Digital Assignment

Page No. | Date

1

Name: Eashan Kaushik
Registration no.: 17BCE2031

NLP Pipeline

Sentence Segmentation	Tokenization	Part of Speech Tagging	Lemmatization	Stop words	Dependency Parsing	Noun Phrases	Named Entity Recognition	Coreferenc Resolution

Input text

input = (variable) "New York City, often simply called New York and abbreviated as NYC, is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over 302.6 square miles, New York City is also the most densely populated major city in United States. Located at the southern tip of the U.S state of New York, the city is the centre of the New York metropolitan area, largest metropolitan area in the world by urban land mass. With almost 20 million people in its metropolitan statistical area and appro 23 million in its combined statistical area, it is one of the world's most populous megacities. New York City has been described as the cultural, financial and media capital of the world, significantly influencing commerce, entertainment, research, technology and sports. Home to the headquarters of the United Nations, New York is an important centre for international diplomacy."

Word count - 156 words

Step 1: Sentence Segmentation

The first step in the pipeline is to break the text into seperate sentences which gives us:

Code

```
import spacy

nlp= spacy.load("en_core_web-sm")
doc = nlp (input)     # input is our 156 word string
for sent in doc.sents:
        print(sent.text)
```

Output of sentence Segmentation

1. "New York city, often called simply New York and abbreviated as NYC, is the most populous city in United States."
2. "With an estimated 2019 population 8,336,817 distributed over about 302.6 square miles, New York City is also the most densely populated major city in the United States"
3. "Located at Southern tip of the U.S State of New York, the city is the center of the New York metropolitan area, largest metropolitan area in the world by landmass"
4. "With almost 20 million people in its metropolitan statistical area and approx 23 million in its combined statistical area it is one of the worlds most populous megacities"
5. "New York City has been described as the cultural, financial and media capital of the world, significantly influencing commerce, entertainment, research and sp

6. "Home to the headquarters of the United Nations, New York is an important center for international diplomacy"

Step 2: Word Tokenization

Now that we've split our document into sentences, we can process them one at a time. The next step in our pipeline is to break sentences into seperate words or tokens. This is called tokenization. The code to do this is as follows:

code

```
import spacy

nlp= spacy.load("en-core-web-sm")
doc=nlp(text)
for token in doc:
        print(token.text)
```

output of tokenization:

1. ["New", "York" "city", "often", "called", "simply", "New", "York", "and", "abbreviated", "as", "NYC", ",", "is", "the", "most", "populous", "city", "in", "United", "States", "."]

2. ["With", "an", "estimated", "2019", "population", "8,336,817", "distributed", "over", "about", "302.2", "square", "miles", "New", "York", "city", "is", "also", "the", "most", "densely", "populated", "major", "city", "in", "the", "United", "States", "."]

3. ["Located", "at", "southern", "tip", "of", "the", "U.S", "state", "of", "New", "York", ",", "the", "city", "is", "the", "center", "of", "the", "New", "York", "metropolitan", "area", ",", "largest", "metropolitan", "area", "in", "the", "world", "by", "landmass", "."]

4. ["With", "almost", "20", "million", "people", "in", "its", "metropolitan", "statistical", "area", "and", "approx" "23", "million", "in", "its",

Eashan Kaushik

17BCE2031

(4)

Page No.	
Date	

19 "area", 20 ",", 21 "it", 22 "is", 23 "one", 24 "of", 25 "the", 26 "worlds", 27 "most", 28 "populous",
29 "megacities", 30 "."]

5. [1 "New", 2 "York", 3 "City", 4 "has", 5 "been", 6 "disoribed", 7 "as", 8 "the", 9 "cultural", 10 ",", "fin
12 ",", 13 "and", 14 "media", 15 "capital", 16 "of", 17 "the", 18 "world", 19 ",", 20 "Significantly
21 "commerce", 22 ",", 23 "entetainment", 24 ",", 25 "research", 26 "and", 27 "sports", 28 ".

6. [0 "Home", 1 "to", 2 "the", 3 "headquaters", 4 "of", 5 "the", 6 "United", 7 "Nations"
8 ",", 9 "New", 10 "York", 11 "is", 12 "an", 13 "important", 14 "center", 15 "for", 16 "internati
17 "diplomacy", 18 "."]

Step 3: Predicting Parts of speech for Each Token

Next, we'll look at each token and try to guess its part of speech – whether it is a noun, a ver, an adjective and so on. Knowing the role of each word in the sentence will help us start to figure out what the sentence is talking about. We can do this by feeding each word (and some extra words around it for context) into pre trained part-of-speech classification model:

Code

```
import spacy

nlp = spacy.load("en-core-wbe_sm")
doc = nlp(text)
for token in doc:
        print(token.text, token.pos_)
```

output of POS tagging:

1. [PROPN, PROPN, PROPN, ADV, VERB, AD

Eashan Kaushik
17BCE2031

VERB, SCONJ, PROPN, PUNCT, AUX, DET, ADV, ADJ NOUN, ADP, DET, PROPN, PROPN, PUNCT]

2. [ADP, DET, VERB, NUM, NOUN, ADP, NUM, VERB, ADP, ADV, NUM, ADJ, NOUN, puut, PROPN, PROPN, PROPN, AUX, ADV, DET, ADV, ADV, ADJ, NOUN, ADP, DET, PROPN, PROPN, PUNCT]

3. [VERB, ADP, DET, ADJ, NOUN, ADP, DET, PROPN, NOUN, ADP, PROPN, PROPN, PUNCT, DET, NOUN, AUX, DET, NOUN, ADP, DET, PROPN, PROPN, ADJ, NOUN, PUNCT, DET, ADJ, ADJ, NOUN, ADP, DET, NOUN, ADP, ADJ, NOUN, PUNCT]

4. [ADP, ADV, NUM, NUM, NOUN, ADP, DET, ADJ, ADJ, NOUN, CCONJ ADV, NUM, NUM, ADP, DET, VERB, ADJ, NOUN, PUNCT, PRON, AUX, NUM, ADP, DET, NOUN, PART, ADV, ADJ, NOUN, PUNCT]

5. [PROPN, PROAN, PROPN, AUX, AUX, VERB, SCONJ, DET, APJ, PUNCT, ADJ, PUNCT, CCONJ, NOUN, NOUN, ADP, DET, NOUN, PUNCT, ADV, VERB NOUN, PUNCT, NOUN, PUNCT, NOUN, PUNCT, NOUN, PUNCT, CCONJ NOUN PUNCT]

6. [ADV, ADP, DET, NOUN, ADP, DET, PROPN, PROPN, PUNCT PROPN PROPN, AUX, DET, ADJ, NOUN, ADP, ADJ NOUN, PUNCT]

Index

PROPN	- Proper Noun
ADP	- Preposition
ADV	- Verb
DET	- Determinant
NOUN	- Noun
CCONJ	- Conjection
AUX	- Auxillary verb
PUNCT	- punctuation
SCONJ	- preposition.

Eashan Kaushik
17BCE2031

Step 4: Text Lemmatization

Lemmatization usually refers to doing things prope with the use of vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, called a lemma.

Output of Lemmatization:

1. New York City, often call, simply New York and abbreviate as NYC, be the most populous city in the United States.

2. With an estimate 2019 population of 8,336817 distribute over about 302.6 square mile, New York City be also the most densely populated major city in the United States.

3. Locate at the southern tip of the US. state of New York, the city be the center of the New York Metropolitan area, the large metropolitan area in the world by urban landma

4. With almost 20 million people in -PRON- metropolitan statistical area and approximately 23 million in -PRON- combine statistical area, -PRON- be one of the world's most populous megacitie.

5. New York City have be describe as the cultural, financial and medium capital of the world, significantly influence commerce, research, and sports.

6. Home to the headquaters of the Unite

be an important center for international diplomacy.

code

```
import spacy

nlp= spacy.load("en_core_web_sm")
doc=nlp(text)
 for token in doc:
        print(token.text, token.lemma_)
```

Step 5: Identifying Stop words

Next, we want to consider the importance of each word in the sentence. English has lots of filler words that appear very frequently like "and", "the", and "a" when doing statistics on text, these words introduce a lot of noise since they appear way more frequently than other words. Some NLP pipelines will flag them as stop words - that is, words that you might want to filter out before doing any statistical analysis.

code

```
import spacy

nlp= spacy.load("en_core_web_sm")
doc= nlp(text)

for token in doc:
        print(token.text, token.is_stop)
```

1. "often", "and", "as", "is", "the", "most", "in", "the"

2. "with", "an", "of", "over", "about", "is", "also", "the", "most", "in", "the"

3. "at", "the" "of", "the", "of", "the", "is", "the", "of", "the", "the", "in", "the", "by",

4. "with" "almost" "in", "its", "and", "in", "its", "it", "
"one", "of", "the", "'s"

5. "has", "been", "as", "the", "and", "of", "the", "and"

6. "to", "the", "of", "the", "is", "an", "for"

Step 6: Dependency Parsing

The next step is to figure out how all the words in your sentence relate to each other. This is called dependency parsing.

The goal is to build a tree that assigns a single parent word to each word in sentence. The root of the tree will be main verb in sentence.

Code

```
import spacy

nlp = spacy.load("en_core-web_sm")
doc = nlp(text)
```

Eashan Kaushik
17BCE2031

9

Page No.
Date

```
for chunk in doc.noun_chunks:
        print(chunk.text, chunk.root.text)

for token in doc:
        print(token.text, token.dep_, token.head.text, token.head
              .pos_)
display.serve(doc, style="dep")
```

Sentence 1

output:

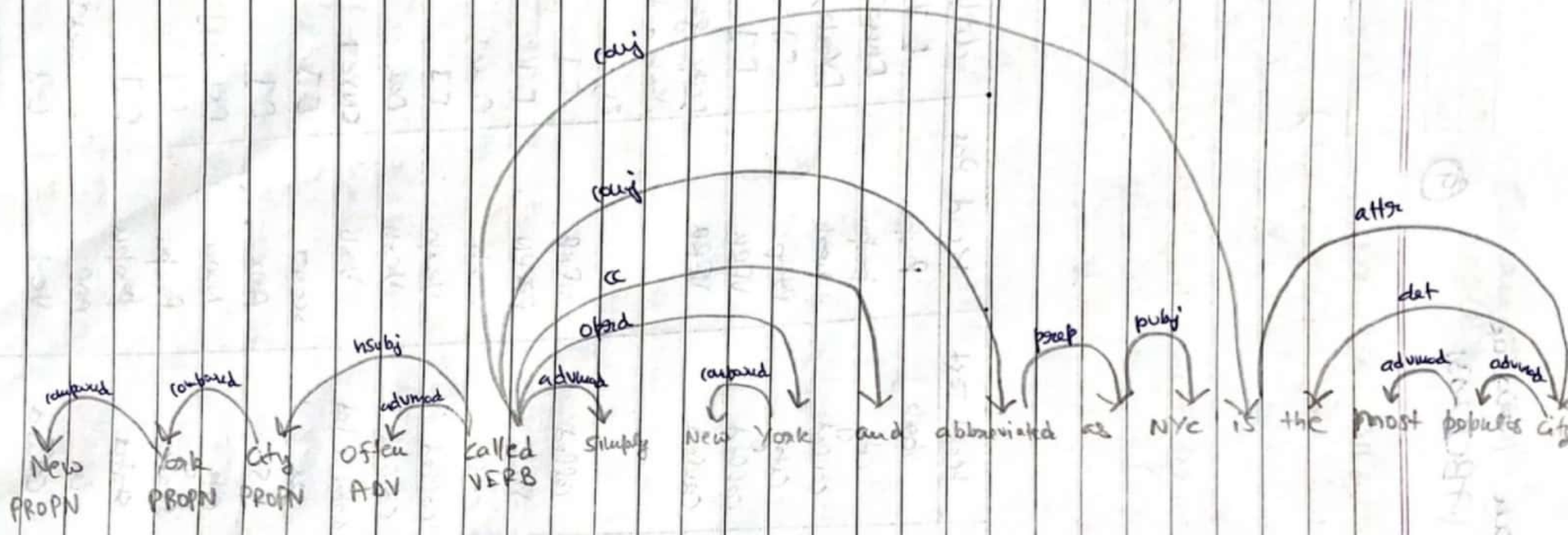
Text	Dep	Head Text	Head Pos	children
New	compound	York	Propn	[ ]
York	compound	city	propn	[new]
city	nsubj	called	verb	[York, ,]
,	punct	city	Propn	[]
often	advmod	called	VERB	[]
called	ROOT	called	VERB	[city, often, simply, York, and, abb, is, .]
simply	advmod	called	VERB	[]
New	compound	York	PROPN	[]
York	oprd	called	verb	[new]
and	cc	called	verb	[]
abbreviated	[conj]	called	verb	[as
as	prep	abbreviated	verb	[NYC]
NYC	pobj	as	SCONJ	[]
,	punct	is	AUX	[]
in	prep	city	Noun	[states]
the	det	states	Propn	[]
united	compound	states	propn	[ ]
states	pobj	in	ADP	[the, united]
.	punct	called	verb	[ ]

10

Eashan Kaushik
17BCE2031

Page No.

Date



Step 6b: Finding Noun Phrases

So far, we've treated every word in our sentence as a seperate entity. But sometimes it makes sense to group words together that represent single idea or thing. We can use the information from dependency parse tree to automatically group together words that are all talking about the same thing.

code:

```
import spacy

nlp=("en_core_web_sm")
doc=nlp(text)
print("Before:"[token.text for token in doc])

with do.retokenize() as retokenizer:
        retokenizer.merge(doc[3:5], attrs={"LEMMA"; "New York
                                                          })

print("After", [token.text for token in doc])
```

```
Before: ['New', 'York', 'City', '(', 'often', 'called', 'simply'
         'New', 'York', 'and', 'abbreviated', 'as', 'NYC', ')', 'i
         , 'the', 'most', 'populous', 'city', 'in', 'the',
         'United', 'States', '.']

After: ['New York City', '(', 'often', 'called', 'simply', 'New
        York', 'and', 'abbreviated', 'as', 'NYC', 'is', 'the'
        'most', 'populous', 'city', 'in', 'the', 'United Sates', '.']
```

Step 7: Named Entity Recognition (NER)

The goal of Named Entity Recognition, or NER, is to detect and label these nouns with real world concepts that they represent.

- Peoples Name
- Company name
- Product Name
- Date/time

code

```
import spacy
nep = spacy load('en_core_web_sm')
doc = nep(text)

for ent in doc.ents:
        print (ent.text, ent.start, ent.end_char,
                        ent.label_)
```

```
1.  New York City              GPE
    New York                   GPE
    NYC                        ORG
    The United states          GPE

2.  8,336,817                  CARDINAL
    about 362.6 Square miles   Quantity
    New York city              GPE
    United States              GPE

3.  U.S.        GPE
    New York    GPE
```

New York GPE

4. almost 20 million CARDINAL
approximately 23 million CARDINAL
one CARDINAL

5. the United Nations ORG
New York GPE

GPE - Geography
ORG - Organization

Step 8: Conference Resolution

At this point we have already taken out much useful information. However we still have a big problem. English is full of pronouns like, he, she, it. These are shortcuts we use instead of writing out the names over & over again. Humans can keep track of these word given context. Machines cant do it as it works on one sentence at a time.

This is conference resolution.

~~NLP Repo~~

PTO →

NLP Pipeline

code 1:

```
impor spacy
nlp= load('en_core_web_lg')
text= input
doc = nlp(text)
for entity in doc.ents:
        print('{0} ({1})'.format(entity.text, entity.
                                          label_))
```

Output:

```
New York City (GPE)
New York (GPE)
NYC (org)
United States (GPE)
an estimated 2019 (CARDINAL)
8,336,817 (CARDINAL)
about 302.6 square mile (Quantity)
New York City (GPE)
United States (GPE)
U.S. (GPE)
New York (GPE)
New York (GPE)
almost 20 million (Cardinal)
approx 23 million. (cardinal)
New York City (GPE)
the United Nations (ORG)
New York (GPE)
```

Extracting facts

code2:
```
import spacy
import textacy.extract

nlp = spay.load('en-core web lg')

text = input

doc = nlp(text)

statements = textacy.extract.semistructured_statements(doc,
                                                      New York)
print("Here are the things I know about london")

  for statement in statements
          sub, verb, fact
            print (f"(- {facts}")
```

output:
```
Here are the things I know about New York city
      - simply New York and abbreviated as
                        NYC
      - most populous city in United states.
      - most populous mega city
      - head quaters of united nations
      - important center of international
                  diplomacy.
```

Name: Eashan Kaushik

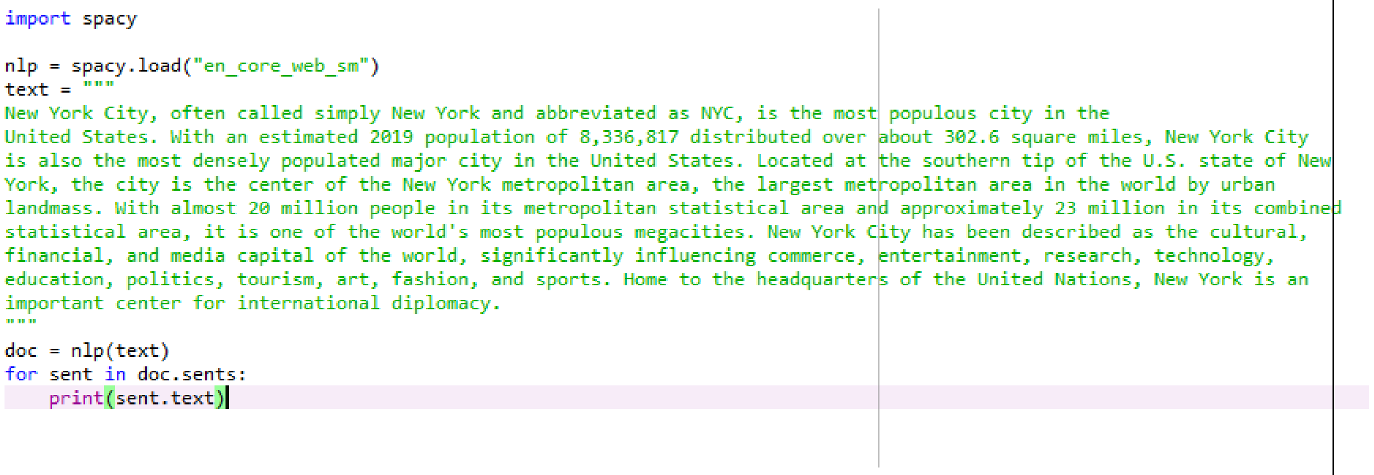
Reg. No.: 17BCE2031

Stepwise Code and output. (Step 1 to Step 8)

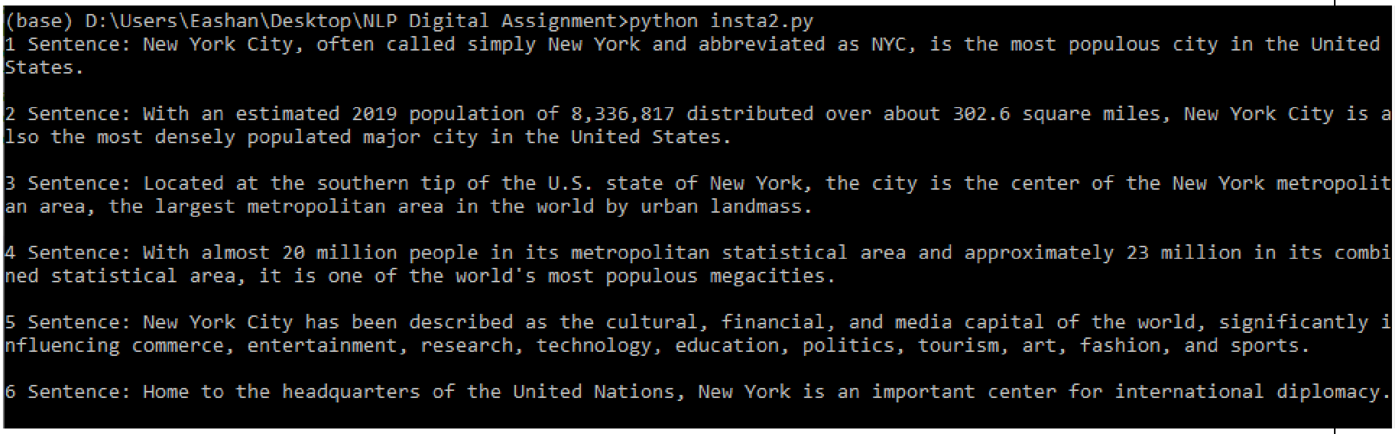
Final Pipeline Output ---- Page 8

Stepwise Code and output. (Step 1 to Step 8)

Step 1: Sentence Segmentation

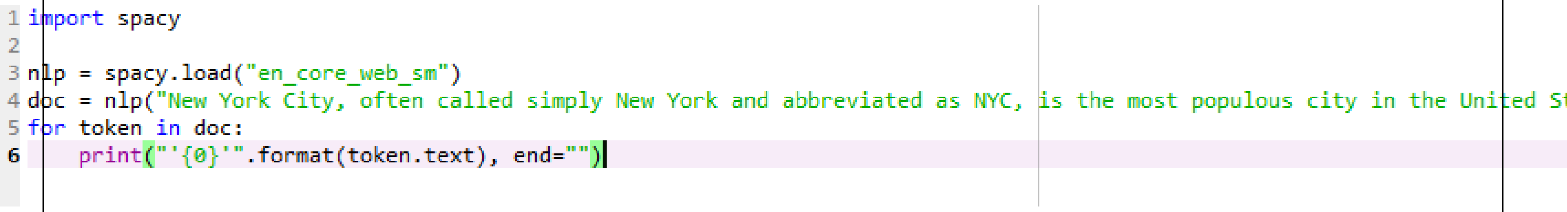


```
import spacy

nlp = spacy.load("en_core_web_sm")
text = """
New York City, often called simply New York and abbreviated as NYC, is the most populous city in the
United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles, New York City
is also the most densely populated major city in the United States. Located at the southern tip of the U.S. state of New
York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban
landmass. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined
statistical area, it is one of the world's most populous megacities. New York City has been described as the cultural,
financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology,
education, politics, tourism, art, fashion, and sports. Home to the headquarters of the United Nations, New York is an
important center for international diplomacy.
"""
doc = nlp(text)
for sent in doc.sents:
    print(sent.text)
```

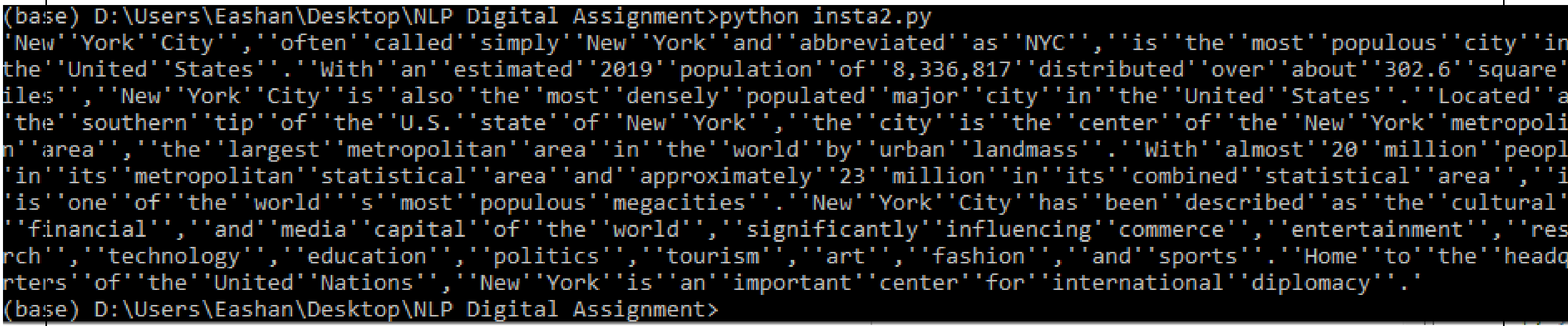


```
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>python insta2.py
1 Sentence: New York City, often called simply New York and abbreviated as NYC, is the most populous city in the United
States.

2 Sentence: With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles, New York City is a
lso the most densely populated major city in the United States.

3 Sentence: Located at the southern tip of the U.S. state of New York, the city is the center of the New York metropolit
an area, the largest metropolitan area in the world by urban landmass.

4 Sentence: With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combi
ned statistical area, it is one of the world's most populous megacities.

5 Sentence: New York City has been described as the cultural, financial, and media capital of the world, significantly i
nfluencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

6 Sentence: Home to the headquarters of the United Nations, New York is an important center for international diplomacy.
```

Step 2: Word Tokenization

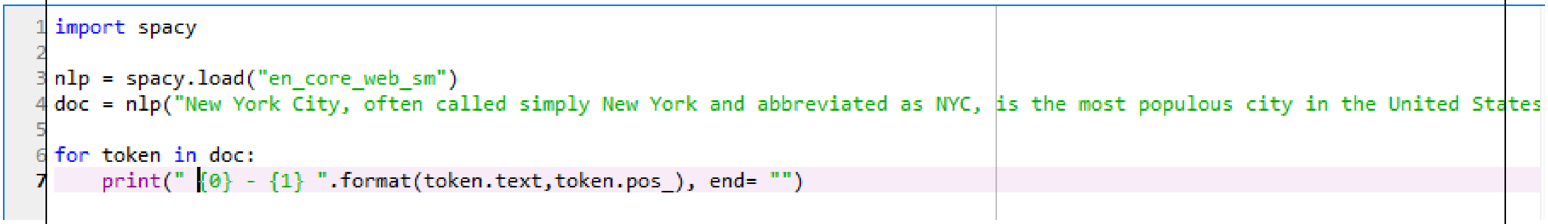


```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("New York City, often called simply New York and abbreviated as NYC, is the most populous city in the United St
for token in doc:
    print("'{0}'".format(token.text), end="")
```

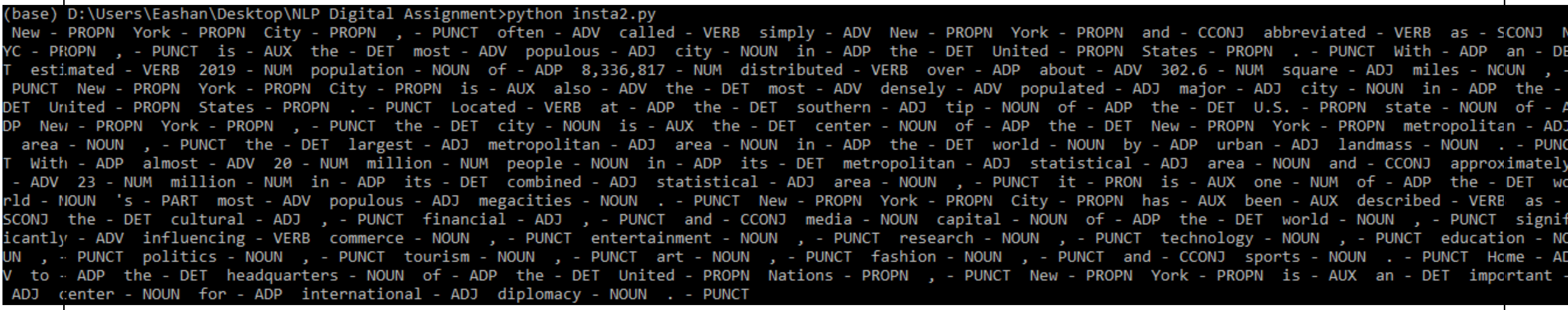


```
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>python insta2.py
'New''York''City'',''often''called''simply''New''York''and''abbreviated''as''NYC'',''is''the''most''populous''city''in
the''United''States''.''With''an''estimated''2019''population''of''8,336,817''distributed''over''about''302.6''square'
iles'',''New''York''City''is''also''the''most''densely''populated''major''city''in''the''United''States''.''Located''a
'the''southern''tip''of''the''U.S.''state''of''New''York'',''the''city''is''the''center''of''the''New''York''metropoli
n''area'',''the''largest''metropolitan''area''in''the''world''by''urban''landmass''.''With''almost''20''million''peopl
'in''its''metropolitan''statistical''area''and''approximately''23''million''in''its''combined''statistical''area'',''i
'is''one''of''the''world''s''most''populous''megacities''.''New''York''City''has''been''described''as''the''cultural'
''financial'',''and''media''capital''of''the''world'',''significantly''influencing''commerce'',''entertainment'',''res
rch'',''technology'',''education'',''politics'',''tourism'',''art'',''fashion'',''and''sports''.''Home''to''the''headq
rters''of''the''United''Nations'',''New''York''is''an''important''center''for''international''diplomacy''.'
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>
```

Step 3: Predicting Part of Speech for each Token

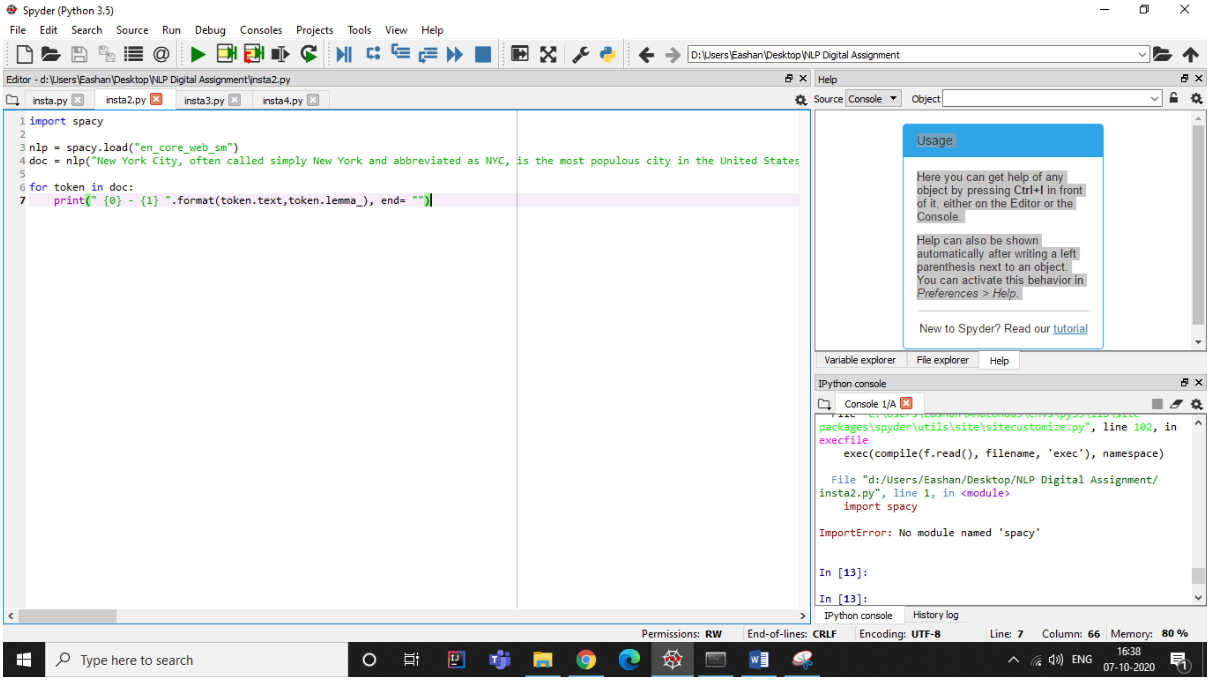


```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("New York City, often called simply New York and abbreviated as NYC, is the most populous city in the United States

for token in doc:
    print(" {0} - {1} ".format(token.text,token.pos_), end= "")
```



```
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>python insta2.py
 New - PROPN  York - PROPN  City - PROPN  , - PUNCT  often - ADV  called - VERB  simply - ADV  New - PROPN  York - PROPN  and - CCONJ  abbreviated - VERB  as - SCONJ  N
YC - PROPN  , - PUNCT  is - AUX  the - DET  most - ADV  populous - ADJ  city - NOUN  in - ADP  the - DET  United - PROPN  States - PROPN  . - PUNCT  With - ADP  an - DE
T  estimated - VERB  2019 - NUM  population - NOUN  of - ADP  8,336,817 - NUM  distributed - VERB  over - ADP  about - ADV  302.6 - NUM  square - ADJ  miles - NOUN  , -
 PUNCT  New - PROPN  York - PROPN  City - PROPN  is - AUX  also - ADV  the - DET  most - ADV  densely - ADV  populated - ADJ  major - ADJ  city - NOUN  in - ADP  the -
DET  United - PROPN  States - PROPN  . - PUNCT  Located - VERB  at - ADP  the - DET  southern - ADJ  tip - NOUN  of - ADP  the - DET  U.S. - PROPN  state - NOUN  of - A
DP  New - PROPN  York - PROPN  , - PUNCT  the - DET  city - NOUN  is - AUX  the - DET  center - NOUN  of - ADP  the - DET  New - PROPN  York - PROPN  metropolitan - ADJ
  area - NOUN  , - PUNCT  the - DET  largest - ADJ  metropolitan - ADJ  area - NOUN  in - ADP  the - DET  world - NOUN  by - ADP  urban - ADJ  landmass - NOUN  . - PUNC
T  With - ADP  almost - ADV  20 - NUM  million - NUM  people - NOUN  in - ADP  its - DET  metropolitan - ADJ  statistical - ADJ  area - NOUN  and - CCONJ  approximately
 - ADV  23 - NUM  million - NUM  in - ADP  its - DET  combined - ADJ  statistical - ADJ  area - NOUN  , - PUNCT  it - PRON  is - AUX  one - NUM  of - ADP  the - DET  wo
rld - NOUN  's - PART  most - ADV  populous - ADJ  megacities - NOUN  . - PUNCT  New - PROPN  York - PROPN  City - PROPN  has - AUX  been - AUX  described - VERB  as -
SCONJ  the - DET  cultural - ADJ  , - PUNCT  financial - ADJ  , - PUNCT  and - CCONJ  media - NOUN  capital - NOUN  of - ADP  the - DET  world - NOUN  , - PUNCT  signif
icantly - ADV  influencing - VERB  commerce - NOUN  , - PUNCT  entertainment - NOUN  , - PUNCT  research - NOUN  , - PUNCT  technology - NOUN  , - PUNCT  education - NO
UN  , - PUNCT  politics - NOUN  , - PUNCT  tourism - NOUN  , - PUNCT  art - NOUN  , - PUNCT  fashion - NOUN  , - PUNCT  and - CCONJ  sports - NOUN  . - PUNCT  Home - AD
V  to - ADP  the - DET  headquarters - NOUN  of - ADP  the - DET  United - PROPN  Nations - PROPN  , - PUNCT  New - PROPN  York - PROPN  is - AUX  an - DET  important -
 ADJ  center - NOUN  for - ADP  international - ADJ  diplomacy - NOUN  . - PUNCT
```

Step 4: Text Lemmatization


Spyder (Python 3.5)
File Edit Search Source Run Debug Consoles Projects Tools View Help
D:\Users\Eashan\Desktop\NLP Digital Assignment
Editor - d:\Users\Eashan\Desktop\NLP Digital Assignment\insta2.py
insta.py insta2.py insta3.py insta4.py
Help
Source Console Object
Variable explorer File explorer Help
IPython console
Console 1/A
IPython console History log
Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 7 Column: 66 Memory: 80 %
Type here to search
ENG 16:38 07-10-2020

```python
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("New York City, often called simply New York and abbreviated as NYC, is the most populous city in the United States

for token in doc:
    print(" {0} - {1} ".format(token.text,token.lemma_), end= "")
```

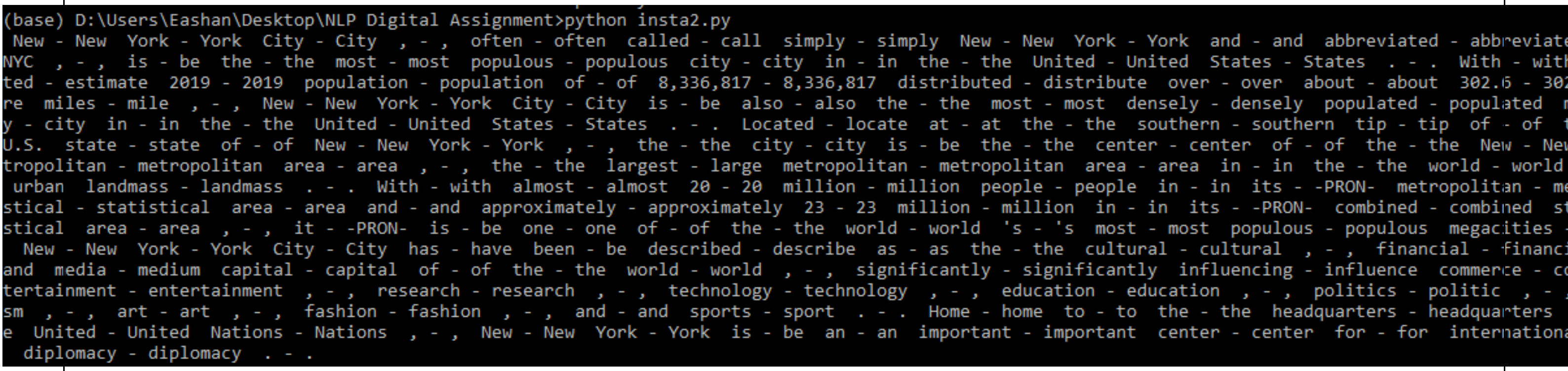
Usage

Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.

Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in *Preferences > Help*.

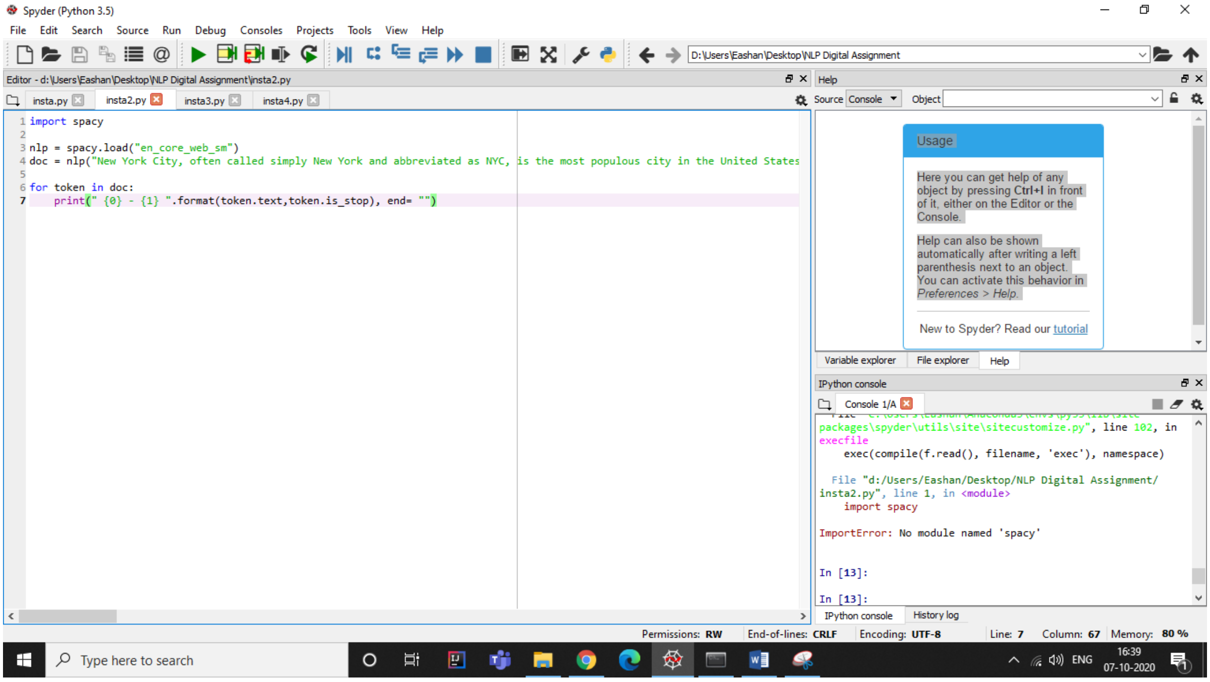
New to Spyder? Read our tutorial

```
packages\spyder\utils\site\sitecustomize.py", line 102, in
execfile
    exec(compile(f.read(), filename, 'exec'), namespace)

  File "d:/Users/Eashan/Desktop/NLP Digital Assignment/
insta2.py", line 1, in <module>
    import spacy

ImportError: No module named 'spacy'


In [13]:

In [13]:
```



```
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>python insta2.py
 New - New  York - York  City - City  , - ,  often - often  called - call  simply - simply  New - New  York - York  and - and  abbreviated - abbreviate
NYC  , - ,  is - be  the - the  most - most  populous - populous  city - city  in - in  the - the  United - United  States - States  . - .  With - with
ted - estimate  2019 - 2019  population - population  of - of  8,336,817 - 8,336,817  distributed - distribute  over - over  about - about  302.6 - 302
re  miles - mile  , - ,  New - New  York - York  City - City  is - be  also - also  the - the  most - most  densely - densely  populated - populated
y - city  in - in  the - the  United - United  States - States  . - .  Located - locate  at - at  the - the  southern - southern  tip - tip  of - of
U.S.  state - state  of - of  New - New  York - York  , - ,  the - the  city - city  is - be  the - the  center - center  of - of  the - the  New - New
tropolitan - metropolitan  area - area  , - ,  the - the  largest - large  metropolitan - metropolitan  area - area  in - in  the - the  world - world
 urban  landmass - landmass  . - .  With - with  almost - almost  20 - 20  million - million  people - people  in - in  its - -PRON-  metropolitan - me
stical - statistical  area - area  and - and  approximately - approximately  23 - 23  million - million  in - in  its - -PRON-  combined - combined  st
stical  area - area  , - ,  it - -PRON-  is - be  one - one  of - of  the - the  world - world  's - 's  most - most  populous - populous  megacities -
  New - New  York - York  City - City  has - have  been - be  described - describe  as - as  the - the  cultural - cultural  , - ,  financial - financ
and  media - medium  capital - capital  of - of  the - the  world - world  , - ,  significantly - significantly  influencing - influence  commerce - co
tertainment - entertainment  , - ,  research - research  , - ,  technology - technology  , - ,  education - education  , - ,  politics - politic  , -
sm  , - ,  art - art  , - ,  fashion - fashion  , - ,  and - and  sports - sport  . - .  Home - home  to - to  the - the  headquarters - headquarters
e  United - United  Nations - Nations  , - ,  New - New  York - York  is - be  an - an  important - important  center - center  for - for  internation
  diplomacy - diplomacy  . - .
```

Step 5: Identifying Stop Words


Spyder (Python 3.5)
File Edit Search Source Run Debug Consoles Projects Tools View Help
D:\Users\Eashan\Desktop\NLP Digital Assignment
Editor - d:\Users\Eashan\Desktop\NLP Digital Assignment\insta2.py
insta.py
insta2.py
insta3.py
insta4.py
Help
Source Console Object
Variable explorer
File explorer
Help
IPython console
Console 1/A
IPython console
History log

```python
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("New York City, often called simply New York and abbreviated as NYC, is the most populous city in the United States

for token in doc:
    print(" {0} - {1} ".format(token.text,token.is_stop), end= "")
```

Usage

Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.

Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in Preferences > Help.

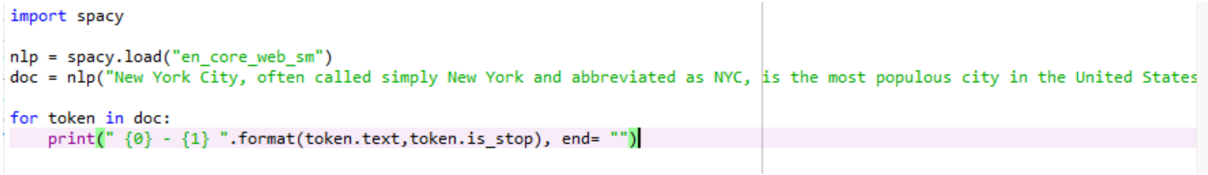
New to Spyder? Read our tutorial

```
packages\spyder\utils\site\sitecustomize.py", line 102, in
execfile
    exec(compile(f.read(), filename, 'exec'), namespace)

  File "d:/Users/Eashan/Desktop/NLP Digital Assignment/
insta2.py", line 1, in <module>
    import spacy

ImportError: No module named 'spacy'


In [13]:

In [13]:
```

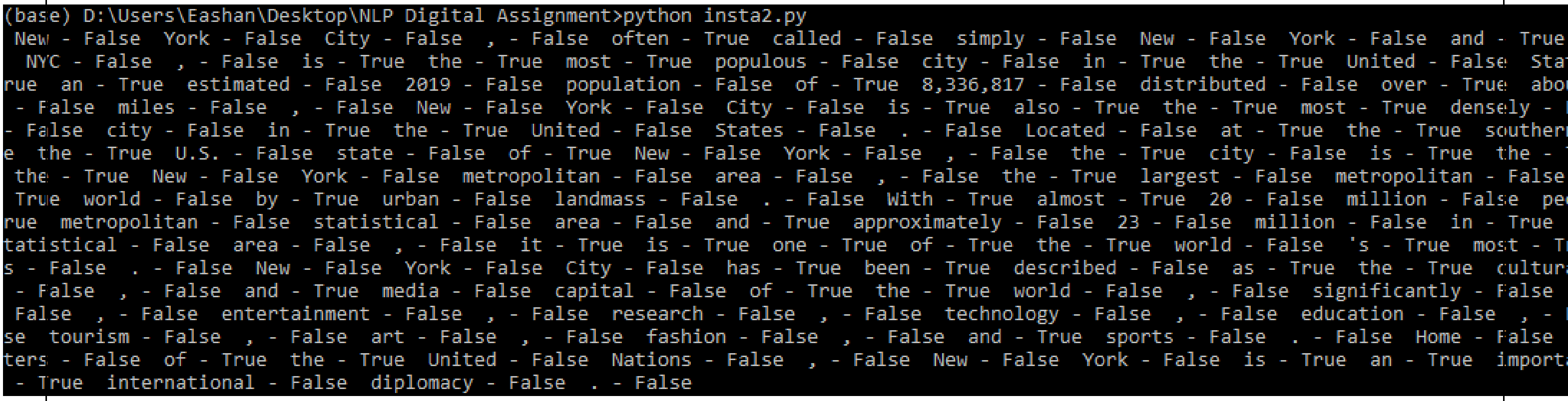
Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 7 Column: 67 Memory: 80 %

Type here to search

16:39
07-10-2020

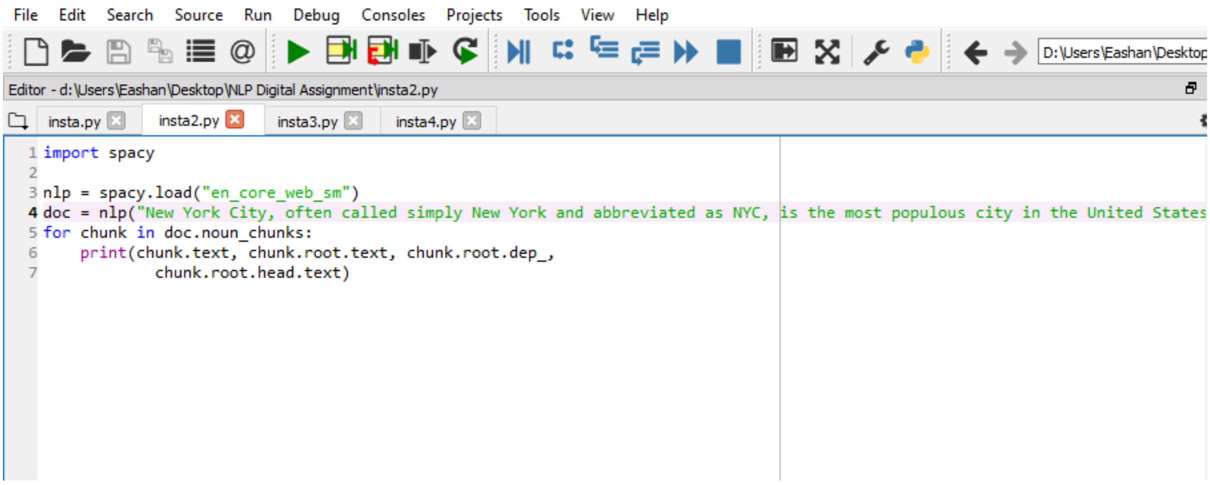


```python
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("New York City, often called simply New York and abbreviated as NYC, is the most populous city in the United States

for token in doc:
    print(" {0} - {1} ".format(token.text,token.is_stop), end= "")
```

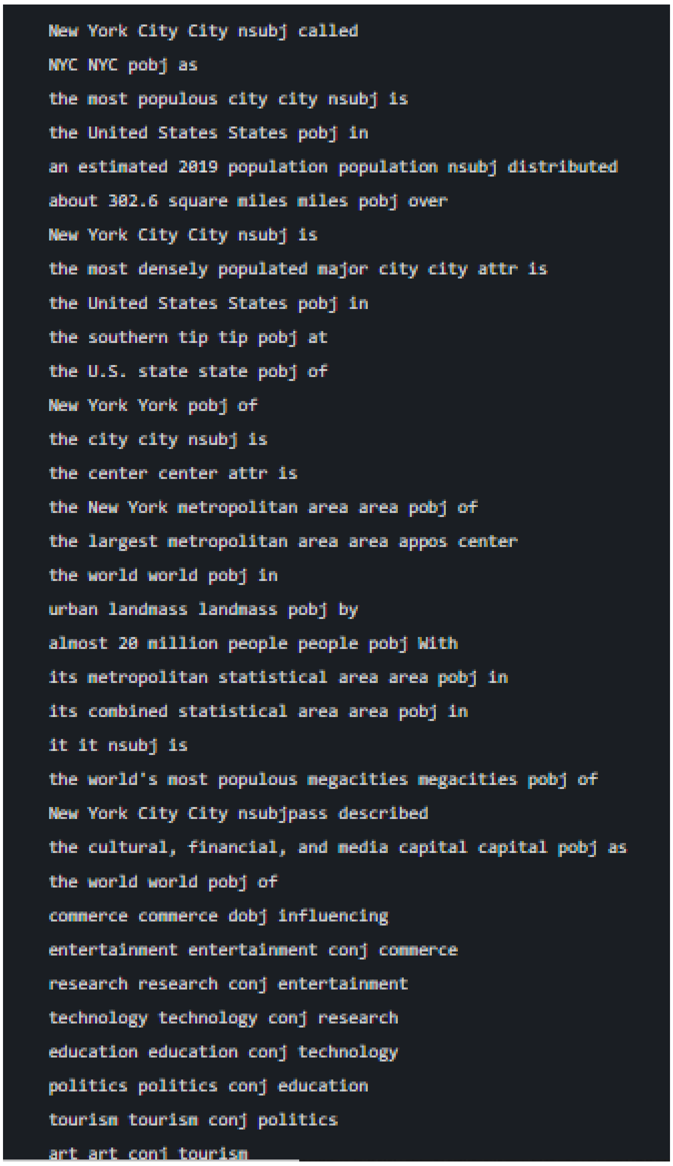


```
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>python insta2.py
 New - False  York - False  City - False  , - False  often - True  called - False  simply - False  New - False  York - False  and - True
  NYC - False  , - False  is - True  the - True  most - True  populous - False  city - False  in - True  the - True  United - False  Sta
rue  an - True  estimated - False  2019 - False  population - False  of - True  8,336,817 - False  distributed - False  over - True  abo
 - False  miles - False  , - False  New - False  York - False  City - False  is - True  also - True  the - True  most - True  densely -
- False  city - False  in - True  the - True  United - False  States - False  . - False  Located - False  at - True  the - True  souther
e  the - True  U.S. - False  state - False  of - True  New - False  York - False  , - False  the - True  city - False  is - True  the -
 the - True  New - False  York - False  metropolitan - False  area - False  , - False  the - True  largest - False  metropolitan - False
 True  world - False  by - True  urban - False  landmass - False  . - False  With - True  almost - True  20 - False  million - False  pe
rue  metropolitan - False  statistical - False  area - False  and - True  approximately - False  23 - False  million - False  in - True
tatistical - False  area - False  , - False  it - True  is - True  one - True  of - True  the - True  world - False  's - True  most - T
s - False  . - False  New - False  York - False  City - False  has - True  been - True  described - False  as - True  the - True  cultur
 - False  , - False  and - True  media - False  capital - False  of - True  the - True  world - False  , - False  significantly - False
 False  , - False  entertainment - False  , - False  research - False  , - False  technology - False  , - False  education - False  , -
se  tourism - False  , - False  art - False  , - False  fashion - False  , - False  and - True  sports - False  . - False  Home - False
ters - False  of - True  the - True  United - False  Nations - False  , - False  New - False  York - False  is - True  an - True  import
 - True  international - False  diplomacy - False  . - False
```

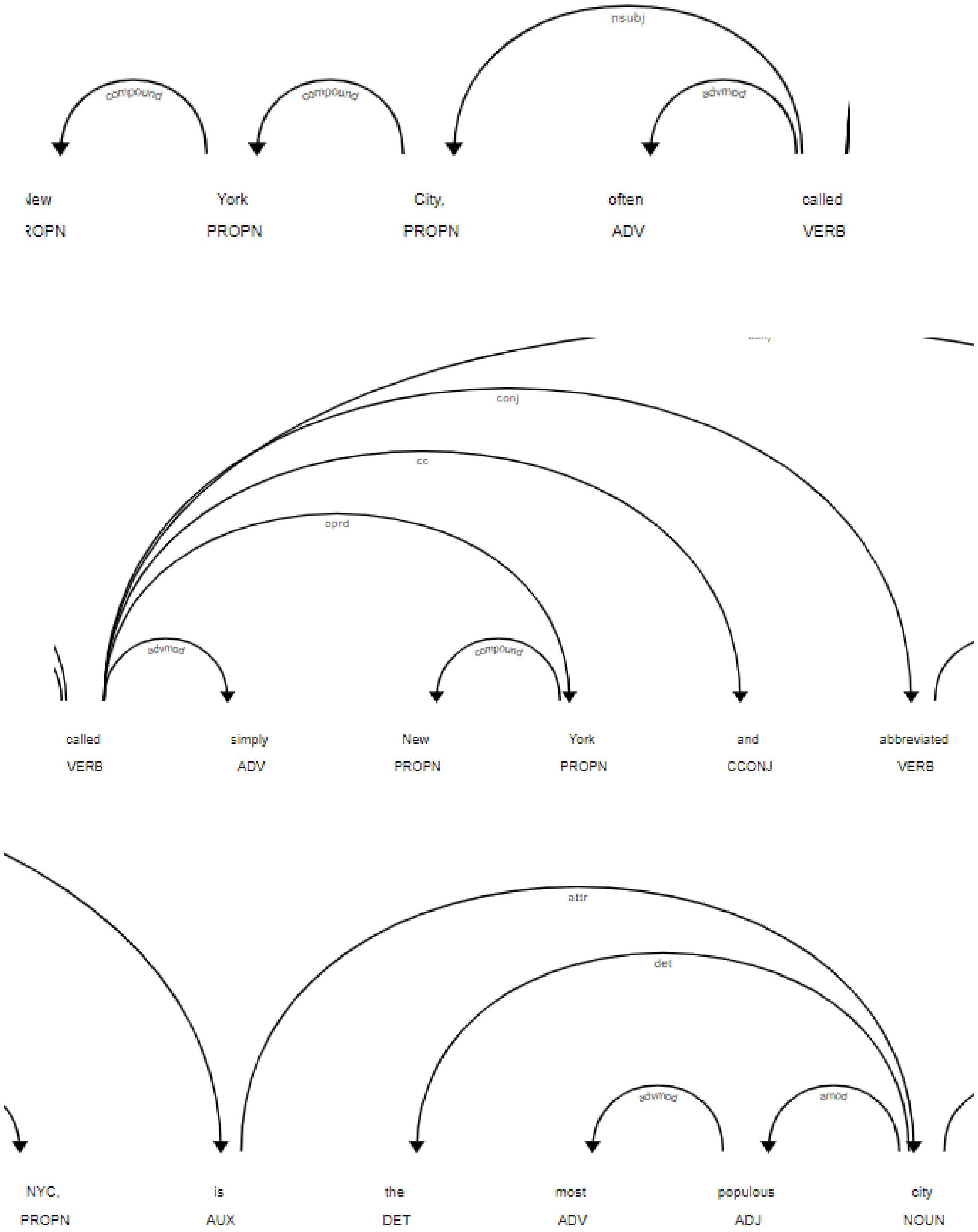
Step 6: Dependency Parsing


File Edit Search Source Run Debug Consoles Projects Tools View Help
D:\Users\Eashan\Desktop
Editor - d:\Users\Eashan\Desktop\NLP Digital Assignment\insta2.py
insta.py
insta2.py
insta3.py
insta4.py

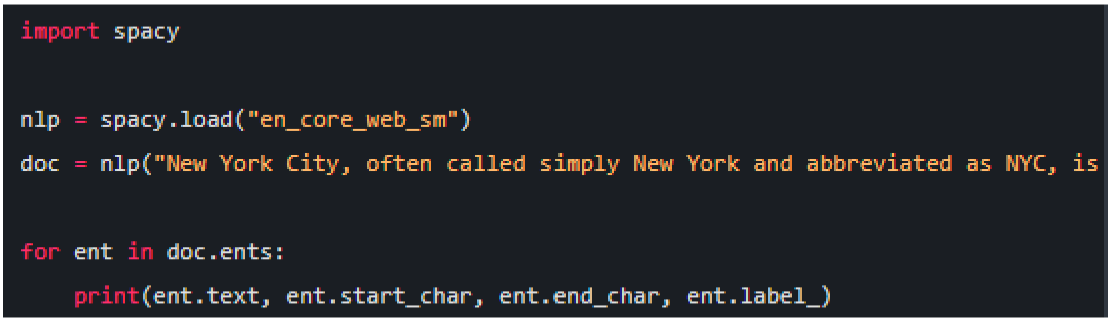
```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("New York City, often called simply New York and abbreviated as NYC, is the most populous city in the United States
for chunk in doc.noun_chunks:
    print(chunk.text, chunk.root.text, chunk.root.dep_,
            chunk.root.head.text)
```



```
New York City City nsubj called
NYC NYC pobj as
the most populous city city nsubj is
the United States States pobj in
an estimated 2019 population population nsubj distributed
about 302.6 square miles miles pobj over
New York City City nsubj is
the most densely populated major city city attr is
the United States States pobj in
the southern tip tip pobj at
the U.S. state state pobj of
New York York pobj of
the city city nsubj is
the center center attr is
the New York metropolitan area area pobj of
the largest metropolitan area area appos center
the world world pobj in
urban landmass landmass pobj by
almost 20 million people people pobj With
its metropolitan statistical area area pobj in
its combined statistical area area pobj in
it it nsubj is
the world's most populous megacities megacities pobj of
New York City City nsubjpass described
the cultural, financial, and media capital capital pobj as
the world world pobj of
commerce commerce dobj influencing
entertainment entertainment conj commerce
research research conj entertainment
technology technology conj research
education education conj technology
politics politics conj education
tourism tourism conj politics
art art conj tourism
```

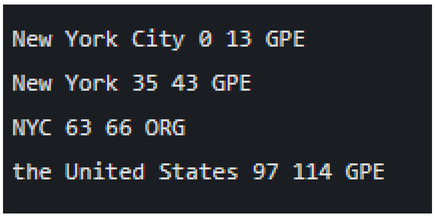

compound
compound
nsubj
advmod
New
PROPN
York
PROPN
City,
PROPN
often
ADV
called
VERB
conj
cc
oprd
advmod
compound
called
VERB
simply
ADV
New
PROPN
York
PROPN
and
CCONJ
abbreviated
VERB
attr
det
advmod
amod
NYC,
PROPN
is
AUX
the
DET
most
ADV
populous
ADJ
city
NOUN

Step 7: Named Entity Recognition



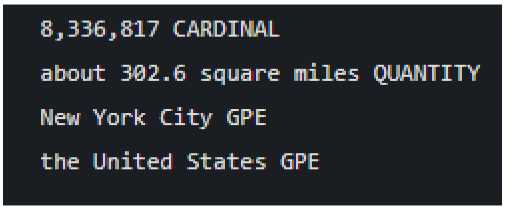
```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("New York City, often called simply New York and abbreviated as NYC, is

for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

1.



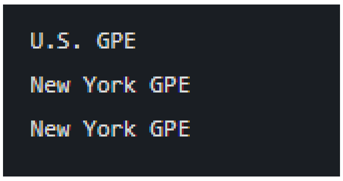
```
New York City 0 13 GPE
New York 35 43 GPE
NYC 63 66 ORG
the United States 97 114 GPE
```

2.



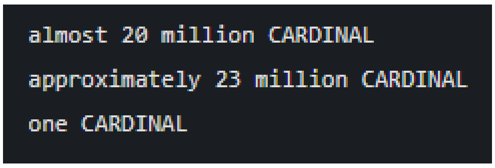
```
8,336,817 CARDINAL
about 302.6 square miles QUANTITY
New York City GPE
the United States GPE
```

3.



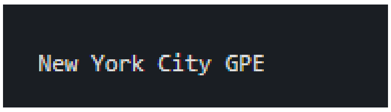
```
U.S. GPE
New York GPE
New York GPE
```

4.



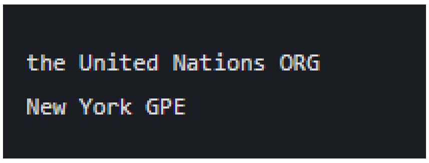
```
almost 20 million CARDINAL
approximately 23 million CARDINAL
one CARDINAL
```

5.



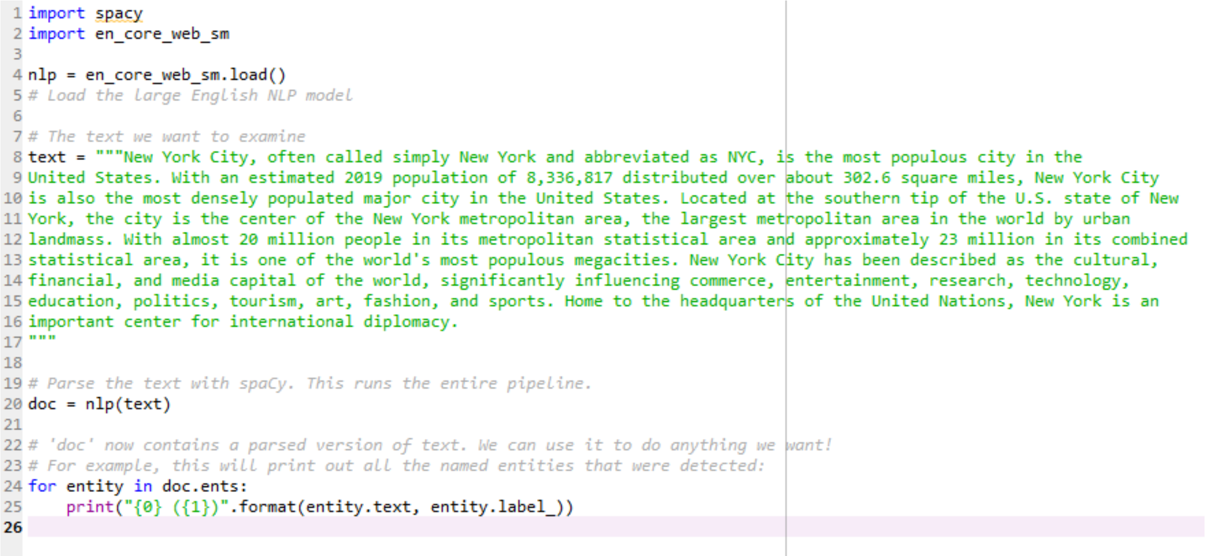
```
New York City GPE
```

6.

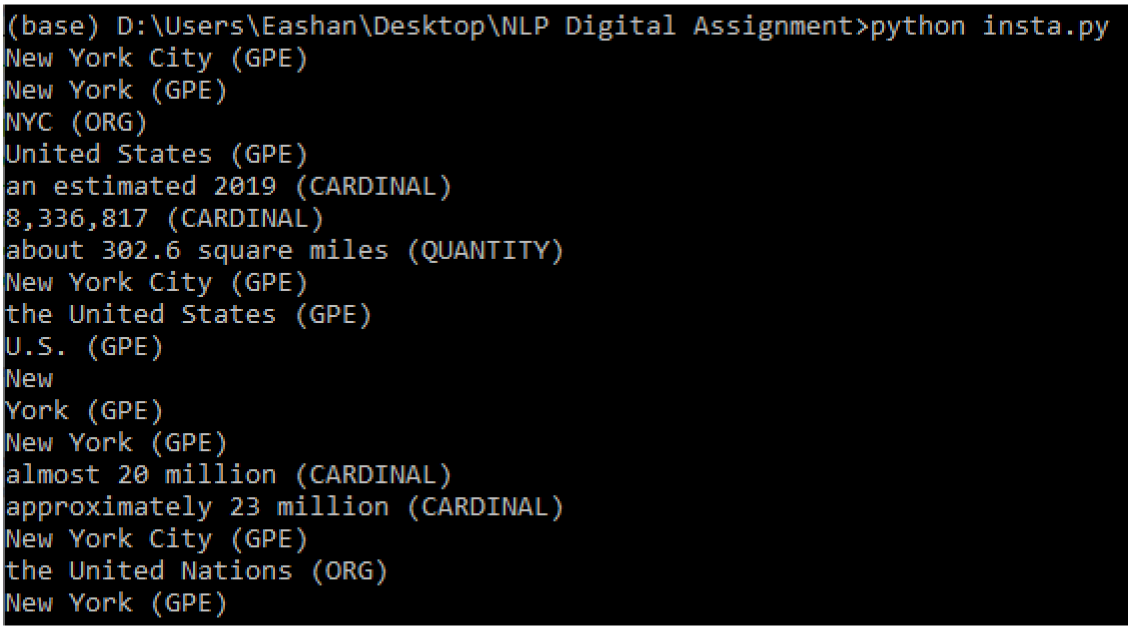


```
the United Nations ORG
New York GPE
```

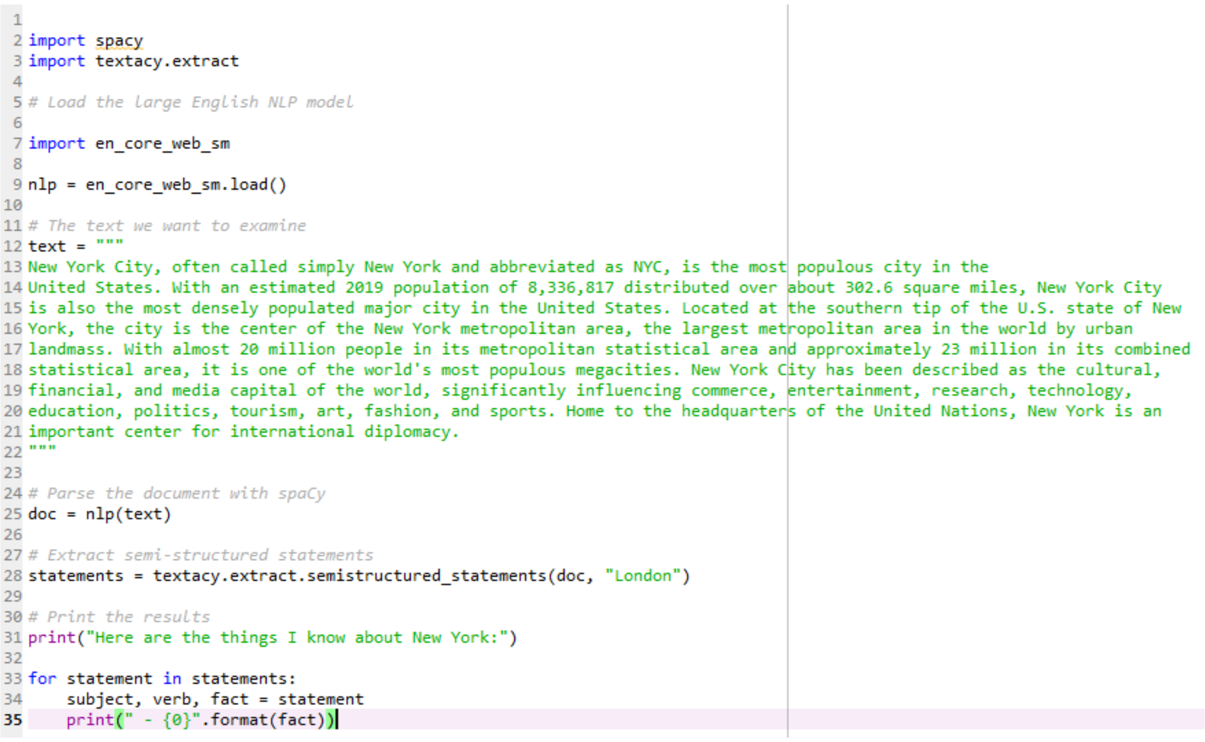
Final Pipeline Output



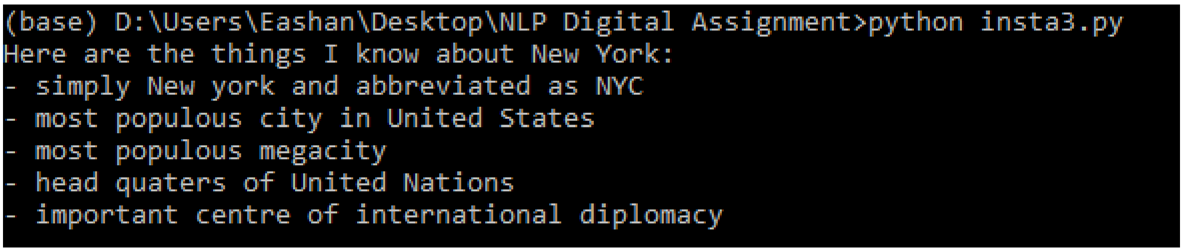
```python
import spacy
import en_core_web_sm

nlp = en_core_web_sm.load()
# Load the large English NLP model

# The text we want to examine
text = """New York City, often called simply New York and abbreviated as NYC, is the most populous city in the
United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles, New York City
is also the most densely populated major city in the United States. Located at the southern tip of the U.S. state of New
York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban
landmass. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined
statistical area, it is one of the world's most populous megacities. New York City has been described as the cultural,
financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology,
education, politics, tourism, art, fashion, and sports. Home to the headquarters of the United Nations, New York is an
important center for international diplomacy.
"""

# Parse the text with spaCy. This runs the entire pipeline.
doc = nlp(text)

# 'doc' now contains a parsed version of text. We can use it to do anything we want!
# For example, this will print out all the named entities that were detected:
for entity in doc.ents:
    print("{0} ({1})".format(entity.text, entity.label_))

```



```
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>python insta.py
New York City (GPE)
New York (GPE)
NYC (ORG)
United States (GPE)
an estimated 2019 (CARDINAL)
8,336,817 (CARDINAL)
about 302.6 square miles (QUANTITY)
New York City (GPE)
the United States (GPE)
U.S. (GPE)
New
York (GPE)
New York (GPE)
almost 20 million (CARDINAL)
approximately 23 million (CARDINAL)
New York City (GPE)
the United Nations (ORG)
New York (GPE)
```



```

import spacy
import textacy.extract

# Load the large English NLP model

import en_core_web_sm

nlp = en_core_web_sm.load()

# The text we want to examine
text = """
New York City, often called simply New York and abbreviated as NYC, is the most populous city in the
United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles, New York City
is also the most densely populated major city in the United States. Located at the southern tip of the U.S. state of New
York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban
landmass. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined
statistical area, it is one of the world's most populous megacities. New York City has been described as the cultural,
financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology,
education, politics, tourism, art, fashion, and sports. Home to the headquarters of the United Nations, New York is an
important center for international diplomacy.
"""

# Parse the document with spaCy
doc = nlp(text)

# Extract semi-structured statements
statements = textacy.extract.semistructured_statements(doc, "London")

# Print the results
print("Here are the things I know about New York:")

for statement in statements:
    subject, verb, fact = statement
    print(" - {0}".format(fact))
```



```
(base) D:\Users\Eashan\Desktop\NLP Digital Assignment>python insta3.py
Here are the things I know about New York:
- simply New york and abbreviated as NYC
- most populous city in United States
- most populous megacity
- head quaters of United Nations
- important centre of international diplomacy
```