Hands on NLP

NLP with NLTK in Python

Marco Petolicchio

November 2020

Setting up the environment

Installing NLTK

NLTK is an acronym for *Natural Language ToolKit*, and represents a useful pipeline to process natural language and retrieve quantitative information from texts.

After having installed Python on the machine, it will be necessary to install NTLK.

For Mac and Linux launch the following command from Terminal:

```
pip3 install --user -U nltk numpy ssl
```

Then enter in Python from terminal running `python3`. Now you are into Python environment, run:

```
import nltk

nltk.download()
```

The method `ntlk.download()` should open a new window. If you get an SSL trouble, launch this in python:

```
import nltk
import ssl

try:
    _create_unverified_https_context = ssl._create_unverified_context
except AttributeError:
    pass
else:
    ssl._create_default_https_context = _create_unverified_https_context

nltk.download()
```

Now you can select what do you want to install from NLTK: corpora, models, tagsets, and so on.

For now, we can still with the default environment, then you can close the window.

Preparing the data

In order to process the data, we need to clean it from non significative parts. For example, articles and very common words as prepositions are not significant for obtaining quantitative information about texts.

Lowercase

```
text = ”Hello World”

text = text.lower()

print(text)
```

Remove punctuation

```
import string

print(string.punctuation)

text_p = ””.join([char for char in text if char not in string.punctuation])

print(text_p)
```

Tokenization

```
from nltk import word_tokenize

words = word_tokenize(text_p)

print(words)
```

Stopword filtering

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
print(stop_words)

filtered_words = [word for word in words if word not in stop_words]
print(filtered_words)
```

POS Tagger

```
from nltk import pos_tag

pos = pos_tag(filtered_words)

print(pos)
```

Let’s have fun

In this exercise we will plot the frequency of the words from the corpus `brown` in nltk corpora.

```
import nltk
import matplotlib
from nltk import FreqDist
from nltk.corpus import brown

fd = FreqDist(brown.words())
fd.plot(30)
```