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.

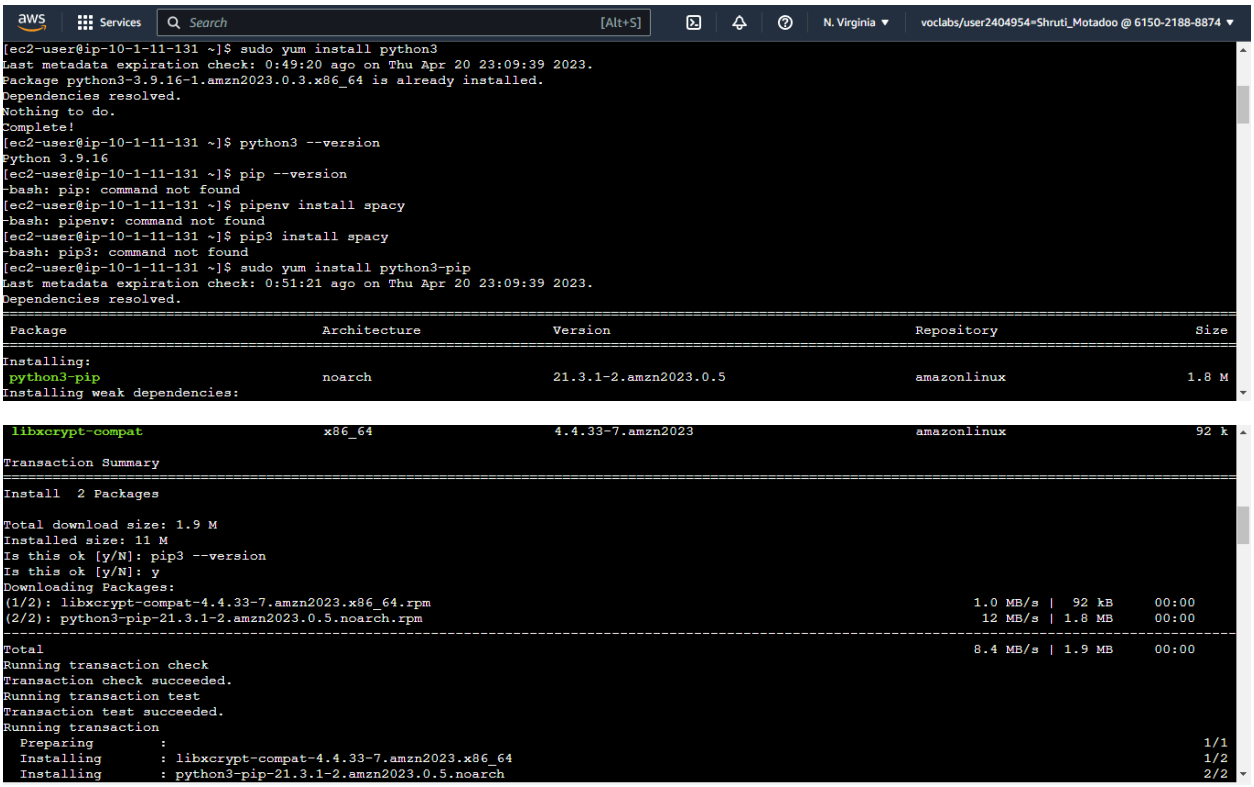
We created an EC2 instance and connected to the instance.

[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



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

```
Transaction test succeeded.
Running transaction
  Preparing        :                                                                                                    1/1
  Installing       : libxcrypt-compat-4.4.33-7.amzn2023.x86_64                                                          1/2
  Installing       : python3-pip-21.3.1-2.amzn2023.0.5.noarch                                                           2/2
  Running scriptlet: python3-pip-21.3.1-2.amzn2023.0.5.noarch                                                           2/2
  Verifying        : libxcrypt-compat-4.4.33-7.amzn2023.x86_64                                                          1/2
  Verifying        : python3-pip-21.3.1-2.amzn2023.0.5.noarch                                                           2/2

Installed:
  libxcrypt-compat-4.4.33-7.amzn2023.x86_64                    python3-pip-21.3.1-2.amzn2023.0.5.noarch

Complete!
[ec2-user@ip-10-1-11-131 ~]$ pip3 install spacy
Defaulting to user installation because normal site-packages is not writeable
Collecting spacy
  Downloading spacy-3.5.2-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (6.6 MB)
     |████████████████████████████████| 6.6 MB 5.5 MB/s
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/lib/python3.9/site-packages (from spacy) (2.25.1)
Collecting smart-open<7.0.0,>=5.2.1
  Downloading smart_open-6.3.0-py3-none-any.whl (56 kB)
     |████████████████████████████████| 56 kB 8.9 MB/s
Collecting srsly<3.0.0,>=2.4.3
```

Now, we download en_core_web_sm module using below command,

python3 -m spacy download en_core_web_sm

```
  Downloading click-8.1.3-py3-none-any.whl (96 kB)
     |████████████████████████████████| 96 kB 10.5 MB/s
Requirement already satisfied: MarkupSafe>=0.23 in /usr/lib64/python3.9/site-packages (from jinja2->spacy) (1.1.1)
Installing collected packages: typing-extensions, catalogue, srsly, pydantic, numpy, murmurhash, cymem, click, wasabi, typer, smart-open, preshed, pac
kaging, confection, blis, tqdm, thinc, spacy-loggers, spacy-legacy, pathy, langcodes, spacy
Successfully installed blis-0.7.9 catalogue-2.0.8 click-8.1.3 confection-0.0.4 cymem-2.0.7 langcodes-3.3.0 murmurhash-1.0.9 numpy-1.24.2 packaging-23.
1 pathy-0.10.1 preshed-3.0.8 pydantic-1.10.7 smart-open-6.3.0 spacy-3.5.2 spacy-legacy-3.0.12 spacy-loggers-1.0.4 srsly-2.4.6 thinc-8.1.9 tqdm-4.65.0
typer-0.7.0 typing-extensions-4.5.0 wasabi-1.1.1
[ec2-user@ip-10-1-11-131 ~]$ python -m spacy download en_core_web_sm
-bash: python: command not found
[ec2-user@ip-10-1-11-131 ~]$ python3 -m spacy download en_core_web_sm
Defaulting to user installation because normal site-packages is not writeable
Collecting en-core-web-sm==3.5.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.5.0/en_core_web_sm-3.5.0-py3-none-any.whl (12.8 MB)
     |████████████████████████████████| 12.8 MB 5.2 MB/s
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in ./.local/lib/python3.9/site-packages (from en-core-web-sm==3.5.0) (3.5.2)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in ./.local/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.1.1)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in ./.local/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm
==3.5.0) (1.10.7)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in ./.local/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (8.1.9)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in ./.local/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.0.8)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in ./.local/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (
3.0.12)
```

```
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in ./.local/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.7)
Requirement already satisfied: jinja2 in /usr/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.11.3)
Requirement already satisfied: typing-extensions>=4.2.0 in ./.local/lib/python3.9/site-packages (from pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4->spacy<3.6
.0,>=3.5.0->en-core-web-sm==3.5.0) (4.5.0)
Requirement already satisfied: idna<3,>=2.5 in /usr/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.
5.0) (2.10)
Requirement already satisfied: chardet<5,>=3.0.2 in /usr/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-s
m==3.5.0) (4.0.0)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /usr/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-w
eb-sm==3.5.0) (1.25.10)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in ./.local/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->en-cor
e-web-sm==3.5.0) (0.0.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in ./.local/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->en-core-web-
sm==3.5.0) (0.7.9)
Requirement already satisfied: click<9.0.0,>=7.1.1 in ./.local/lib/python3.9/site-packages (from typer<0.8.0,>=0.3.0->spacy<3.6.0,>=3.5.0->en-core-web
-sm==3.5.0) (8.1.3)
Requirement already satisfied: MarkupSafe>=0.23 in /usr/lib64/python3.9/site-packages (from jinja2->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.1.1
)
Installing collected packages: en-core-web-sm
Successfully installed en-core-web-sm-3.5.0
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
[ec2-user@ip-10-1-11-131 ~]$
```

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

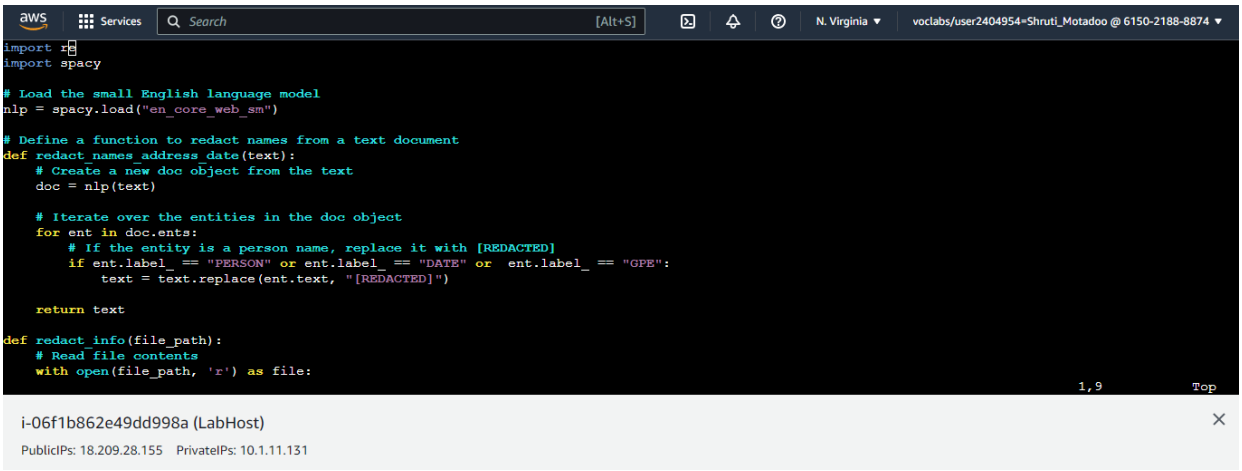
data.txt

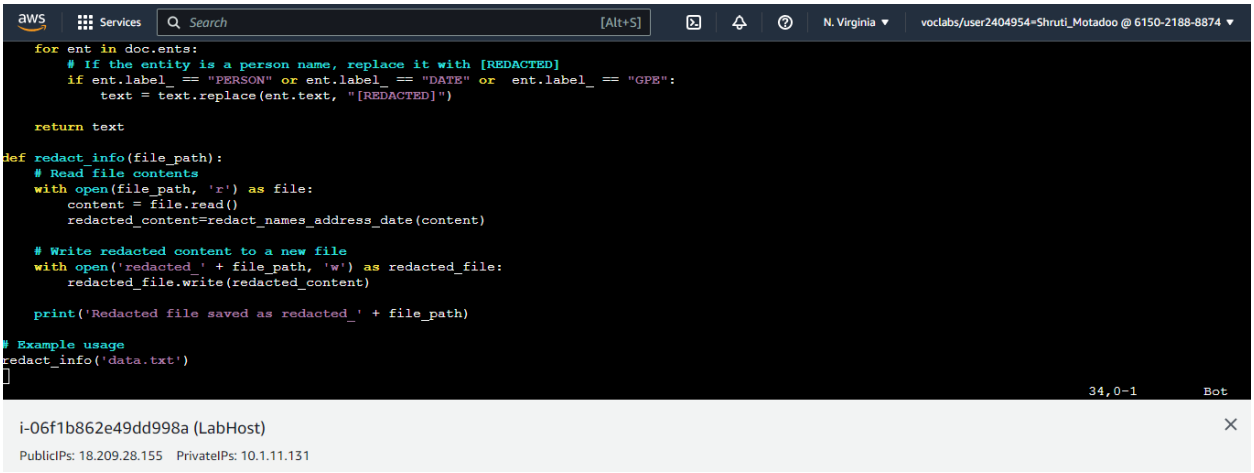
```
The University of Oklahoma (OU) is a public research university located in Norman, Oklahoma. It was founded in 1890 and is the flagship institution of
the University of Oklahoma System. OU is known for its strong academic programs, competitive athletics, and vibrant student life. The president of OU
is Joseph Harroz Jr.

Address:
660 Parrington Oval
Norman, OK 73019
United States
~
~
~
~
```

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].

seppro.py



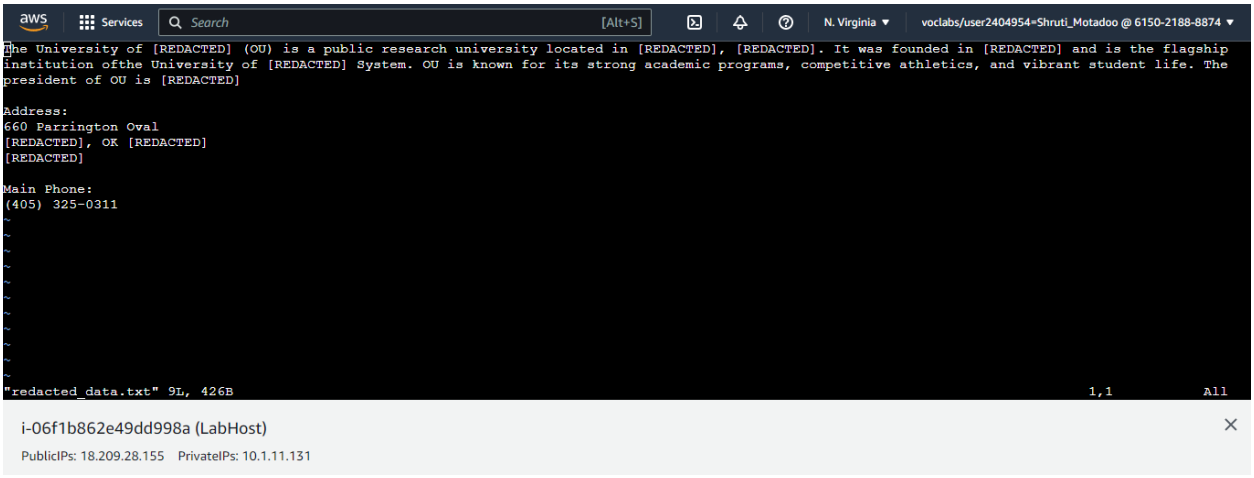


To execute we run the following command,

python3 seppro.py

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

redacted_data.txt


aws
Services
Search
[Alt+S]
N. Virginia
voclabs/user2404954=Shruti_Motadoo @ 6150-2188-8874
The University of [REDACTED] (OU) is a public research university located in [REDACTED], [REDACTED]. It was founded in [REDACTED] and is the flagship institution ofthe University of [REDACTED] System. OU is known for its strong academic programs, competitive athletics, and vibrant student life. The president of OU is [REDACTED]
Address:
660 Parrington Oval
[REDACTED], OK [REDACTED]
[REDACTED]
Main Phone:
(405) 325-0311
"redacted_data.txt" 9L, 426B
1,1
All
i-06f1b862e49dd998a (LabHost)
PublicIPs: 18.209.28.155 PrivateIPs: 10.1.11.131