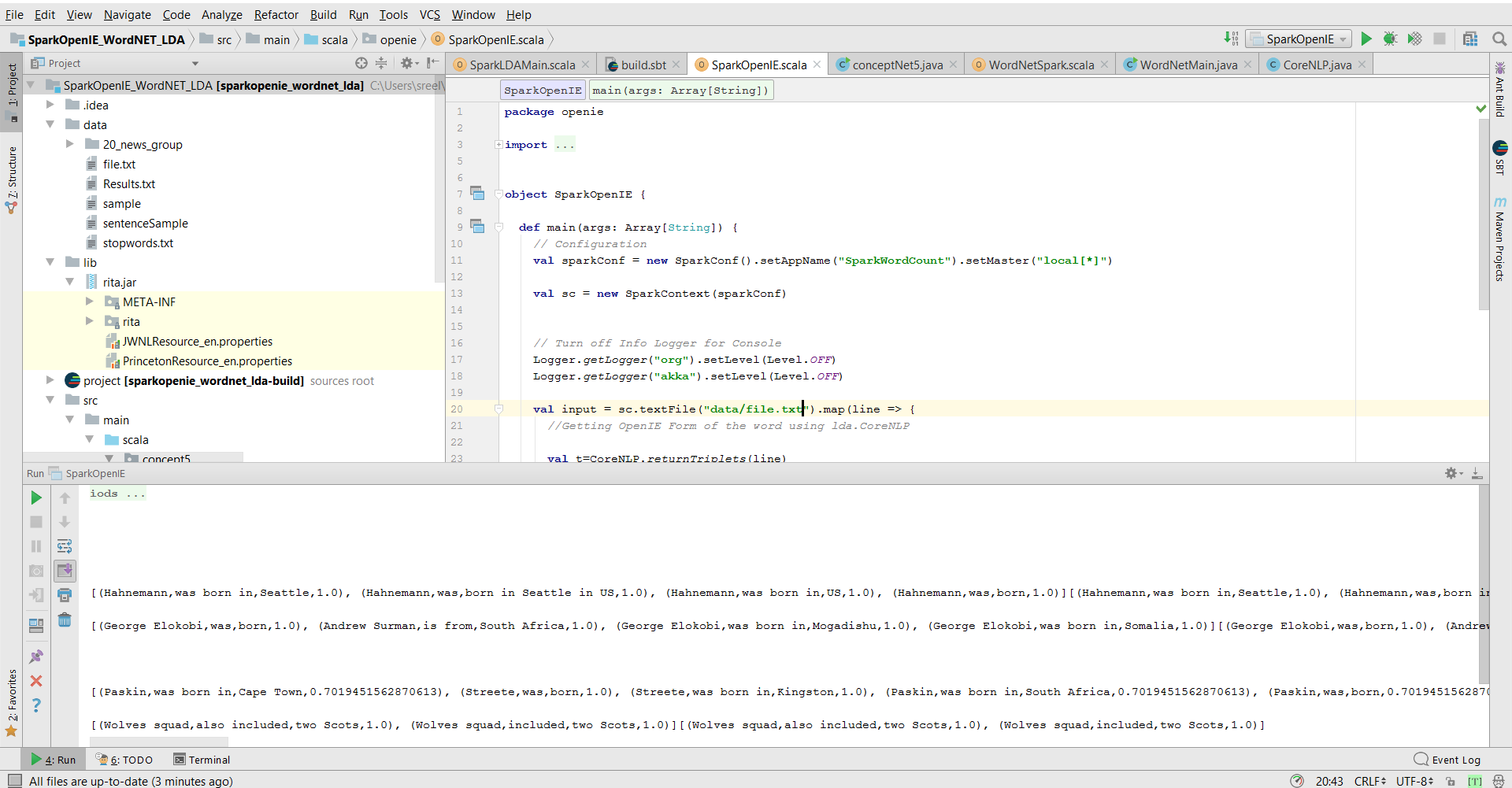


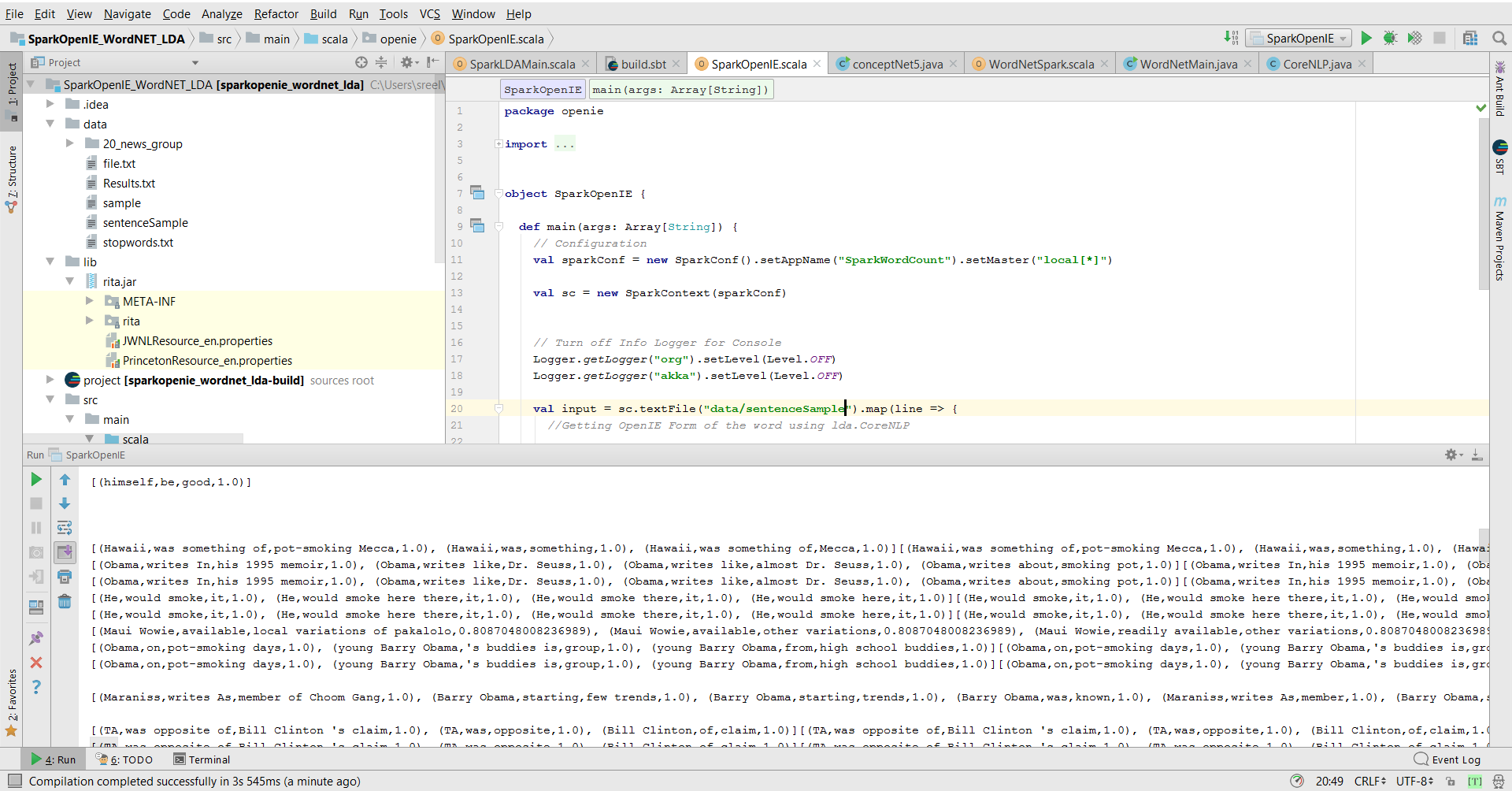
**5.4 Information Extraction**

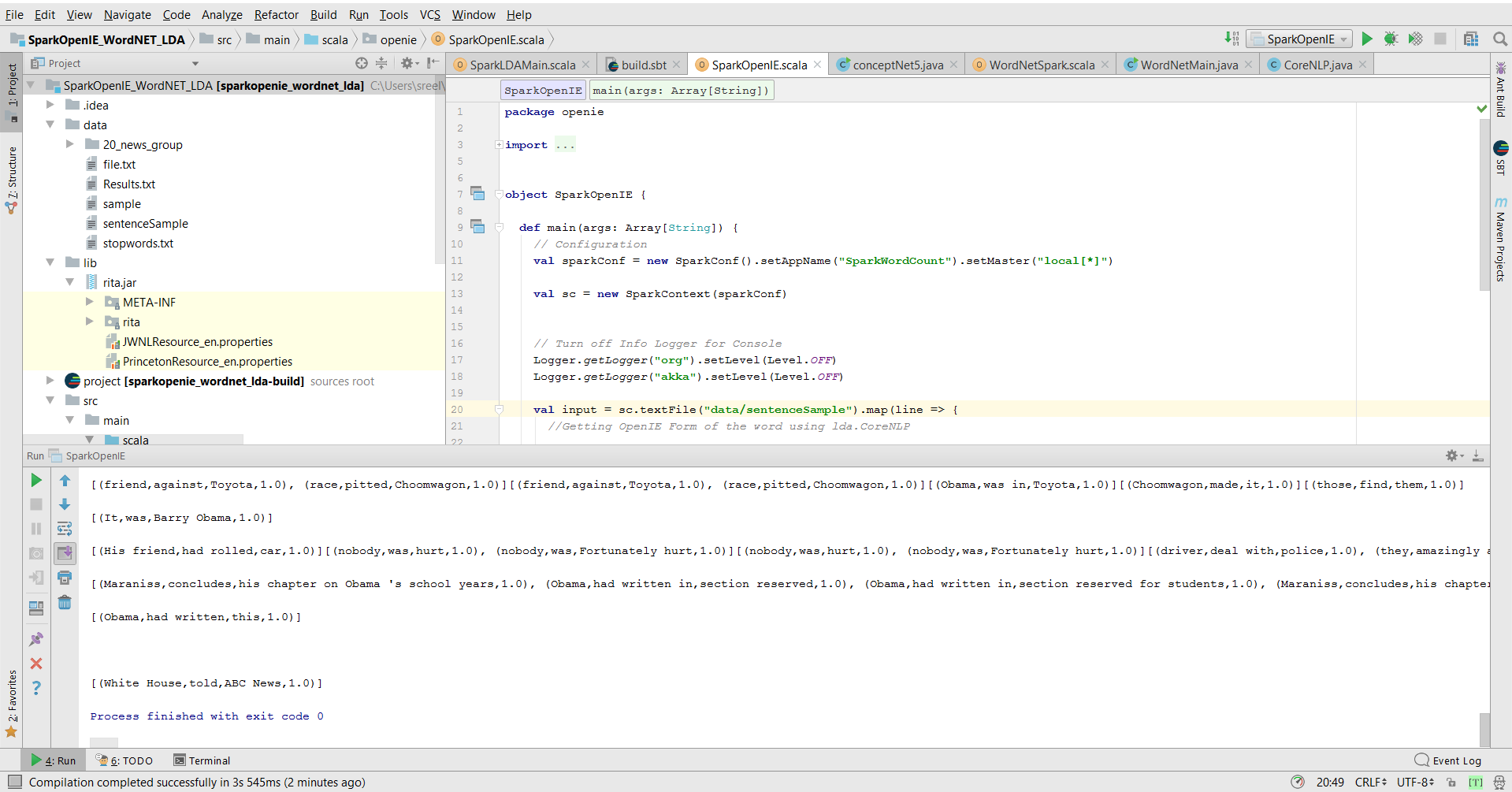
**OpenIE for our dataset:**

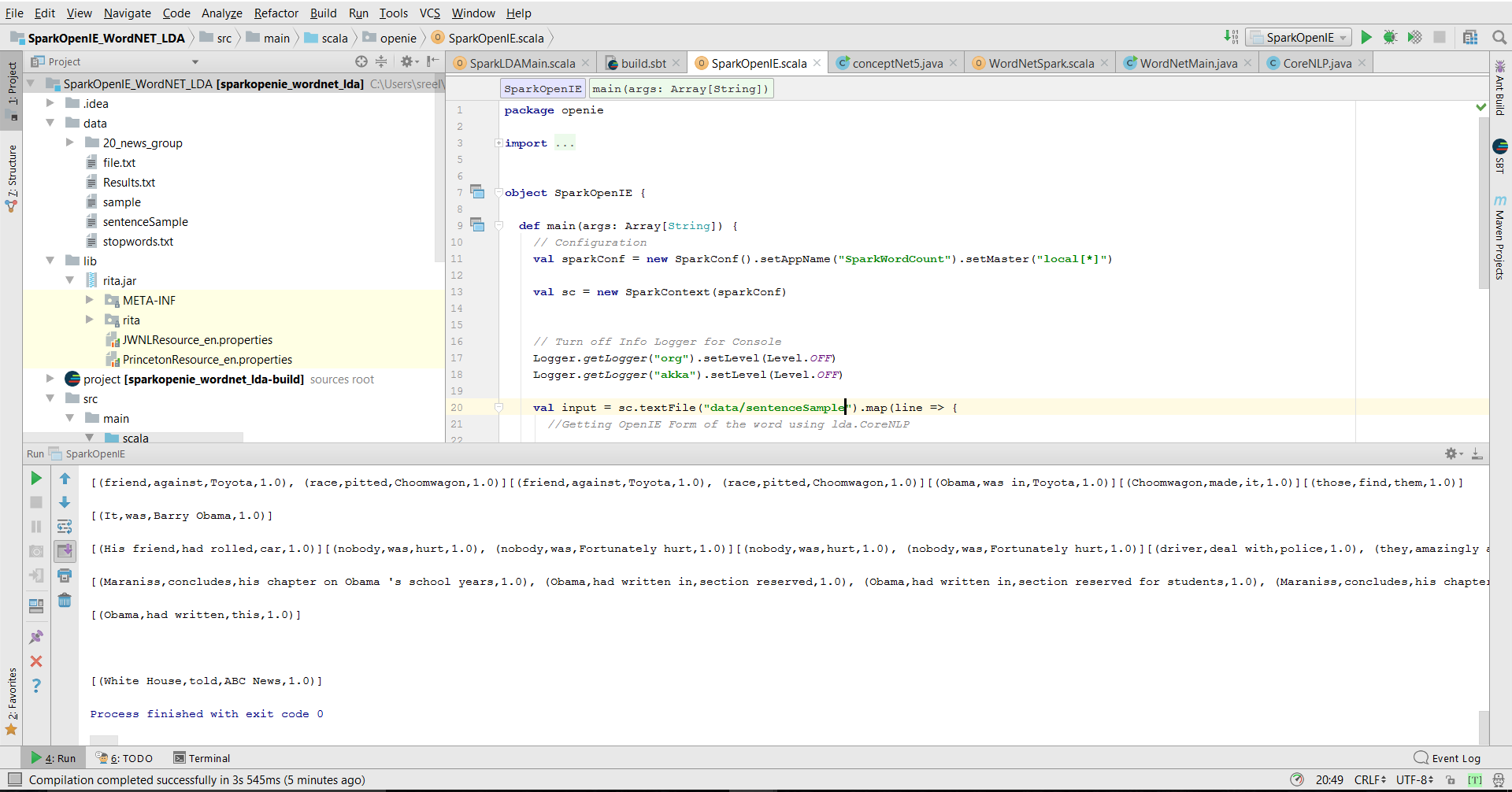




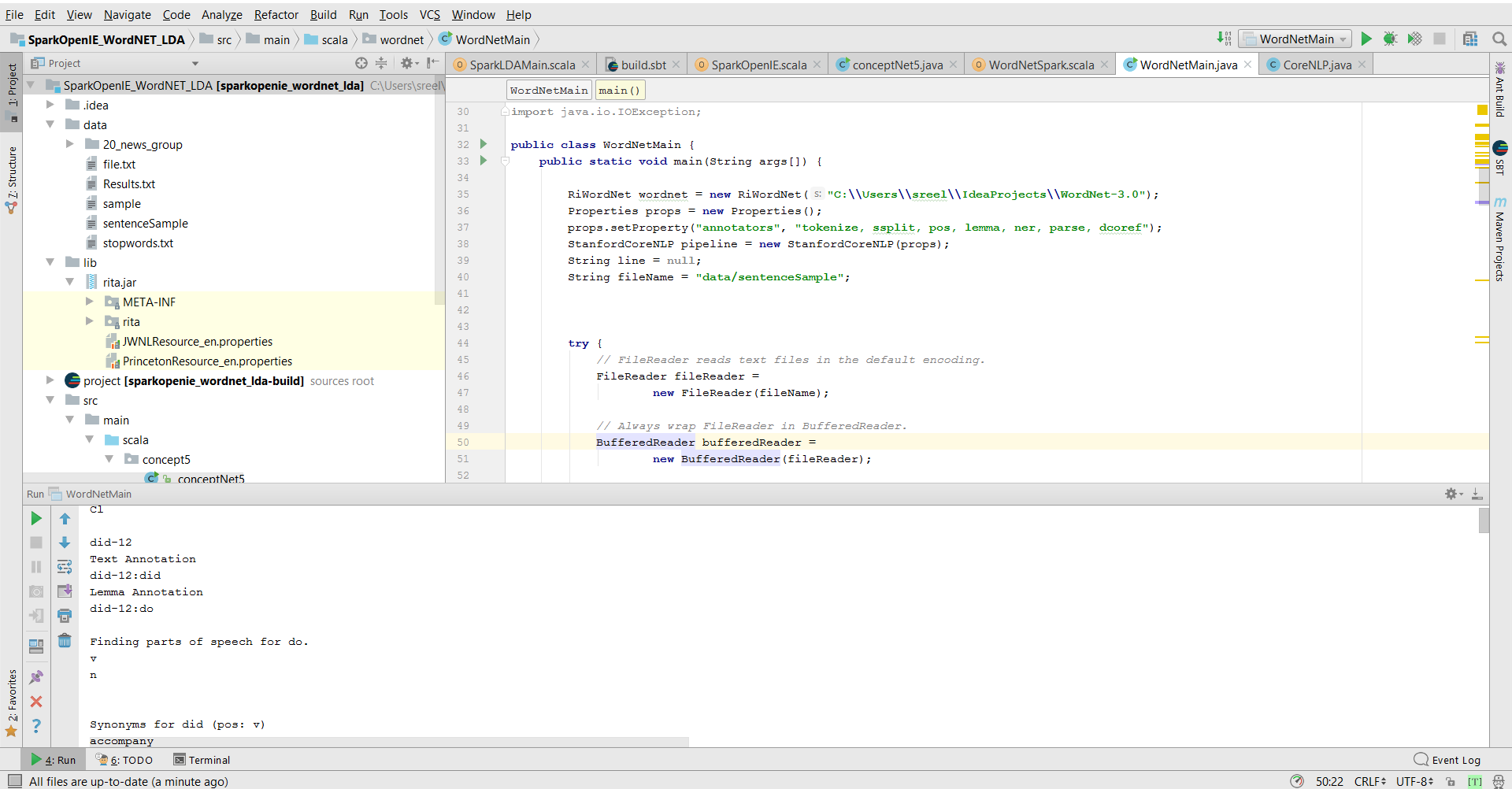


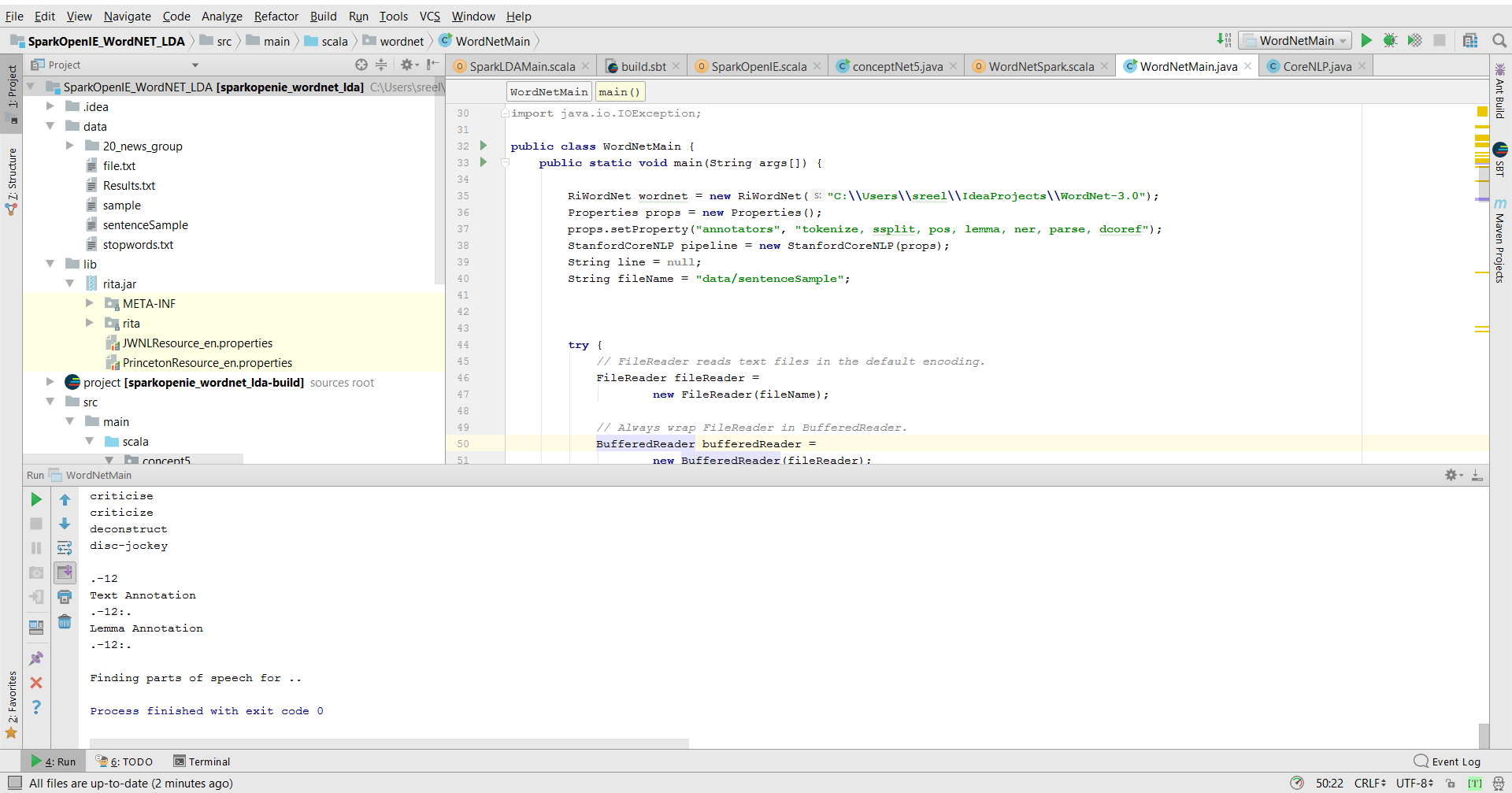


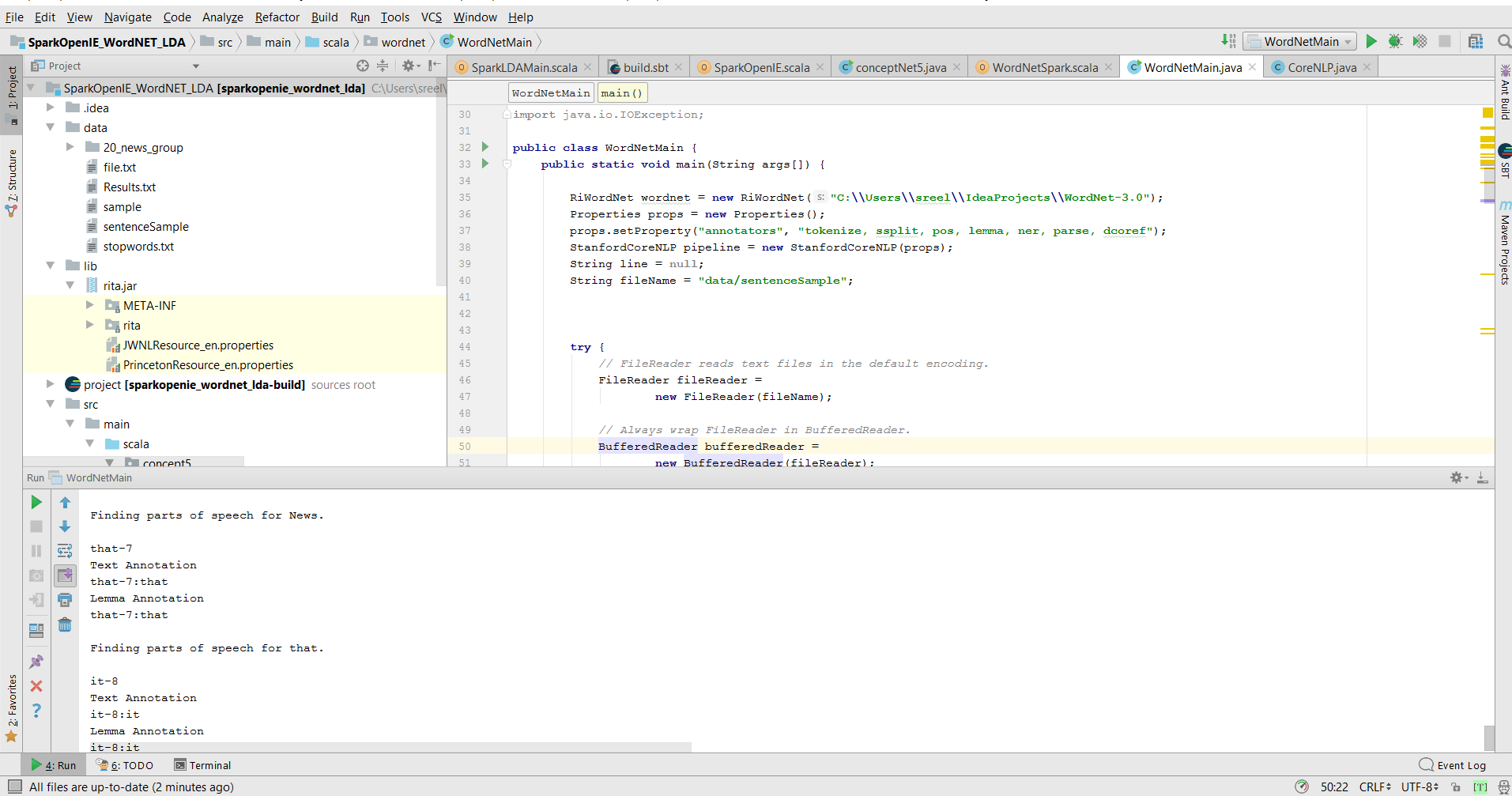


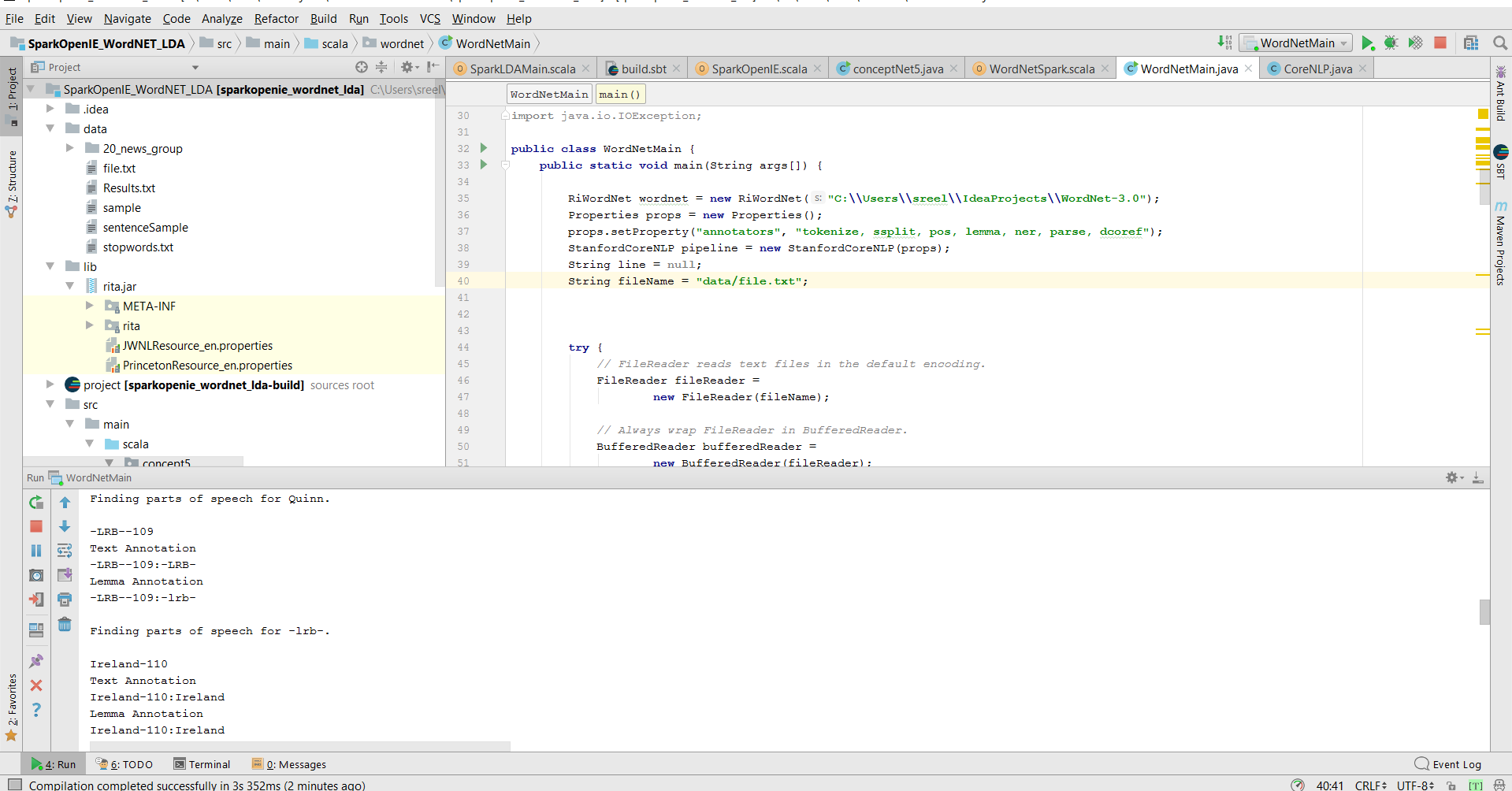


**Wordnet for our dataset**









# 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

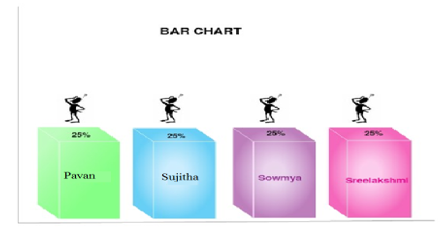
## IDE Used:

* IntelliJ
* PyCharm

## 6.1 Contributors

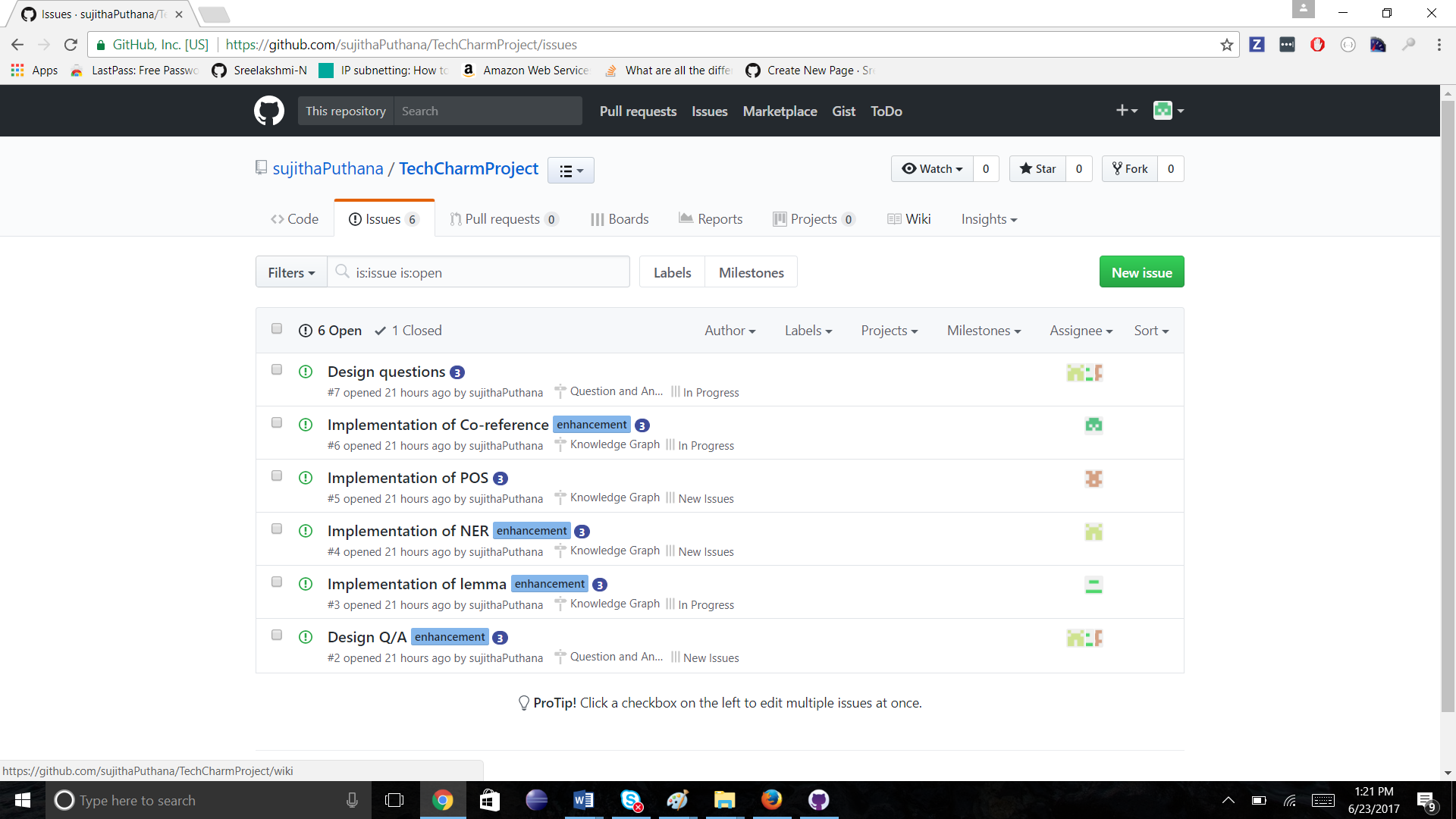
* Jakkepalli, Rama Charan Pavan - **25%**
* Puthana, Sujitha - **25%**
* Yalamanchili, Sowmya **- 25%**
* Nandanamudi, Sreelakshmi **- 25%**

|  |  |  |
| --- | --- | --- |
| Name | Implementation | Documentation |
| Pavan | Question Answer System | Domain, Specific Dataset,  Future Work |
| Sujitha | TF-IDF | Design workflow, Question Answer, Knowledge Graph |
| Sowmya | Core NLP | Project Motivation, Objective, Significance. |
| Sreelakshmi | Named Entity Recognition | Contribution, Milestone, issues creation, Work Completed |



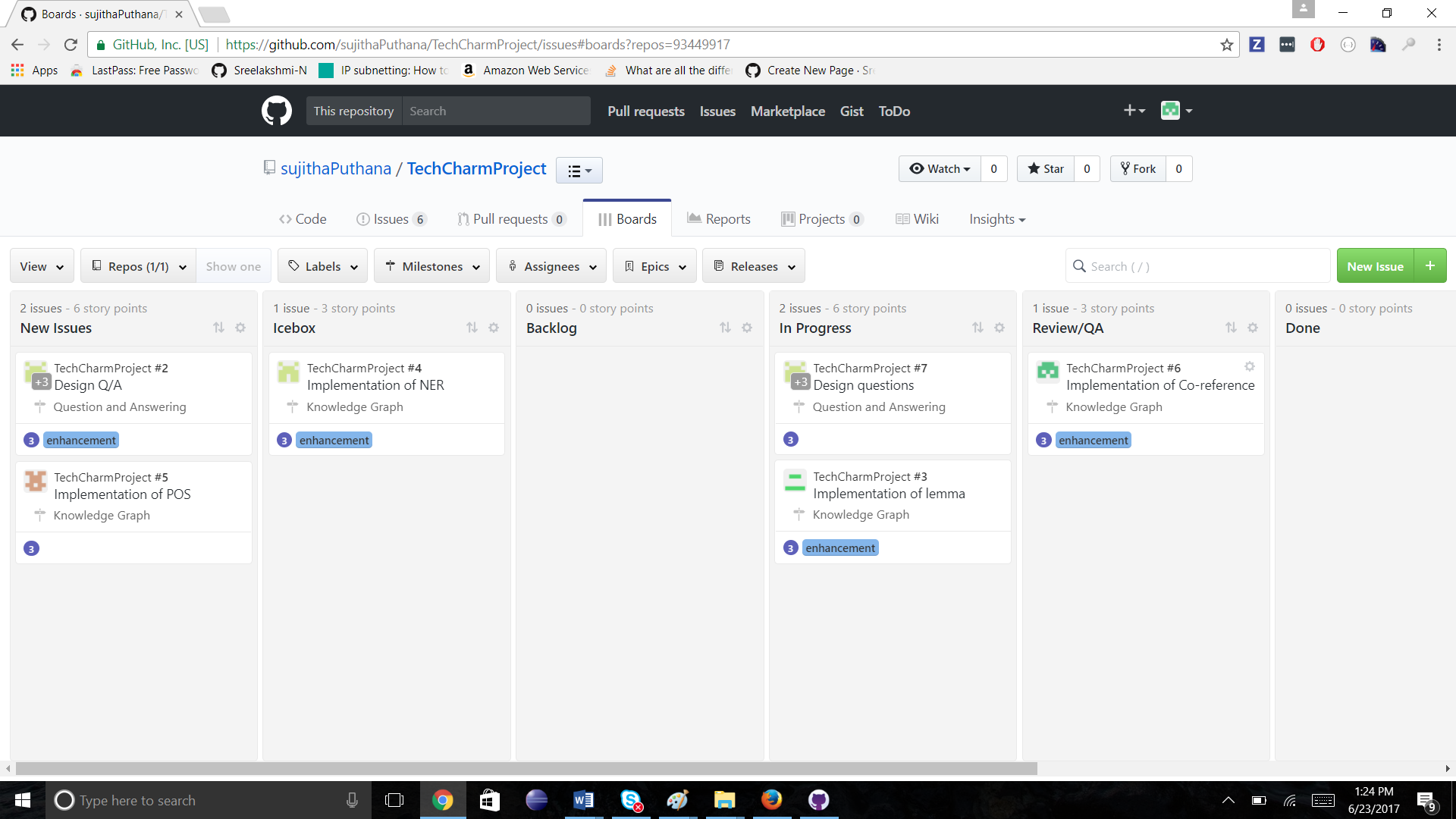
## 6.2 Zen-Hub Screenshots

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.



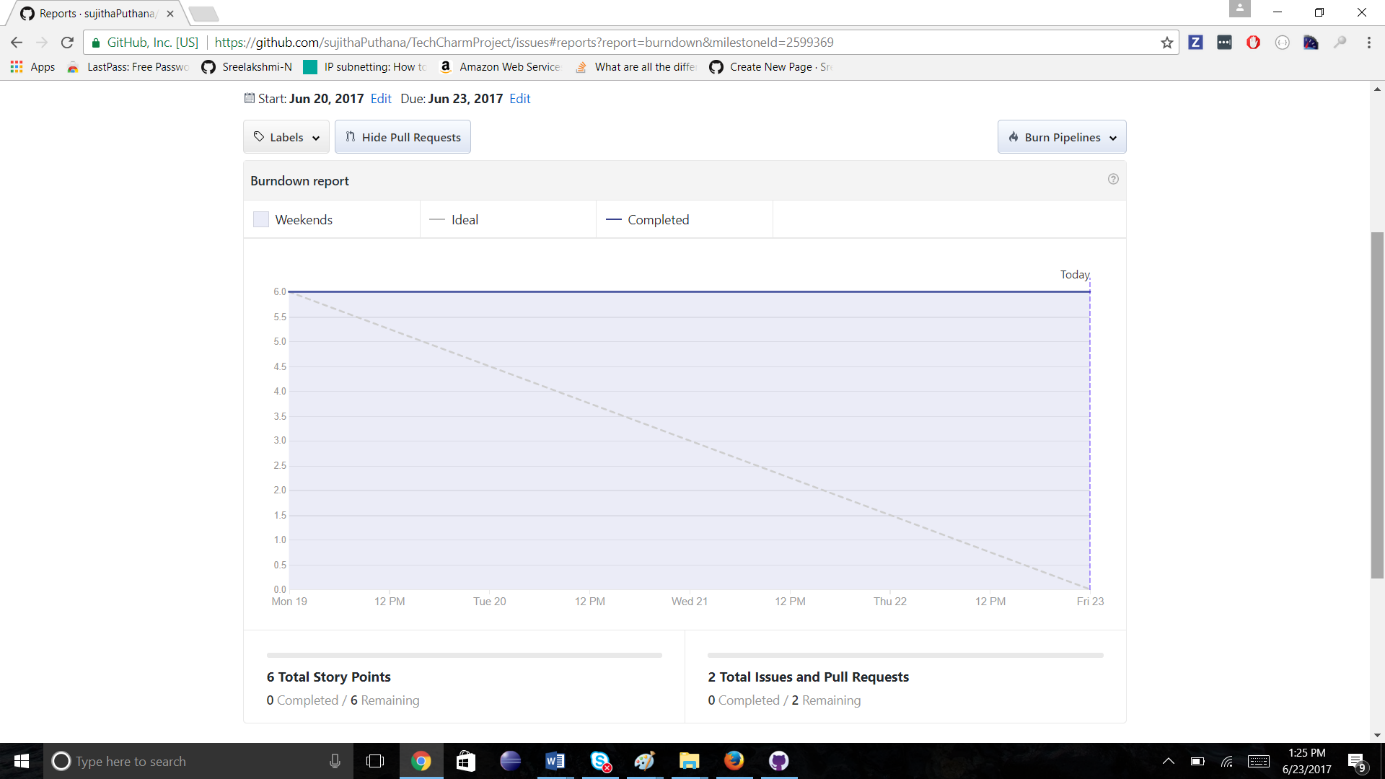
## Project Timeline, Members, and Task Responsibility

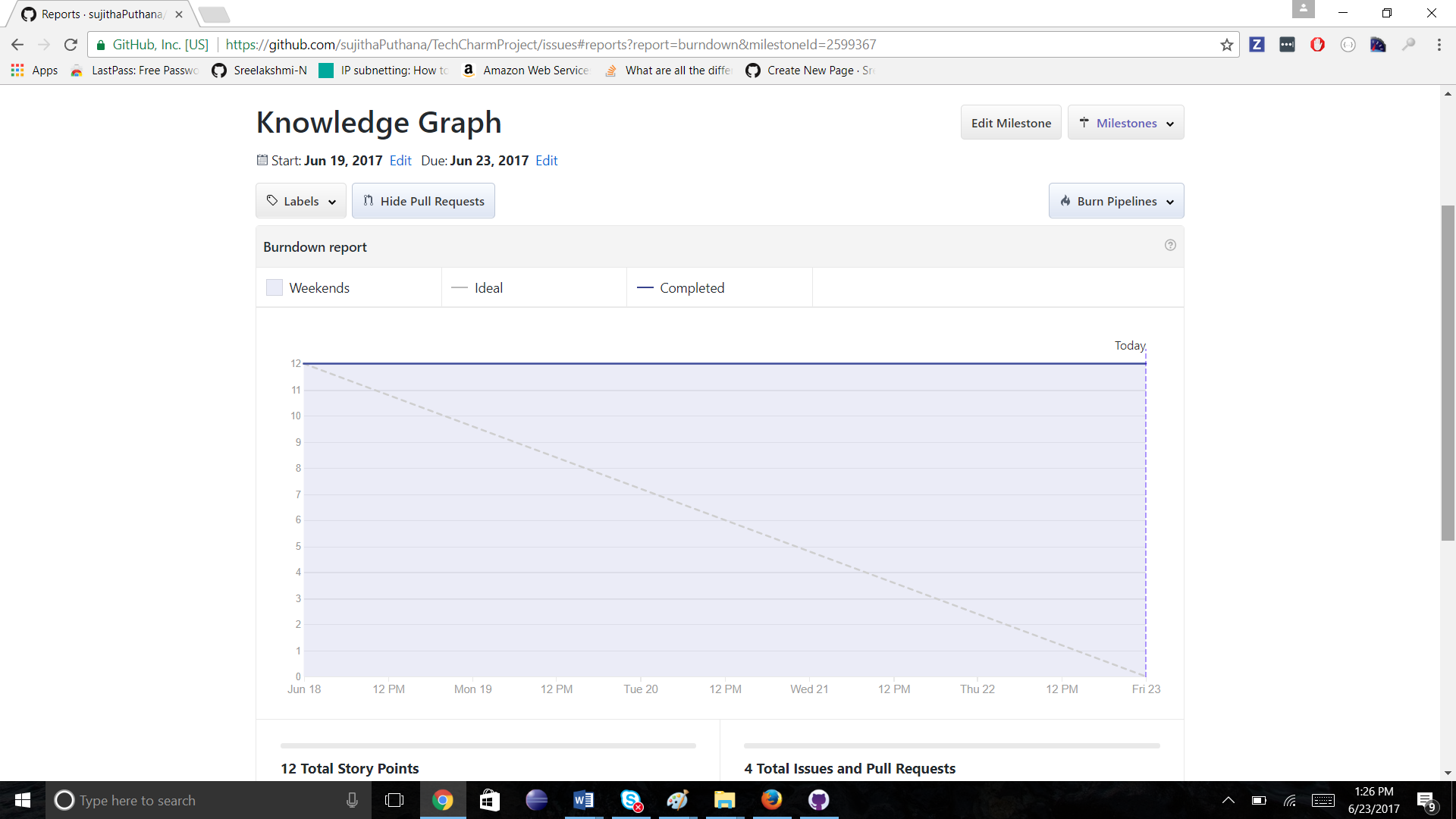
The issues that are registered and current one’s which we are working are updated and can be viewed in GitHub repository. The below screenshot will show you the issues and their respective categorization’s i.e. New issues, Icebox, Backlog, In Progress.



**Burn-Down Chart:**

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





## 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/>