

# Ejemplo

February 3, 2024

```
[1]: !pip install nltk
```

Collecting nltk

Downloading nltk-3.8.1-py3-none-any.whl (1.5 MB)

1.5/1.5 MB 3.0 MB/s eta 0:00:00[31m2.6 MB/s eta  
0:00:010m

Collecting click (from nltk)

Using cached click-8.1.7-py3-none-any.whl.metadata (3.0 kB)

Requirement already satisfied: joblib in /home/reroes/entornos/envpy310-ia-  
maestria/lib/python3.10/site-packages (from nltk) (1.3.2)

Collecting regex>=2021.8.3 (from nltk)

Downloading regex-2023.12.25-cp310-cp310-  
manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (40 kB)

40.9/40.9 kB 1.4 MB/s eta 0:00:00

Collecting tqdm (from nltk)

Downloading tqdm-4.66.1-py3-none-any.whl.metadata (57 kB)

57.6/57.6 kB 3.1 MB/s eta 0:00:00

Downloading

regex-2023.12.25-cp310-cp310-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl (773  
kB)

774.0/774.0 kB 9.4 MB/s eta 0:00:000m eta

0:00:01[36m0:00:01

Using cached click-8.1.7-py3-none-any.whl (97 kB)

Downloading tqdm-4.66.1-py3-none-any.whl (78 kB)

78.3/78.3 kB 4.1 MB/s eta 0:00:00

Installing collected packages: tqdm, regex, click, nltk

Successfully installed click-8.1.7 nltk-3.8.1 regex-2023.12.25 tqdm-4.66.1

```
[ ]: import nltk  
nltk.download('punkt')  
nltk.download('cess_esp') # Para etiquetado POS en español
```

```
[6]: texto = "Mi maestría la estudio en la Universidad Técnica Particular de Loja.␣  
      ↪Loja es una provincia del Ecuador."
```

```
[7]: from nltk.tokenize import word_tokenize  
  
palabras = word_tokenize(texto, language='spanish')  
print(palabras)
```

```
['Mi', 'maestría', 'la', 'estudio', 'en', 'la', 'Universidad', 'Técnica',  
'Particular', 'de', 'Loja', '.', 'Loja', 'es', 'una', 'provincia', 'del',  
'Ecuador']
```

```
[8]: from nltk.tokenize import sent_tokenize  
  
oraciones = sent_tokenize(texto, language='spanish')  
print(oraciones)
```

```
['Mi maestría la estudio en la Universidad Técnica Particular de Loja.', 'Loja  
es una provincia del Ecuador']
```

```
[16]: from nltk.corpus import cess_esp  
      from nltk import pos_tag  
      from nltk.tokenize import word_tokenize  
  
      # Se carga el conjunto de entrenamiento de cess_esp  
      cess_esp.ensure_loaded()  
      sents = cess_esp.tagged_sents()  
  
      # Se entrena un etiquetador TnT con base al corpus cess_esp  
      from nltk.tag import tnt  
      tnt_pos_tagger = tnt.TnT()  
      tnt_pos_tagger.train(sents)  
  
      palabras = word_tokenize(texto, language='spanish')  
  
      # Se etiqueta las palabras tokenizadas  
      etiquetado = tnt_pos_tagger.tag(palabras)
```

```
[17]: for t in etiquetado:  
      print(t)
```

```
('Mi', 'dplcss')  
('maestría', 'ncfs000')  
('la', 'da0fs0')  
('estudio', 'ncms000')  
('en', 'sps00')  
('la', 'da0fs0')  
('Universidad', 'np0000o')
```

('Técnica', 'Unk')  
('Particular', 'Unk')  
('de', 'sps00')  
('Loja', 'Unk')  
('.', 'Fp')  
('Loja', 'Unk')  
('es', 'vsip3s0')  
('una', 'di0fs0')  
('provincia', 'ncfs000')  
('del', 'spcms')  
('Ecuador', 'np0000o')