# **CSYE 7245 - Big Data Systems & Intelligence Analytics**

## **Assignment 2 - Part 2 Tutorial**

## 

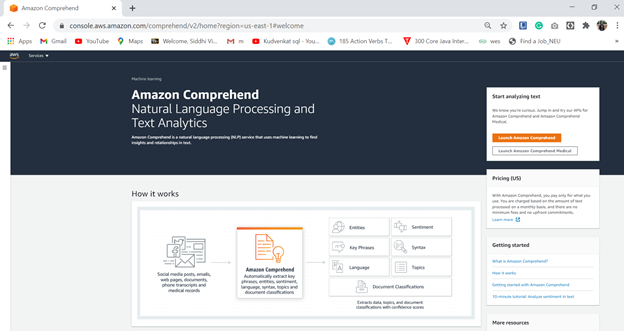
### **Team Information: Team 4**

|  |  |
| --- | --- |
| **Name** | **NUID** |
| Samarth Hadawale | 001053811 |
| Shivendra Shahi | 001393331 |
| Siddhi Prabhu | 001342165 |

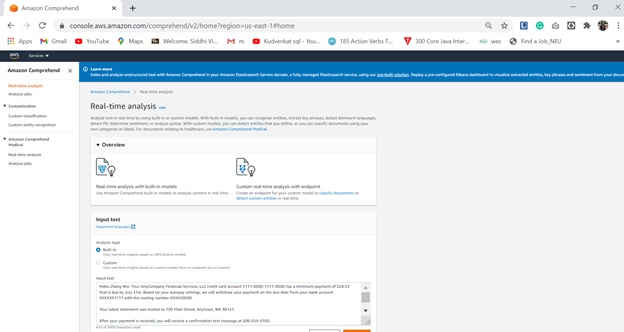
# **Part 1: Detecting and redacting PII using Amazon Comprehend**

## Analyzing text on the Amazon Comprehend console

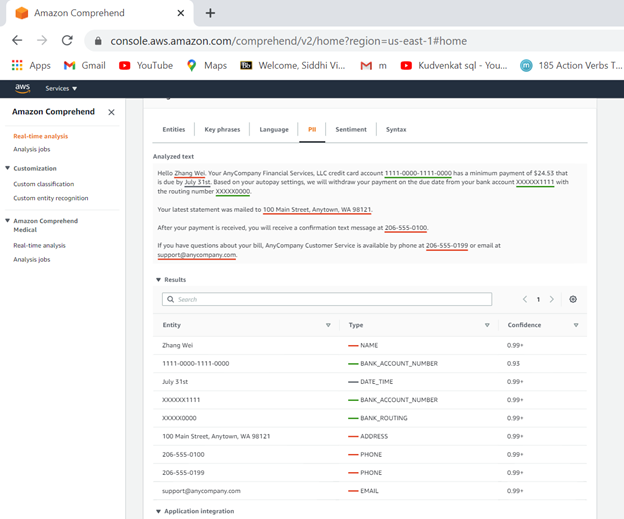
* Launching AWS Comprehend



* After launching amazon comprehend select built-in input text and enter text to analyze in the Input text field



* Analyzing the text -> Viewing PII



## 

## 

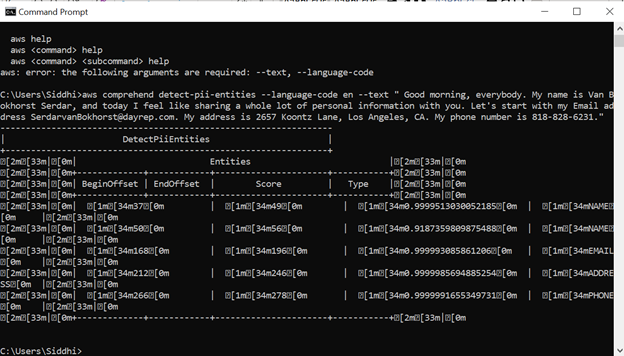
## 

## 

## 

## **Analyzing text via the AWS CLI**

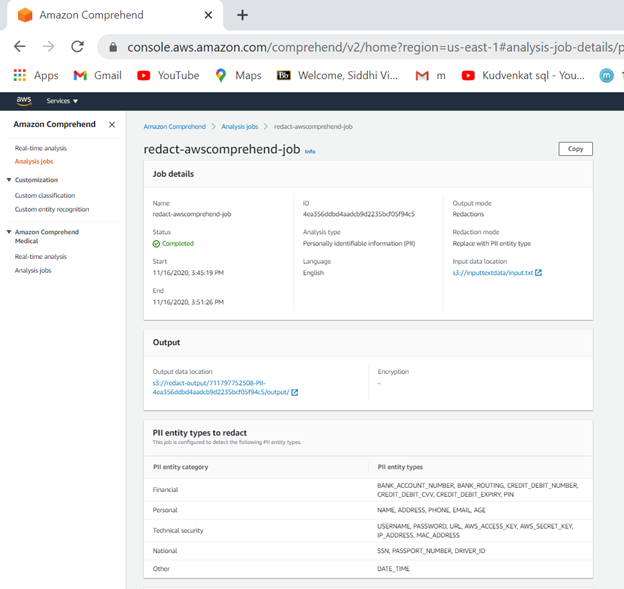
* Realtime text analysis via the AWS CLI



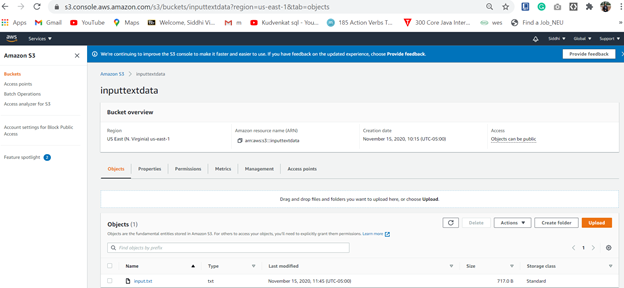
### Asynchronous PII redaction batch processing on the Amazon Comprehend console

**Redaction mode – Replace with PII entity type**

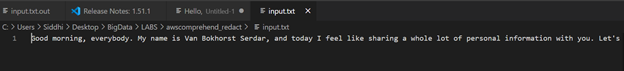
* Created job



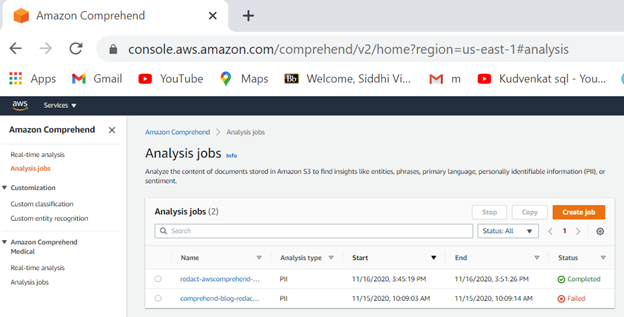
* Input S3 bucket



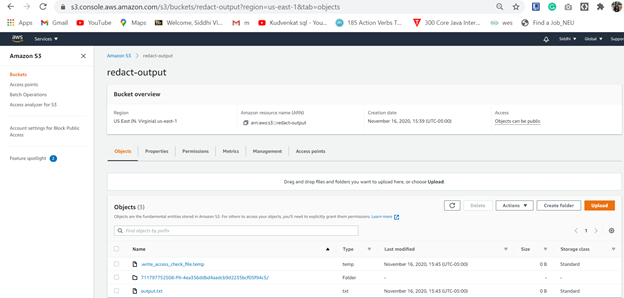
* Input text file in S3 bucket



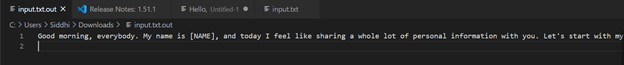
* Job successfully completed



* S3 output bucket



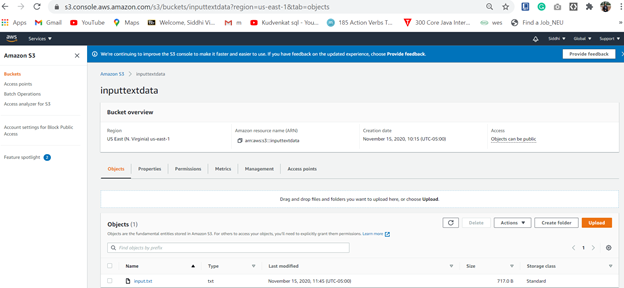
* Output text file

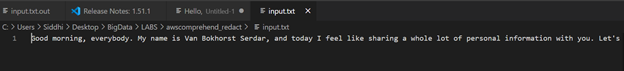




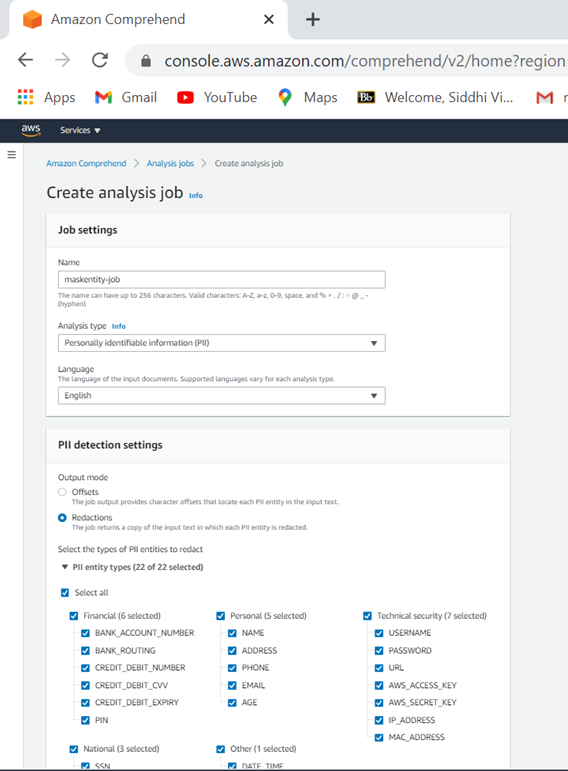
## Redaction mode – Replace with character

* Input text file from AWS S3 Bucket

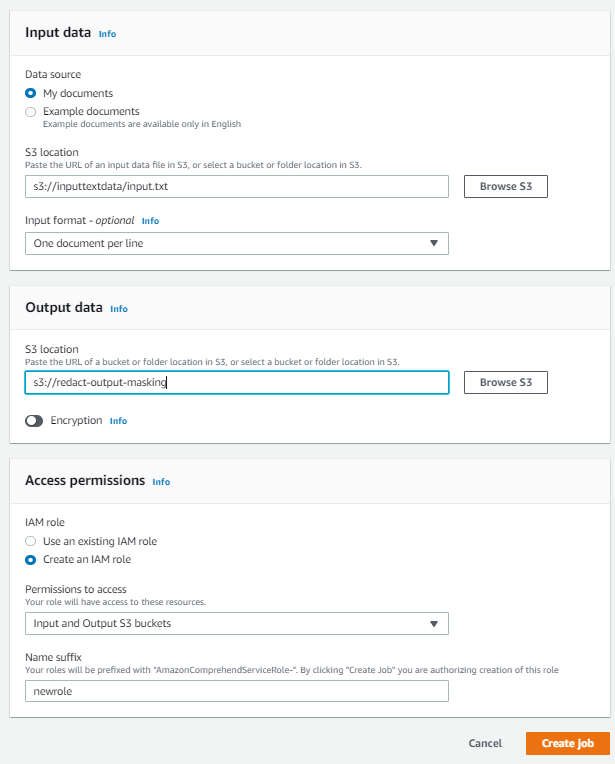




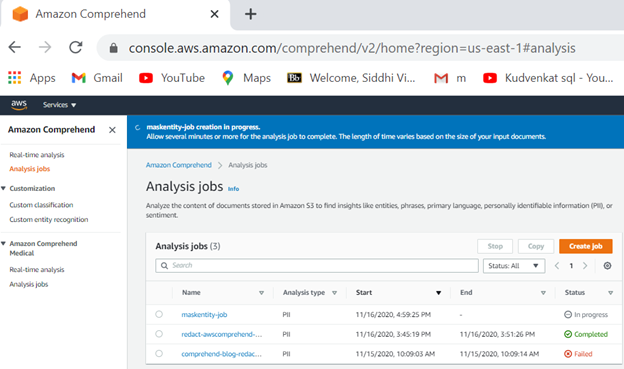
Creating job



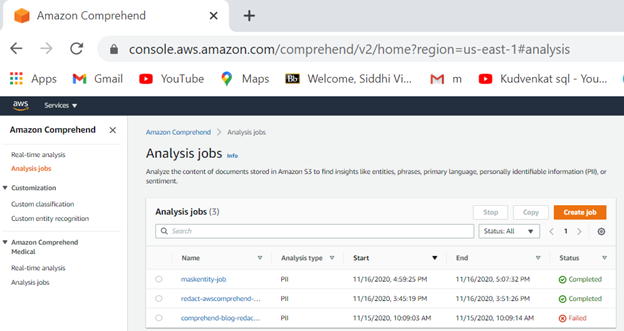


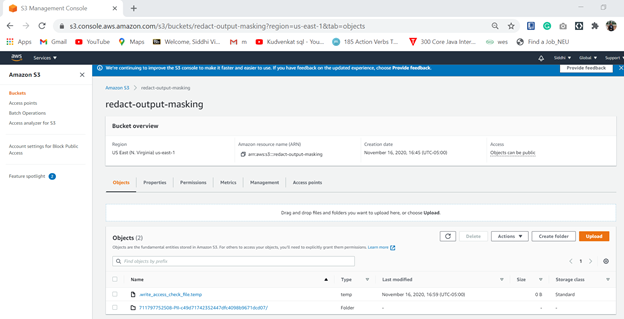


Executing the job

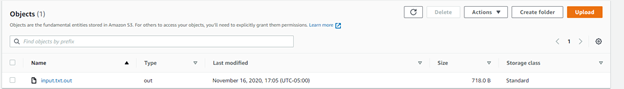


Job successfully completed

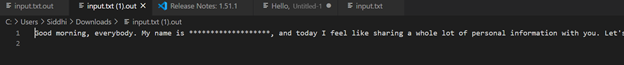


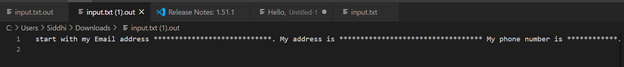


Output text file with masked PII entities



Masked entities

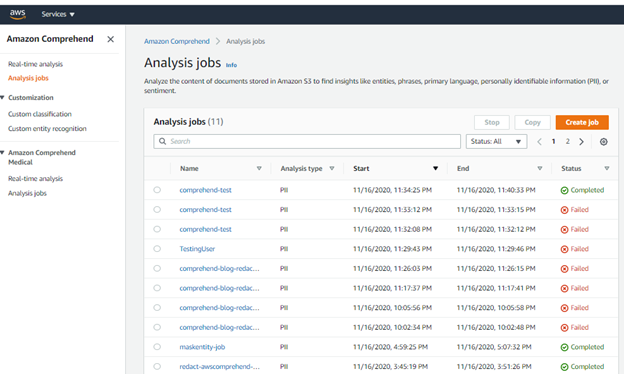


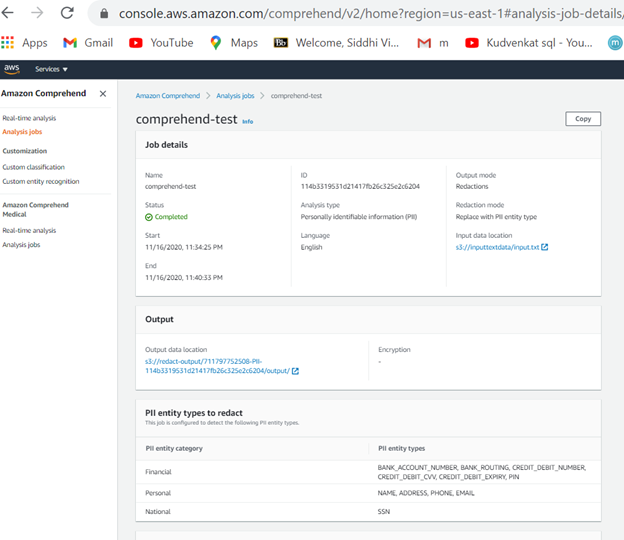


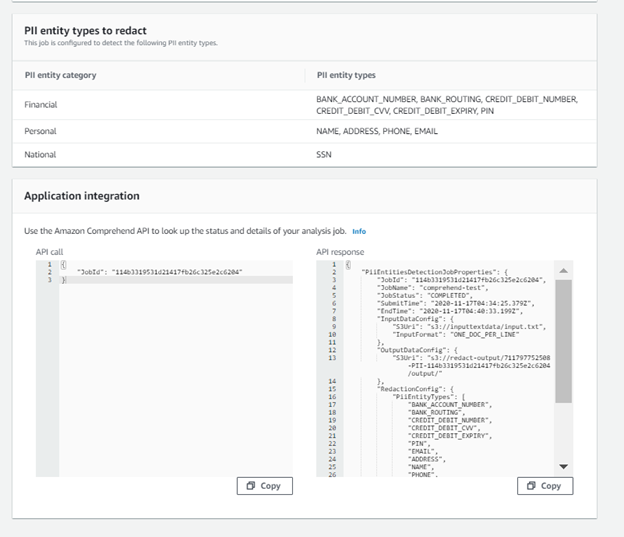
### Asynchronous PII redaction batch processing via the AWS CLI



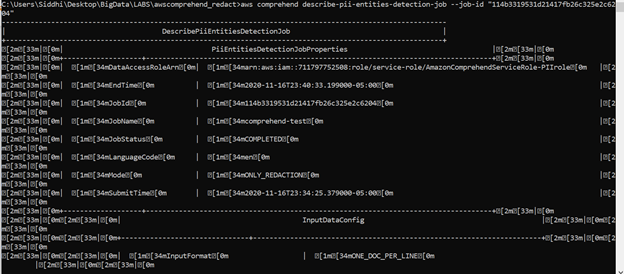
Job completed successfully

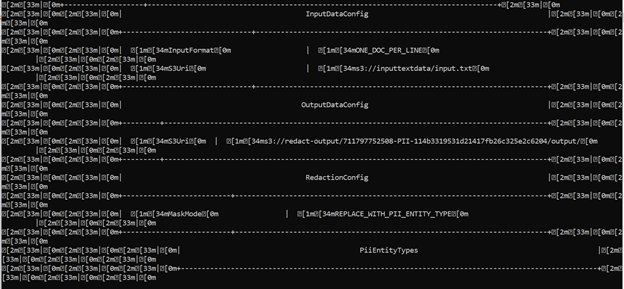


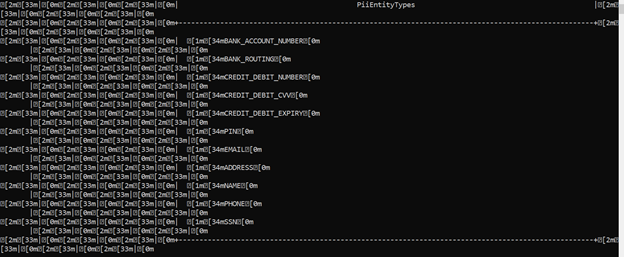




monitor the job request





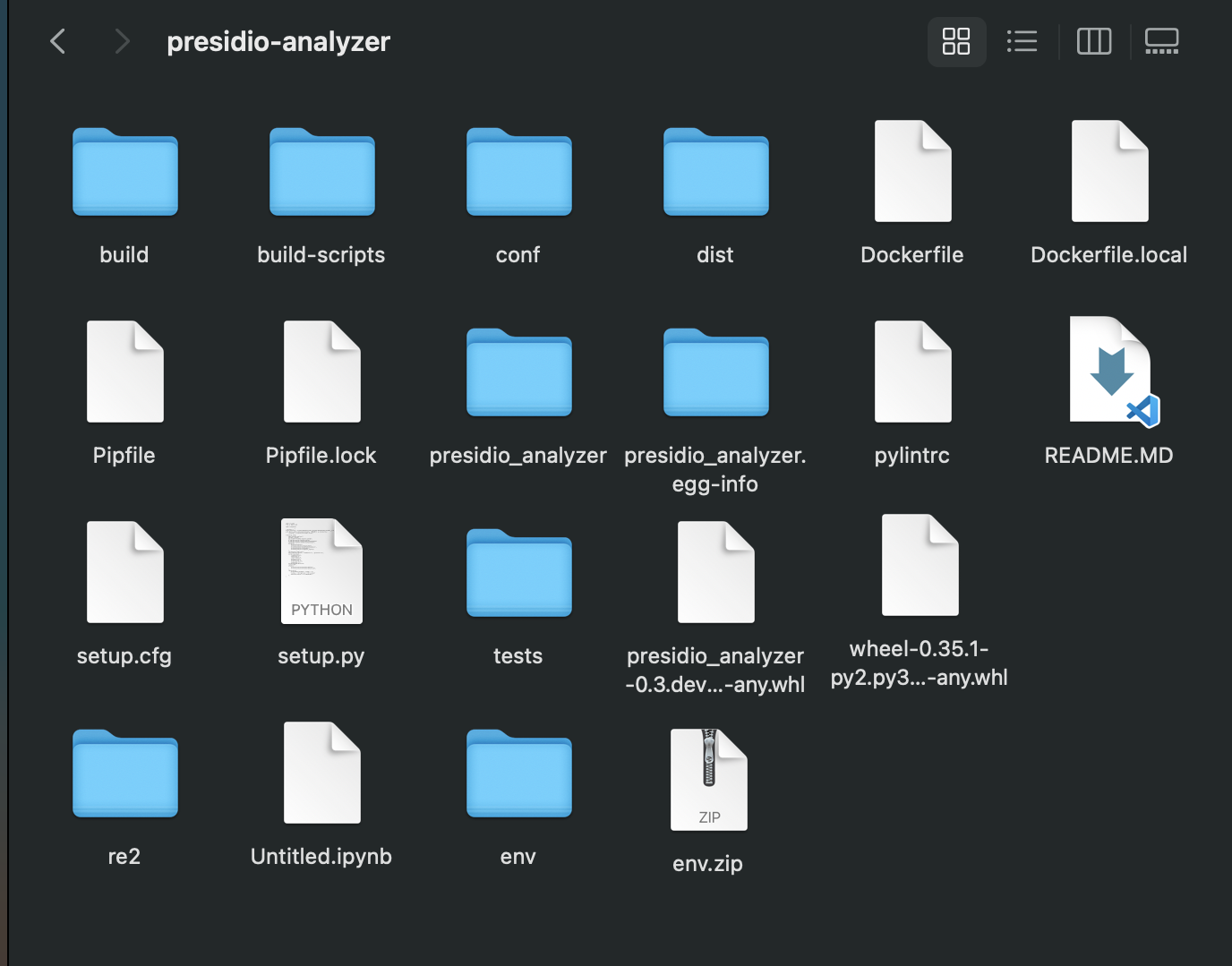


# 

# Part 2 : Presidio Analyzer as a Python Package.

### Step 1 : Creating a Virtual environment by installing all the dependencies & then the Wheel file in the folder

**python setup.py bdist\_wheel**

****

Here we can see that the Wheel file was generated.

### Step 2 : Installing the Wheel file

**pip install wheel**

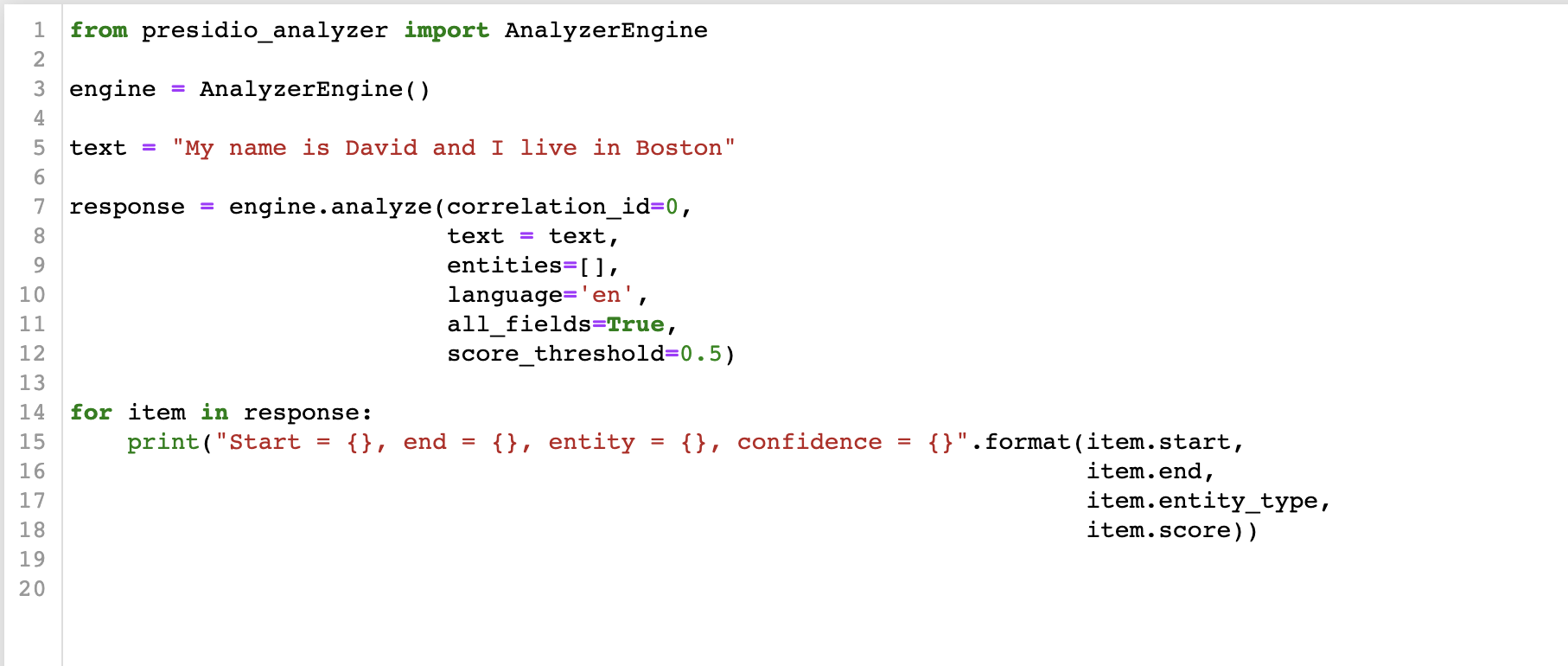
Step 3 : Installing the Presidio Analyzer Wheel file

**pip install ‘WHEEL FILE PATH’**

### Step 4 : Installing the Spacy model from Github.

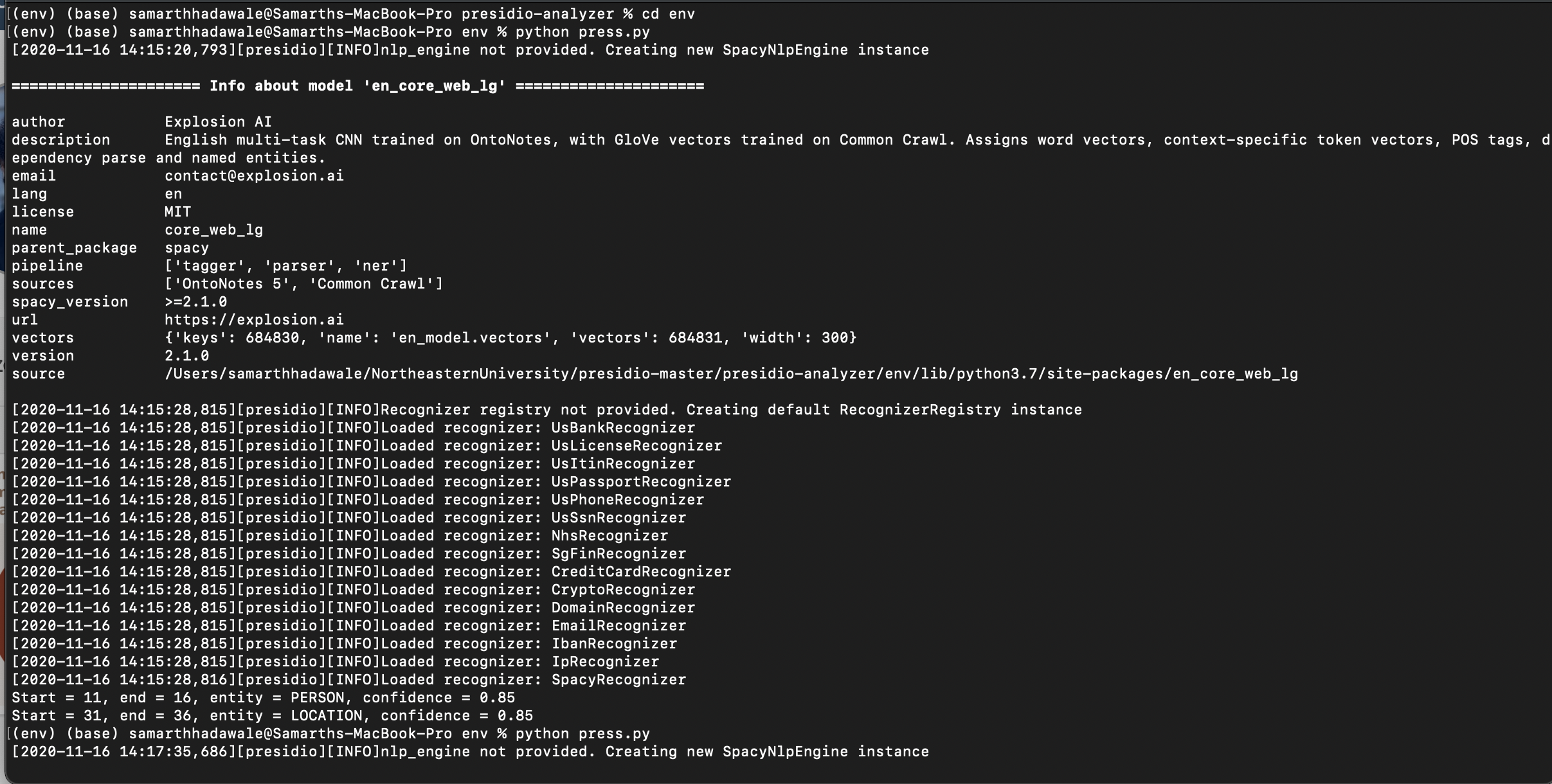
**pip install** [**https://github.com/explosion/spacy-models/releases/download/en\_core\_web\_lg-2.1.0/en\_core\_web\_lg-2.1.0.tar.gz**](https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-2.1.0/en_core_web_lg-2.1.0.tar.gz)

### Step 5 : Writing a code in press.py file to test the presidio python package & it’s working



Here we are passing the text which is given in the above code.   
Now our Presidio python package should be able to figure out those entities and give us back the required output.

### Step 6 : Outcome



Here we can see that David is a Person & Boston is a Location so we have got the required output back after providing the text file to our Presidio Python package.