Problem Statement:

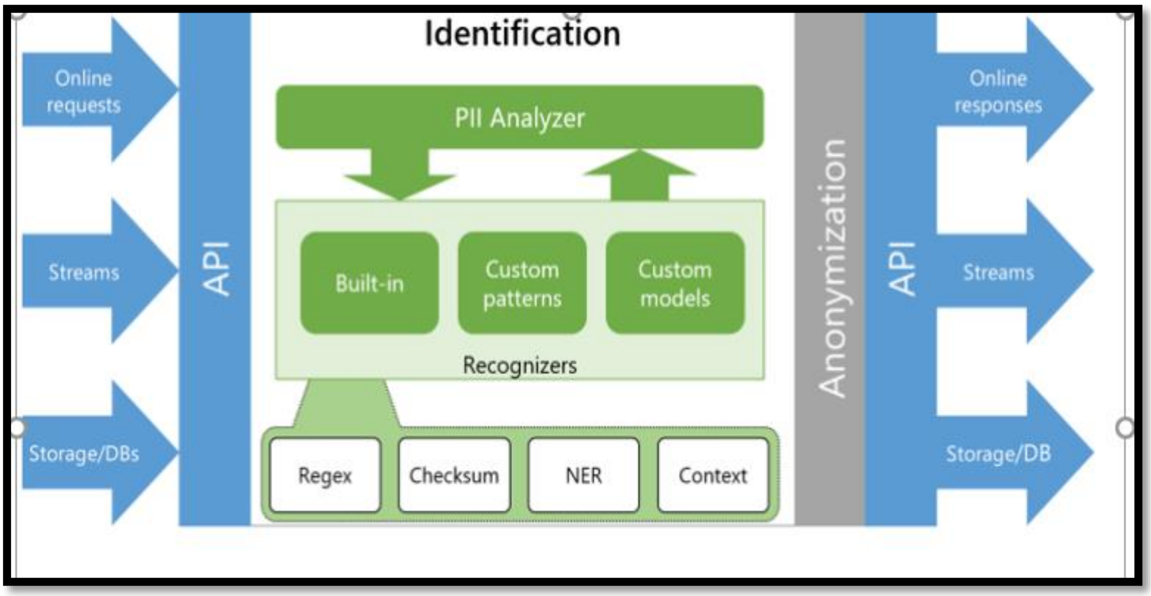
Work on anonymizing using presidio:

1. For example, if there is an email, then it should automatically detect PII information and encrypt/mark it.
2. Build a custom named entity recognizer that can identify the medicines, disease, and age. For Example: “hello I am taking disprin and my age is 25 and i am having bipolar disorder” then it should automatically tell the named identities.

1. **Email encrypts:**

PII stands for Personally Identifiable Information and for detecting this we have Presidio which allows any user to create standard and transparent processes for anonymizing PII entities on structured and unstructured data.

To do so, it exposes a set of predefined PII recognizers (for common entities like emails, credit card numbers and phone numbers), and tools for extending it with new logic for identifying more specific PII entities.



Here we see that after loading the request or files from the DBs, it is sent to PII analyzer, which has several built-in models and custom patterns, and we can also make custom models using machine learning as well.

After analyzer using recognizer through various ways like regex, checksum, it sends it output to anonymizer which tells or encrypts or marks the PII and sent it back to the API's.

Code for Email_Encryption:

Import statements for analyzers:

- pip install presidio_analyzer
- pip install presidio_anonymizer
- from typing import List
- import pprint
- from presidio_analyzer import AnalyzerEngine, PatternRecognizer, EntityRecognizer, Pattern, RecognizerResult
- from presidio_analyzer.recognizer_registry import RecognizerRegistry
- from presidio_analyzer.nlp_engine import NlpEngine, SpacyNlpEngine, NlpArtifacts

1. **Define the regex pattern in a Presidio `Pattern` object:**

- Email_pattern = Pattern(name="Email_pattern",regex="[a-zA-Z0-9+._-]+@[a-zA-Z0-9._-]+\.[a-zA-Z0-9_-]+", score = 0.5)

2. **Define the recognizer with one or more patterns**

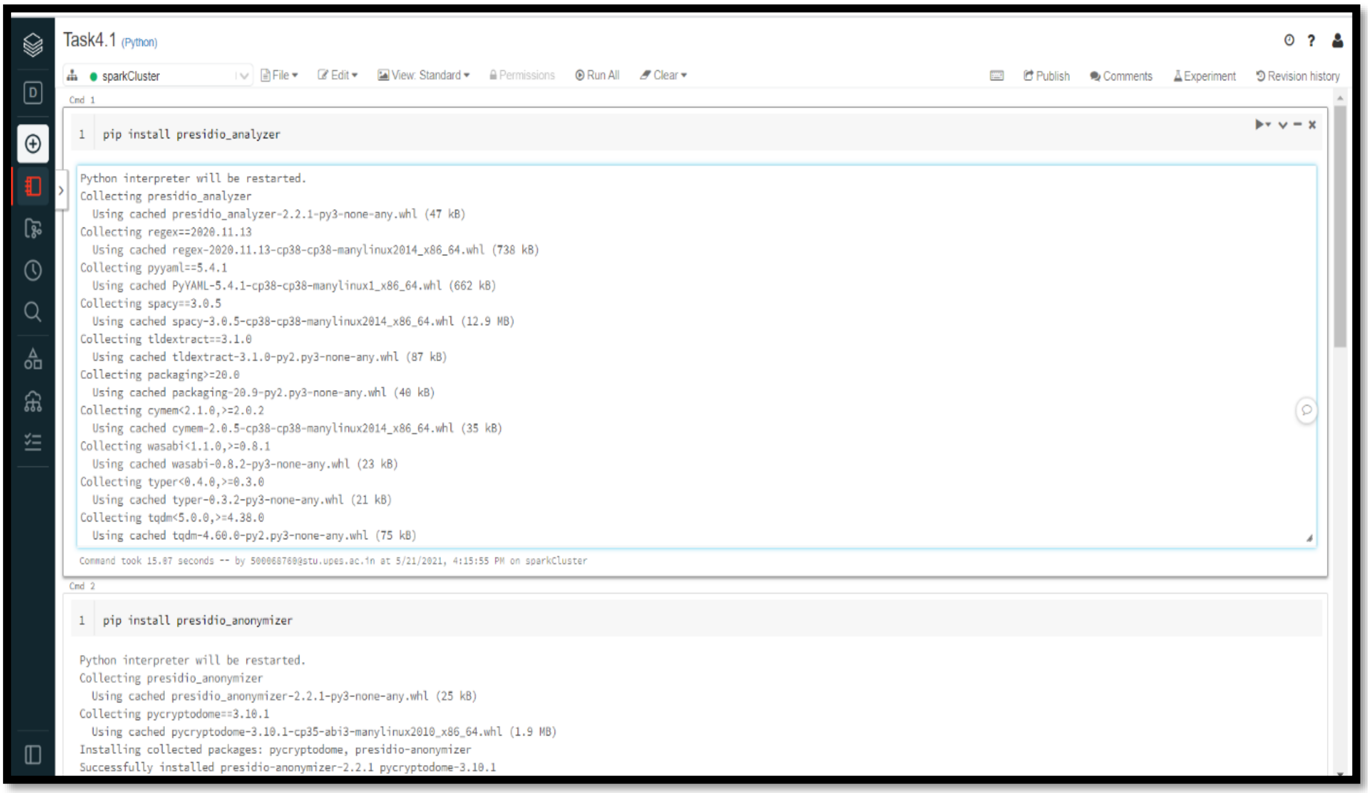
- Email_recognizer = PatternRecognizer(supported_entity="Email", patterns = [Email_pattern])

3. Testing the analyzer:

- myemail = "Feel free to mail me the issue at lakshay.sharma@rani.ai or to the our head aparnesh.gaurav@rani.ai"

- Email_result = Email_recognizer.analyze(text=myemail, entities=["Email"])
- print("Result:")
- print(Email_result.text)

Snapshots:



Cmd 1

```
pip install presidio_analyzer
```

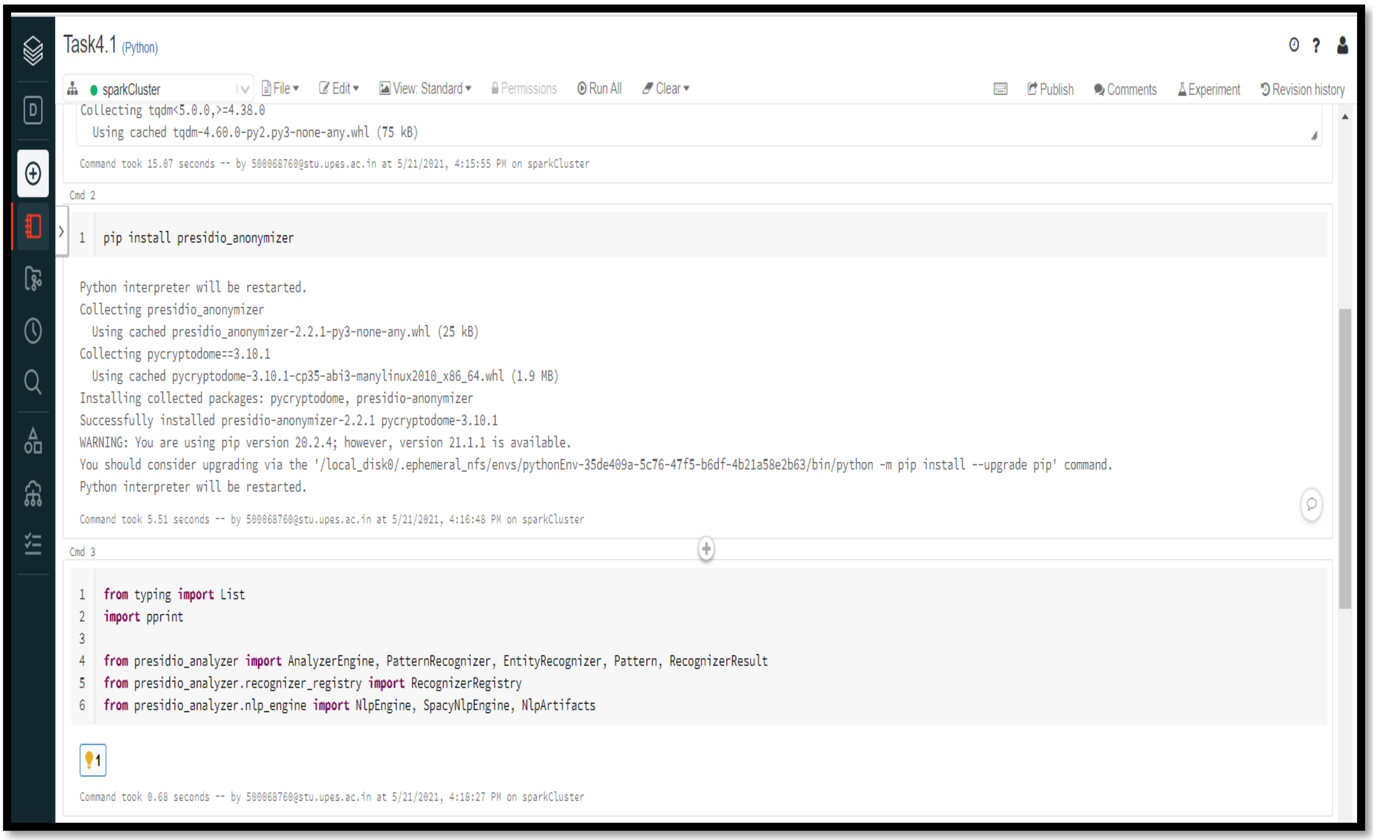
```
Python interpreter will be restarted.
Collecting presidio_analyzer
  Using cached presidio_analyzer-2.2.1-py3-none-any.whl (47 kB)
Collecting regex==2020.11.13
  Using cached regex-2020.11.13-cp38-cp38-manylinux2014_x86_64.whl (738 kB)
Collecting pyyaml==5.4.1
  Using cached PyYAML-5.4.1-cp38-cp38-manylinux1_x86_64.whl (662 kB)
Collecting spacy==3.0.5
  Using cached spacy-3.0.5-cp38-cp38-manylinux2014_x86_64.whl (12.9 MB)
Collecting tldextract==3.1.0
  Using cached tldextract-3.1.0-py2.py3-none-any.whl (87 kB)
Collecting packaging>=20.0
  Using cached packaging-20.9-py2.py3-none-any.whl (40 kB)
Collecting cymem<2.1.0,>=2.0.2
  Using cached cymem-2.0.5-cp38-cp38-manylinux2014_x86_64.whl (35 kB)
Collecting wasabi<1.1.0,>=0.8.1
  Using cached wasabi-0.8.2-py3-none-any.whl (23 kB)
Collecting typer<0.4.0,>=0.3.0
  Using cached typer-0.3.2-py3-none-any.whl (21 kB)
Collecting tqdm<5.0.0,>=4.38.0
  Using cached tqdm-4.60.0-py2.py3-none-any.whl (75 kB)
```

Command took 15.07 seconds -- by 500068760@stu.upes.ac.in at 5/21/2021, 4:15:55 PM on sparkCluster

Cmd 2

```
pip install presidio_anonymizer
```

```
Python interpreter will be restarted.
Collecting presidio_anonymizer
  Using cached presidio_anonymizer-2.2.1-py3-none-any.whl (25 kB)
Collecting pycryptodome==3.10.1
  Using cached pycryptodome-3.10.1-cp35-abi3-manylinux2010_x86_64.whl (1.9 MB)
Installing collected packages: pycryptodome, presidio-anonymizer
Successfully installed presidio-anonymizer-2.2.1 pycryptodome-3.10.1
```



```
Collecting tqdm<5.0.0,>=4.38.0
  Using cached tqdm-4.60.0-py2.py3-none-any.whl (75 kB)
```

Command took 15.07 seconds -- by 500068760@stu.upes.ac.in at 5/21/2021, 4:15:55 PM on sparkCluster

Cmd 2

```
pip install presidio_anonymizer
```

```
Python interpreter will be restarted.
Collecting presidio_anonymizer
  Using cached presidio_anonymizer-2.2.1-py3-none-any.whl (25 kB)
Collecting pycryptodome==3.10.1
  Using cached pycryptodome-3.10.1-cp35-abi3-manylinux2010_x86_64.whl (1.9 MB)
Installing collected packages: pycryptodome, presidio-anonymizer
Successfully installed presidio-anonymizer-2.2.1 pycryptodome-3.10.1
WARNING: You are using pip version 20.2.4; however, version 21.1.1 is available.
You should consider upgrading via the '/local_disk0/.ephemeral_nfs/envs/pythonEnv-35de409a-5c76-47f5-b6df-4b21a58e2b63/bin/python -m pip install --upgrade pip' command.
Python interpreter will be restarted.
```

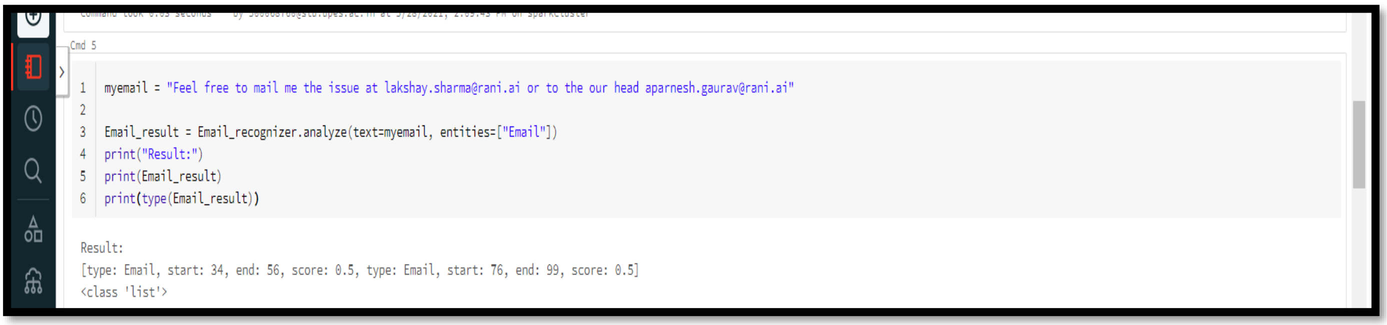
Command took 5.51 seconds -- by 500068760@stu.upes.ac.in at 5/21/2021, 4:16:48 PM on sparkCluster

Cmd 3

```python
from typing import List
import pprint

from presidio_analyzer import AnalyzerEngine, PatternRecognizer, EntityRecognizer, Pattern, RecognizerResult
from presidio_analyzer.recognizer_registry import RecognizerRegistry
from presidio_analyzer.nlp_engine import NlpEngine, SpacyNlpEngine, NlpArtifacts
```

1

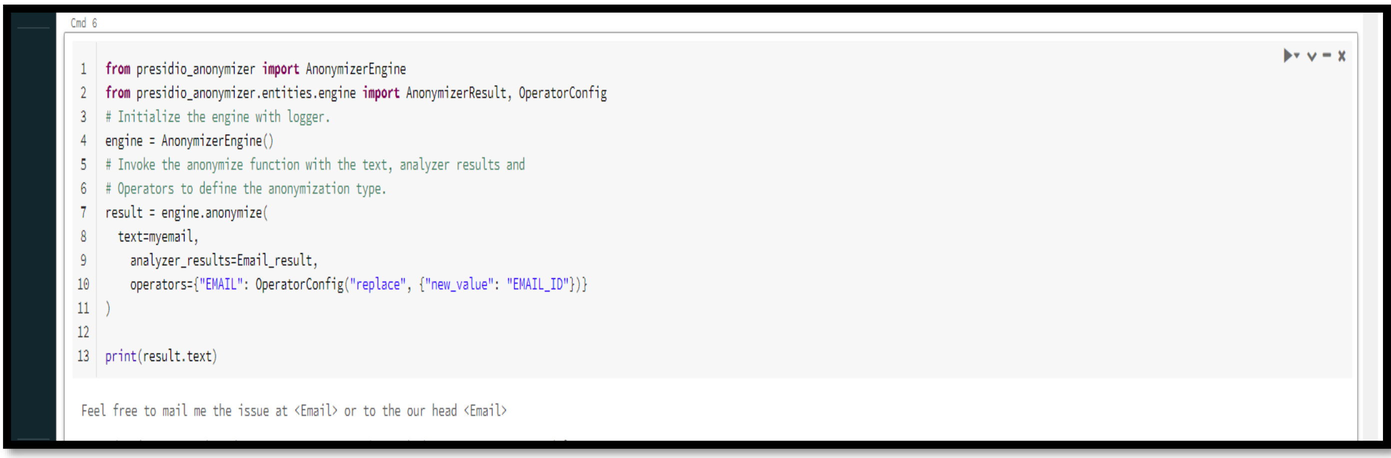
Command took 0.68 seconds -- by 500068760@stu.upes.ac.in at 5/21/2021, 4:18:27 PM on sparkCluster



Cmd 5

```python
myemail = "Feel free to mail me the issue at lakshay.sharma@rani.ai or to the our head aparnesh.gaurav@rani.ai"

Email_result = Email_recognizer.analyze(text=myemail, entities=["Email"])
print("Result:")
print(Email_result)
print(type(Email_result))
```

```
Result:
[type: Email, start: 34, end: 56, score: 0.5, type: Email, start: 76, end: 99, score: 0.5]
<class 'list'>
```



Cmd 6

```python
from presidio_anonymizer import AnonymizerEngine
from presidio_anonymizer.entities.engine import AnonymizerResult, OperatorConfig
# Initialize the engine with logger.
engine = AnonymizerEngine()
# Invoke the anonymize function with the text, analyzer results and
# Operators to define the anonymization type.
result = engine.anonymize(
  text=myemail,
    analyzer_results=Email_result,
    operators={"EMAIL": OperatorConfig("replace", {"new_value": "EMAIL_ID"})}
)

print(result.text)
```

```
Feel free to mail me the issue at <Email> or to the our head <Email>
```

2. Medicine, Disease and Age detection

Named-entity recognition (NER) is the process of automatically identifying the entities discussed in a text and classifying them into pre-defined categories such as ‘person’, ‘organization’, ‘location’ and so on. The spaCy library allows you to train NER models by both updating an existing spacy model to suit the specific context of your text documents and to train a fresh NER model from scratch.

Importing statements and installations

- pip install spacy
- import spacy
- %sh python -m spacy download en_core_web_sm

By default, NLP supports date location, organizations types named entities so we have to customize and add them to the pipelines.

For example:

```
# Perform standard imports

import spacy

nlp = spacy.load('en_core_web_sm')

doc1 = nlp("hello I am taking disprin and my age is 25 and i am having bipolar disorder")

show_ents(doc1)
```

```
25 - 40 - 42 - DATE - Absolute or relative dates or periods
```

Note: here we see age is identified as Date so we need to customize it.

Step1. ner=nlp.get_pipe("ner")

Step2. Training the data:

```
TRAIN_DATA = [

            ("hello i am taking disprin", {"entities": [(18, 25, "MEDICINE")]}),
            ("hello i am taking paracetamol", {"entities": [(18, 29, "MEDICINE")]}),
            ("and i am having fever", {"entities": [(16,21, "DISEASE")]}),
            ("and my age is 25 ", {"entities": [(14,16, "AGE")]}),
            ("and my age is 32 ", {"entities": [(14,16, "AGE")]}),
            ("hello i am taking Ibuprofen", {"entities": [(19,27, "MEDICINE")]}),
            ("and i am having Cramps", {"entities": [(16,22, "DISEASE")]}),
            ("and my age is 18 ", {"entities": [(14,16, "AGE")]}),
            ("hello i am taking Acetaminophen", {"entities": [(18,31, "MEDICINE")]}),
            ("and i am having cold", {"entities": [(16,20, "DISEASE")]}),
            ("and my age is 13 ", {"entities": [(14,16, "AGE")]}),
            ("hello i am taking naproxen ", {"entities": [(18,26, "MEDICINE")]}),
            ("and my age is 15 ", {"entities": [(14,16, "AGE")]}),
            ("and i am having headache", {"entities": [(16,24, "DISEASE")]}),
            ("and my age is 22 ", {"entities": [(14,16, "AGE")]}),
            ("hello i am taking aspirin ", {"entities": [(18,25, "MEDICINE")]}),
            ("and i am having bipolar disorder", {"entities": [(16,32, "DISEASE")]}),
            ("and my age is 26 ",{"entities": [(14,16, "AGE")]}),
            ("hello i am taking disprin", {"entities": [(18,25, "MEDICINE")]})]
```

Step3.Checking the entities

```
for _, annotations in TRAIN_DATA:
  for ent in annotations.get("entities"):
    print(ent[2])
    ner.add_label(ent[2])
```

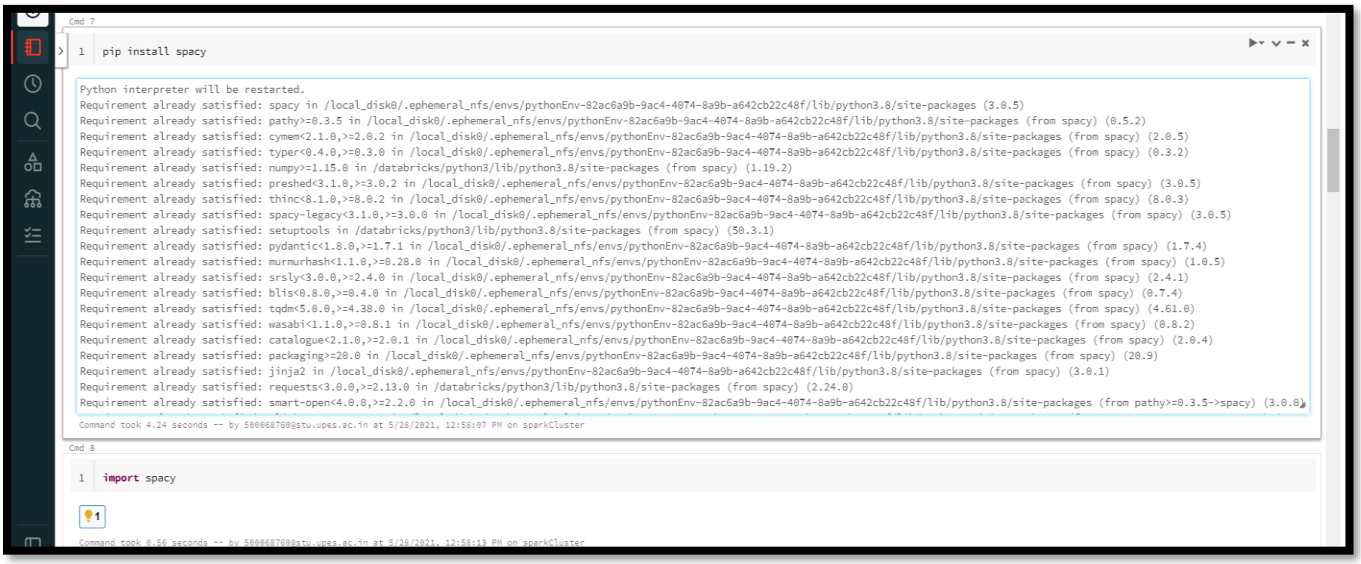
Step4. Training the model in batches and using random to have unbiased results

```
# Import requirements
import random
from spacy.training import Example
from spacy.util import minibatch, compounding
from pathlib import Path

# TRAINING THE MODEL
with nlp.disable_pipes(*unaffected_pipes):

  # Training for 30 iterations
  for iteration in range(30):

    # shuufling examples  before every iteration
    random.shuffle(TRAIN_DATA)
    losses = {}
    # batch up the examples using spaCy's minibatch
    batches = minibatch(TRAIN_DATA, size=compounding(4.0, 32.0, 1.001))
    examples = []
    for batch in batches:
        for texts,annotations in batch:
          print(type(annotations))

examples.append(Example.from_dict(nlp.make_doc(texts),annotations))
          nlp.update(examples,
                    drop=0.5,
                    losses=losses)
          print("Losses", losses)
```

Step 5: Testing the Model

```
# Testing the model
test = " hello I am taking disprin and my age is 25 and i am having bipolar disorder "
doc = nlp(test)
print("Entities", [(ent.text, ent.label_) for ent in doc.ents])
print(doc.text)
```

OUTPUTS:



Cmd 7

```
pip install spacy
```

```
Python interpreter will be restarted.
Requirement already satisfied: spacy in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (3.0.5)
Requirement already satisfied: pathy>=0.3.5 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (0.5.2)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (2.0.5)
Requirement already satisfied: typer<0.4.0,>=0.3.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (0.3.2)
Requirement already satisfied: numpy>=1.15.0 in /databricks/python3/lib/python3.8/site-packages (from spacy) (1.19.2)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (3.0.5)
Requirement already satisfied: thinc<8.1.0,>=8.0.2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (8.0.3)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (3.0.5)
Requirement already satisfied: setuptools in /databricks/python3/lib/python3.8/site-packages (from spacy) (50.3.1)
Requirement already satisfied: pydantic<1.8.0,>=1.7.1 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (1.7.4)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (1.0.5)
Requirement already satisfied: srsly<3.0.0,>=2.4.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (2.4.1)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (0.7.4)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (4.61.0)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (0.8.2)
Requirement already satisfied: catalogue<2.1.0,>=2.0.1 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (2.0.4)
Requirement already satisfied: packaging>=20.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (20.9)
Requirement already satisfied: jinja2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy) (3.0.1)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /databricks/python3/lib/python3.8/site-packages (from spacy) (2.24.0)
Requirement already satisfied: smart-open<4.0.0,>=2.2.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from pathy>=0.3.5->spacy) (3.0.0)
```

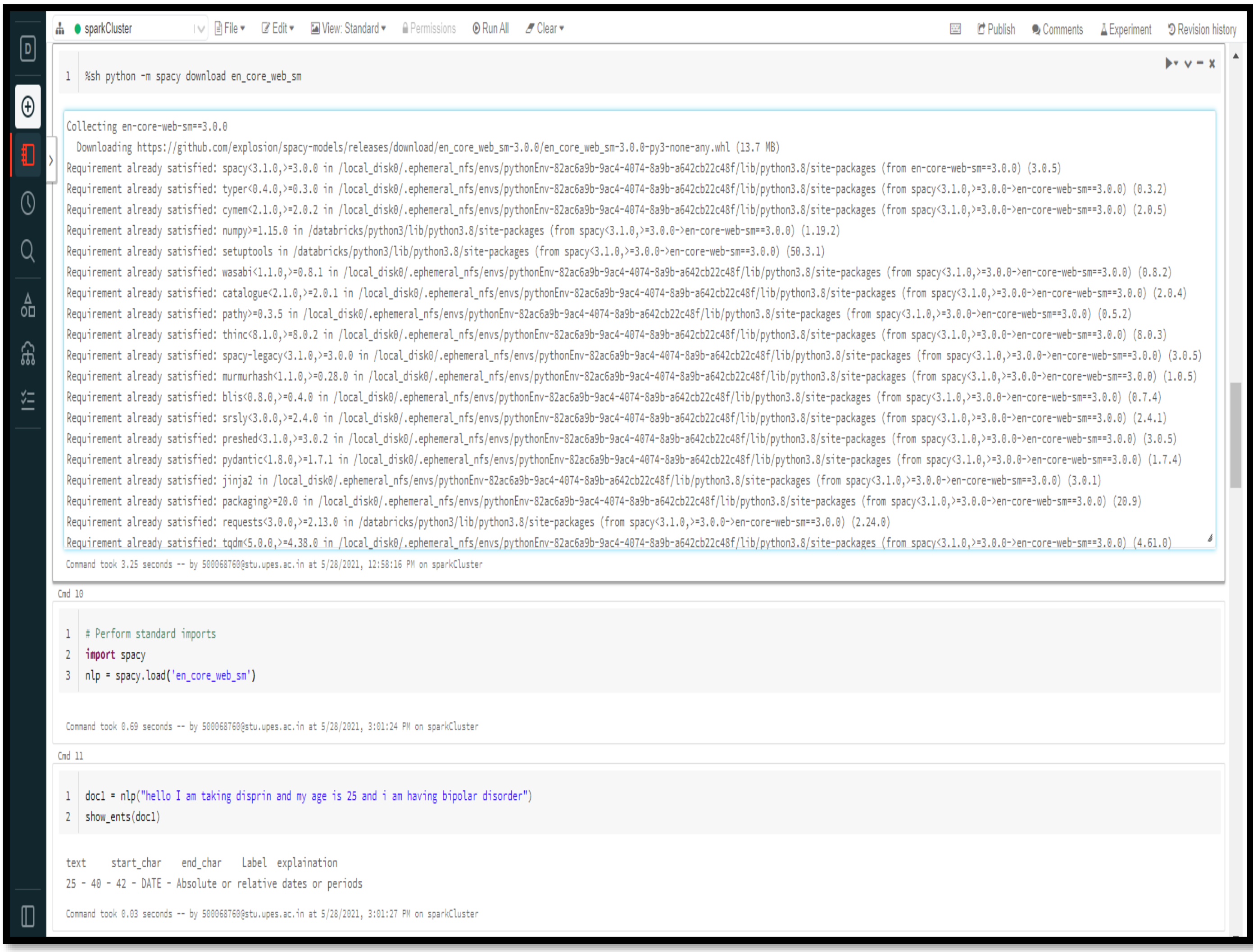
Command took 4.24 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 12:58:07 PM on sparkCluster

Cmd 8

```
import spacy
```

💡1

Command took 0.56 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 12:58:13 PM on sparkCluster



```
%sh python -m spacy download en_core_web_sm
```

```
Collecting en-core-web-sm==3.0.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.0.0/en_core_web_sm-3.0.0-py3-none-any.whl (13.7 MB)
Requirement already satisfied: spacy<3.1.0,>=3.0.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from en-core-web-sm==3.0.0) (3.0.5)
Requirement already satisfied: typer<0.4.0,>=0.3.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (0.3.2)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (2.0.5)
Requirement already satisfied: numpy>=1.15.0 in /databricks/python3/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (1.19.2)
Requirement already satisfied: setuptools in /databricks/python3/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (50.3.1)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (0.8.2)
Requirement already satisfied: catalogue<2.1.0,>=2.0.1 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (2.0.4)
Requirement already satisfied: pathy>=0.3.5 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (0.5.2)
Requirement already satisfied: thinc<8.1.0,>=8.0.2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (8.0.3)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (3.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (1.0.5)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (0.7.4)
Requirement already satisfied: srsly<3.0.0,>=2.4.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (2.4.1)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (3.0.5)
Requirement already satisfied: pydantic<1.8.0,>=1.7.1 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (1.7.4)
Requirement already satisfied: jinja2 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (3.0.1)
Requirement already satisfied: packaging>=20.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (20.9)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /databricks/python3/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (2.24.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /local_disk0/.ephemeral_nfs/envs/pythonEnv-82ac6a9b-9ac4-4074-8a9b-a642cb22c48f/lib/python3.8/site-packages (from spacy<3.1.0,>=3.0.0->en-core-web-sm==3.0.0) (4.61.0)
```

Command took 3.25 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 12:58:16 PM on sparkCluster

Cmd 10

```
# Perform standard imports
import spacy
nlp = spacy.load('en_core_web_sm')
```

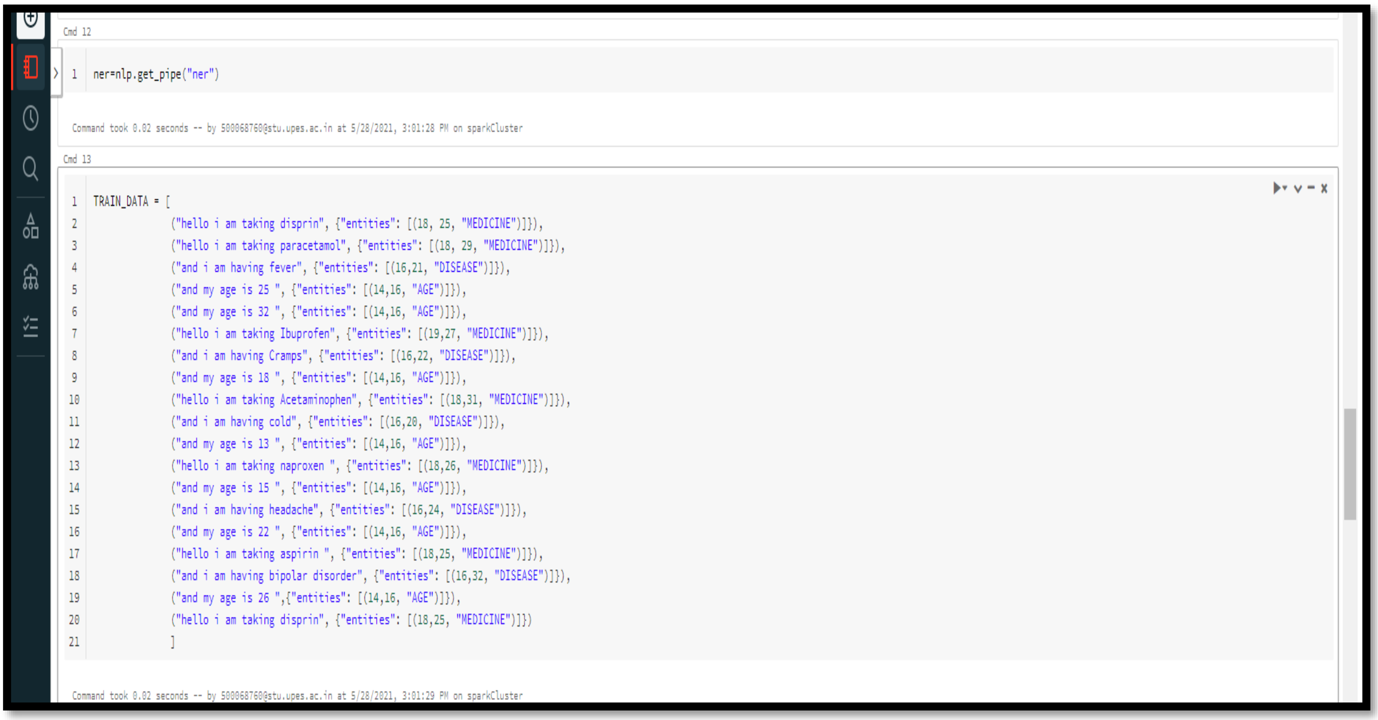
Command took 0.69 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 3:01:24 PM on sparkCluster

Cmd 11

```
doc1 = nlp("hello I am taking disprin and my age is 25 and i am having bipolar disorder")
show_ents(doc1)
```

```
text     start_char    end_char    Label  explaination
25 - 40 - 42 - DATE - Absolute or relative dates or periods
```

Command took 0.03 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 3:01:27 PM on sparkCluster



Cmd 12

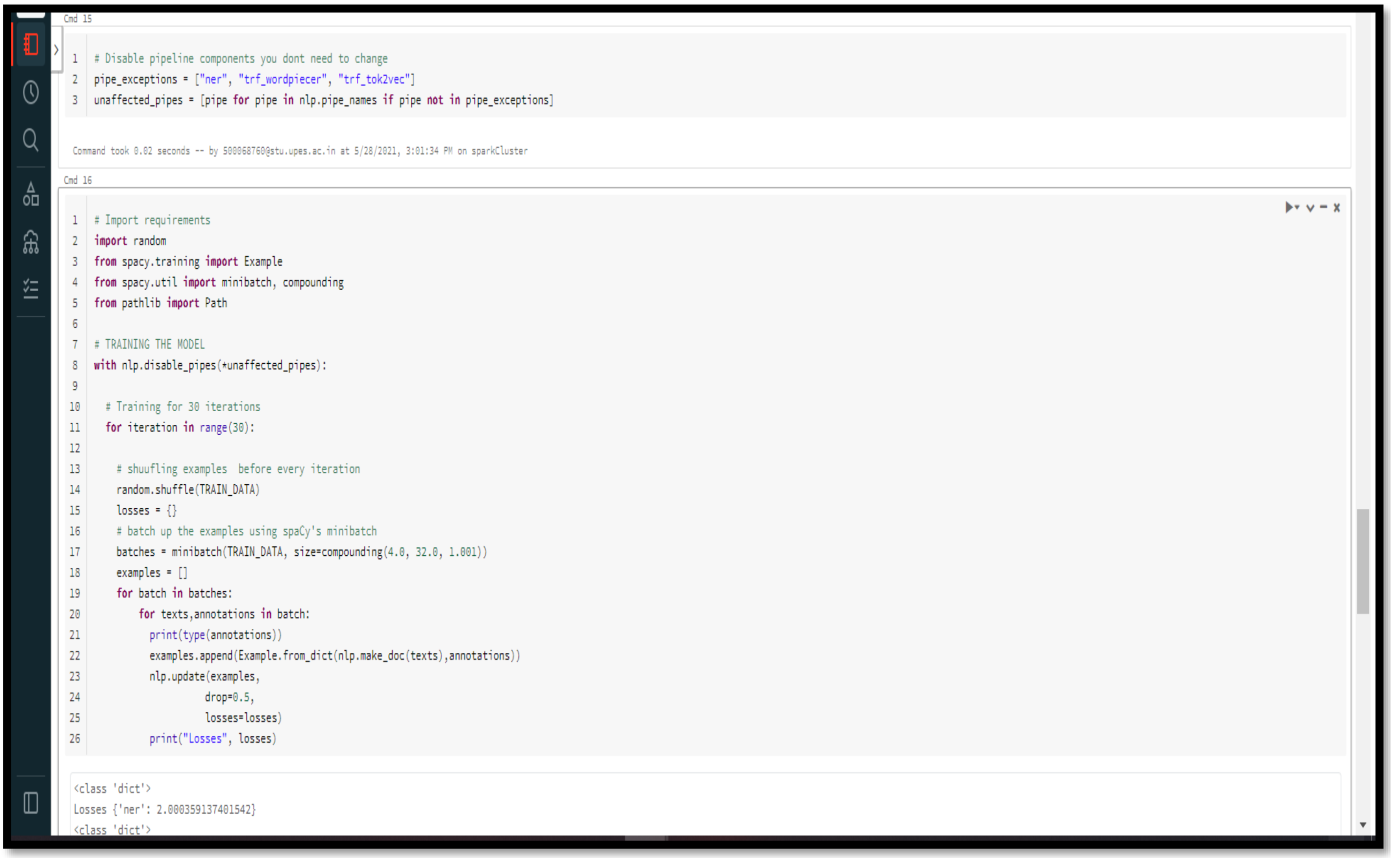
```
ner=nlp.get_pipe("ner")
```

Command took 0.02 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 3:01:28 PM on sparkCluster

Cmd 13

```
TRAIN_DATA = [
              ("hello i am taking disprin", {"entities": [(18, 25, "MEDICINE")]}),
              ("hello i am taking paracetamol", {"entities": [(18, 29, "MEDICINE")]}),
              ("and i am having fever", {"entities": [(16,21, "DISEASE")]}),
              ("and my age is 25 ", {"entities": [(14,16, "AGE")]}),
              ("and my age is 32 ", {"entities": [(14,16, "AGE")]}),
              ("hello i am taking Ibuprofen", {"entities": [(19,27, "MEDICINE")]}),
              ("and i am having Cramps", {"entities": [(16,22, "DISEASE")]}),
              ("and my age is 18 ", {"entities": [(14,16, "AGE")]}),
              ("hello i am taking Acetaminophen", {"entities": [(18,31, "MEDICINE")]}),
              ("and i am having cold", {"entities": [(16,20, "DISEASE")]}),
              ("and my age is 13 ", {"entities": [(14,16, "AGE")]}),
              ("hello i am taking naproxen ", {"entities": [(18,26, "MEDICINE")]}),
              ("and my age is 15 ", {"entities": [(14,16, "AGE")]}),
              ("and i am having headache", {"entities": [(16,24, "DISEASE")]}),
              ("and my age is 22 ", {"entities": [(14,16, "AGE")]}),
              ("hello i am taking aspirin ", {"entities": [(18,25, "MEDICINE")]}),
              ("and i am having bipolar disorder", {"entities": [(16,32, "DISEASE")]}),
              ("and my age is 26 ",{"entities": [(14,16, "AGE")]}),
              ("hello i am taking disprin", {"entities": [(18,25, "MEDICINE")]})
              ]
```

Command took 0.02 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 3:01:29 PM on sparkCluster



Cmd 15

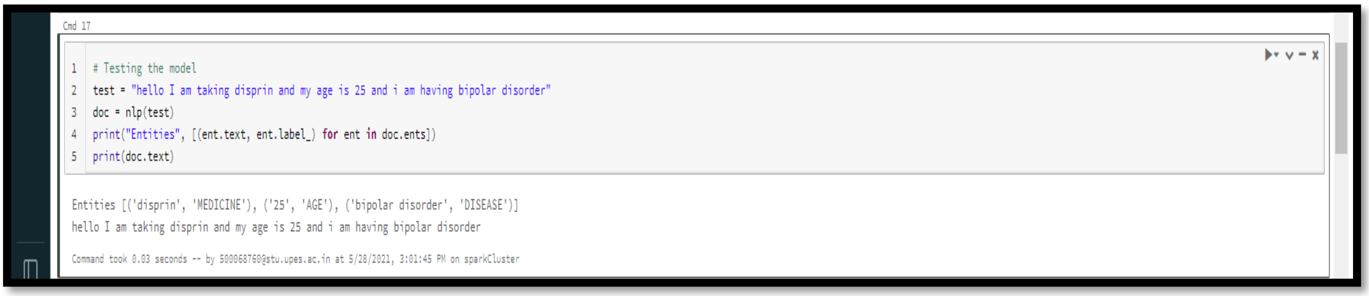
```python
# Disable pipeline components you dont need to change
pipe_exceptions = ["ner", "trf_wordpiecer", "trf_tok2vec"]
unaffected_pipes = [pipe for pipe in nlp.pipe_names if pipe not in pipe_exceptions]
```

```
Command took 0.02 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 3:01:34 PM on sparkCluster
```

Cmd 16

```python
# Import requirements
import random
from spacy.training import Example
from spacy.util import minibatch, compounding
from pathlib import Path

# TRAINING THE MODEL
with nlp.disable_pipes(*unaffected_pipes):

  # Training for 30 iterations
  for iteration in range(30):

    # shuufling examples  before every iteration
    random.shuffle(TRAIN_DATA)
    losses = {}
    # batch up the examples using spaCy's minibatch
    batches = minibatch(TRAIN_DATA, size=compounding(4.0, 32.0, 1.001))
    examples = []
    for batch in batches:
        for texts,annotations in batch:
          print(type(annotations))
          examples.append(Example.from_dict(nlp.make_doc(texts),annotations))
          nlp.update(examples,
                    drop=0.5,
                    losses=losses)
          print("Losses", losses)
```

```
<class 'dict'>
Losses {'ner': 2.000359137401542}
<class 'dict'>
```



Cmd 17

```python
# Testing the model
test = "hello I am taking disprin and my age is 25 and i am having bipolar disorder"
doc = nlp(test)
print("Entities", [(ent.text, ent.label_) for ent in doc.ents])
print(doc.text)
```

```
Entities [('disprin', 'MEDICINE'), ('25', 'AGE'), ('bipolar disorder', 'DISEASE')]
hello I am taking disprin and my age is 25 and i am having bipolar disorder
Command took 0.03 seconds -- by 500068760@stu.upes.ac.in at 5/28/2021, 3:01:45 PM on sparkCluster
```