Project: Text Redaction

In this project, we created a system which detects "sensitive" items in the text. Once identified, the system will redact these sensitive items such as names, dates, addresses (city, state and country). After redacting the sensitive items, the program will save redacted file as a new file with the .txt extension in the current folder. We are using SpaCy module from Natural Language Processing with python.

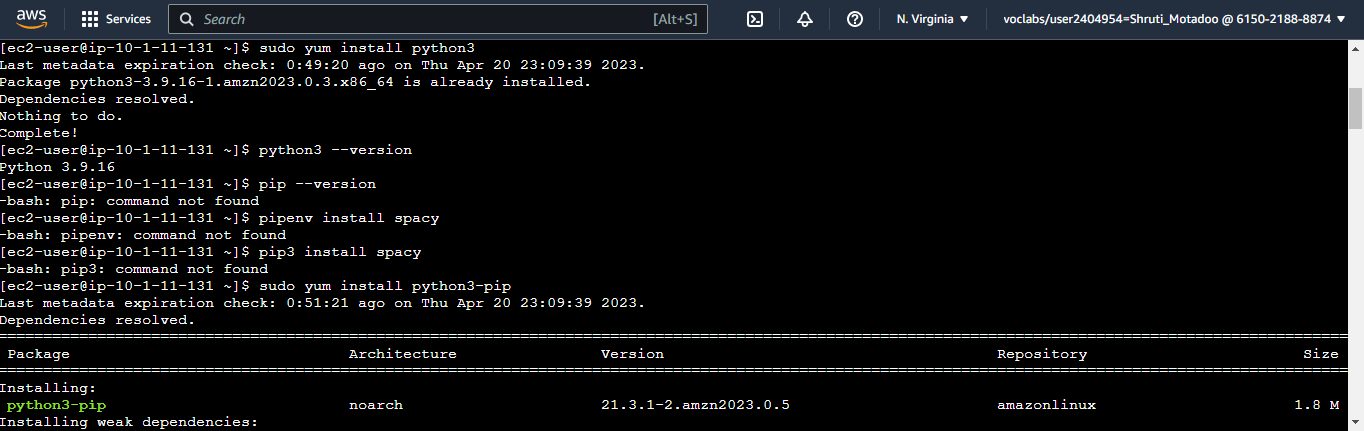
We created an EC2 instance and connected to the instance.

[ec2-user@ip-10-1-11-131 ~]$ python3 --version

Python 3.9.16

Here, we are installing pip.

[ec2-user@ip-10-1-11-131 ~]$ sudo yum install python3-pip



Text

Description automatically generated

Then, we install spacy using the command pip3 install spacy.

Text

Description automatically generated

Now, we download en\_core\_web\_sm module using below command,

python3 -m spacy download en\_core\_web\_sm

Text

Description automatically generated

Text

Description automatically generated

Created a text file using vim data.txt : Input data File



A computer screen capture

Description automatically generated with low confidence

Created a python file using vim seppro.py : It redacts name, address(city, state and country) and date. It replaces name, address(city, state and country) and date with [REDACTED].



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

To execute we run the following command,

python3 seppro.py

Output file redacted\_data.txt is created with redacted name, address and date.



A screenshot of a computer

Description automatically generated with medium confidence