Spatial Databases Exercise : Geoparsing

1. ***Read the contents of the file JackHagelExerciseText.txt and list those named entities that you regard as geographical places.***
   Could include (but subject to discussion):

```
World Trade Center
New York
Raleigh
Lower Manhattan
Washington
Country Club Plaza
Kansas City
Mo.
Raleigh
Cameron Village
Southeast
```

2. ***Write a python script to use the spaCy Named Entity Recognition tool to list all named entities in the file.***

Note:
spaCy can be installed with pip, recommended usage as follows:

```
  pip install -U spacy
```

See : https://spacy.io/usage

The following is an example from the spaCy web site (https://spacy.io/usage/linguistic-features) on basic usage of the NER tool.

<<<<<

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

>>>>>>

The label property stores the entity type as a code such as ORG. See lecture notes for some of the other codes.

If you paste the text of the file into a python script you might have problems with the quote characters that could need to be escaped. This should not be a problem if you read from the file in your python script.

Answer : Named entities listed by spaCy:

```
Jack Hagel PERSON
the World Trade Center ORG
New York GPE
Raleigh ORG
Smedes York PERSON
Urban Land Institute ORG
the World Trade Center and Lower Manhattan Summit ORG
last month DATE
York PERSON
the Urban Land Institute ORG
Washington GPE
1989 to 1991 DATE
J.W. "Willie" York PERSON
the Urban Land Institute ORG
1947 DATE
J.C. Nichols PERSON
Country Club Plaza ORG
Kansas City GPE
Mo. GPE
Willie York PERSON
Raleigh ORG
Cameron Village GPE
Southeast LOC
first ORDINAL
```

The following is a Python script to output the above.
Note that the input file 'JackHagelExerciseText-NotEscaped.txt' has no escape characters in it, while the commented out quoted text in the script does use escape characters

```
import spacy
import json
import os

nlp = spacy.load("en_core_web_sm")

infile = open('JackHagelExerciseText-NotEscaped.txt', "r")
theText = infile.read()
infile.close()

doc = nlp(theText)

#doc = nlp("Jack Hagel, Staff Writer Redevelopment of the World
Trade Center site in New York is getting some input from a Raleigh
real-estate maven.  York Properties President Smedes York was
chairman of an Urban Land Institute panel at the World Trade Center
and Lower Manhattan Summit last month. York was chairman of the
Urban Land Institute, a Washington nonprofit organization, from 1989
to 1991. His dad, J.W. \"Willie\"  York, joined the Urban Land
Institute in 1947. That\'s where he met   J.C. Nichols, the
developer of Country Club Plaza in Kansas City,   Mo. - the center
that inspired Willie York to build Raleigh\'s Cameron Village, the
Southeast\'s first shopping center.")

for ent in doc.ents:
	place = ent.text
	print ("Geo-place: " + str(place) + " " + str(ent.label_)
```

3. ***Modify your script to output only those entities that can be regarded as geographic. These are ones with NER types of GPE, FAC or LOC.***
   List the places in the text that spaCy has failed to categorise correctly as geographic places. How were these missed places categorised if at all?

Solution:
Places identified as GPE, FAC or LOC

```
  New York GPE
  Washington GPE
  Kansas City GPE
  Mo. GPE
  Cameron Village GPE
  Southeast LOC
```

```
Places with other categorisation:
World Trade Center (though doesn’t exist now) – ORG
Lower Manhattan – ORG as part of World Trade Centre entity
Raleigh – ORG
Country Club Plaza – ORG
```

4. ***Enter the text in the file JackHagelExerciseText.txt into the demonstration of the Edinburgh Geoparser software at http://jekyll.inf.ed.ac.uk/geoparser/***
   Consider how well it has succeeded in identifying the place names. List any places that you think that it has failed to identify as geographic places.
   Of the places that it has geocoded – are all of these correctly geocoded?

Solution:

It has failed to identify
Lower Manhattan – but it does identify Manhattan
Country Club Plaza
Cameron Village – but it misidentifies Cameron as a place
Southeast – but this is a vague region

It has made mistakes in geocoding:

- Manhattan – it finds somewhere in Kansas but should be in New York
- Cameron - it finds somewhere in Kansas but the Cameron Village referred to in the text is actually in North Carolina, inside Raleigh, as stated in the text “Raleigh's Cameron Village”
- New York – it has geocoded the state of New York but the text refers to the city.

5. ***Extend your python script to use the GeoPy geocoder tool to attach coordinates to each of the places that spaCy identified as either GPE, FAC or LOC. Use the Nominatim gazetteer and select only the first geocoded place returned by GeoPy – thus set “Limit = 1”***
   When applied to the places that SpaCy identified as either GPE, FAC or LOC, how well has the default first choice geocoded location succeeded in finding the correct place?

   Notes for this question:
   The GeoPy library can be installed with pip as follows (see https://pypi.org/project/geopy/)

   ```
    pip install geopy
   ```

   The following is an example of using GeoPy with the Nominatum gazetteer – taken directly from the GeoPy website at https://pypi.org/project/geopy/

   ```
   from geopy.geocoders import Nominatim
   geolocator =
   Nominatim(user_agent="specify_your_app_name_here")
   location = geolocator.geocode("175 5th Avenue NYC")
   print(location.address)
   >>>Flatiron Building, 175, 5th Avenue, Flatiron, New
   York, NYC, New York, ...
   print((location.latitude, location.longitude))
   >>>(40.7410861, -73.9896297241625)
   ```

   Solutions:
   GPE, FAC or LOC geocoded places

```
Nominatim Geo-place: New York
New York, United States of America
40.7127281 -74.0060152
>>>>> This is the state not the city

Nominatim Geo-place: Washington
Washington, District of Columbia, 20230, United States of America
38.8949924 -77.0365581
>>>>> Correct

Nominatim Geo-place: Kansas City
Kansas City, Jackson County, Missouri, United States of America
39.100105 -94.5781416
>>>>> Correct

Nominatim Geo-place: Mo.
Missouri, United States of America
38.7604815 -92.5617875
>>>>> Correct


Nominatim Geo-place: Cameron Village
Cameron Village, Baltimore, Maryland, 21239, United States of America
39.356734 -76.5992851
```

```
>>>>> error – the Cameron Village in the text is in Raleigh in North
Carolina, not in Baltimore, Maryland

Nominatim Geo-place: Southeast
Sverige
59.6749712 14.5208584

>>>>> error – the Southeast in the text is the south region of the USA,
not somewhere in Sweden.
```

6. ***Modify the script in 5 to treat entities classes as ORG as places and consider how well it has geocoded these organisations.***

Solution:

```
Nominatim Geo-place: the World Trade Center
World Trade Center, 180, Greenwich Street, Financial District, Manhattan
Community Board 1, Westfield World Trade Center, New York County, New
York, 10048, United States of America
40.7118877 -74.0124412
>>>>> Correct

Nominatim Geo-place: Raleigh
Raleigh, Wake County, North Carolina, United States of America
35.7803977 -78.6390989
>>>>> Correct

Nominatim Geo-place: Country Club Plaza
Country Club Plaza, Kansas City, Jackson County, Missouri, 64112, United
States of America
39.0420441 -94.5927959
>>>>> Correct

Nominatim Geo-place: Raleigh
Raleigh, Wake County, North Carolina, United States of America
35.7803977 -78.6390989
>>>>> Correct
```