



# UNIVERSIDAD DE ANTIOQUIA

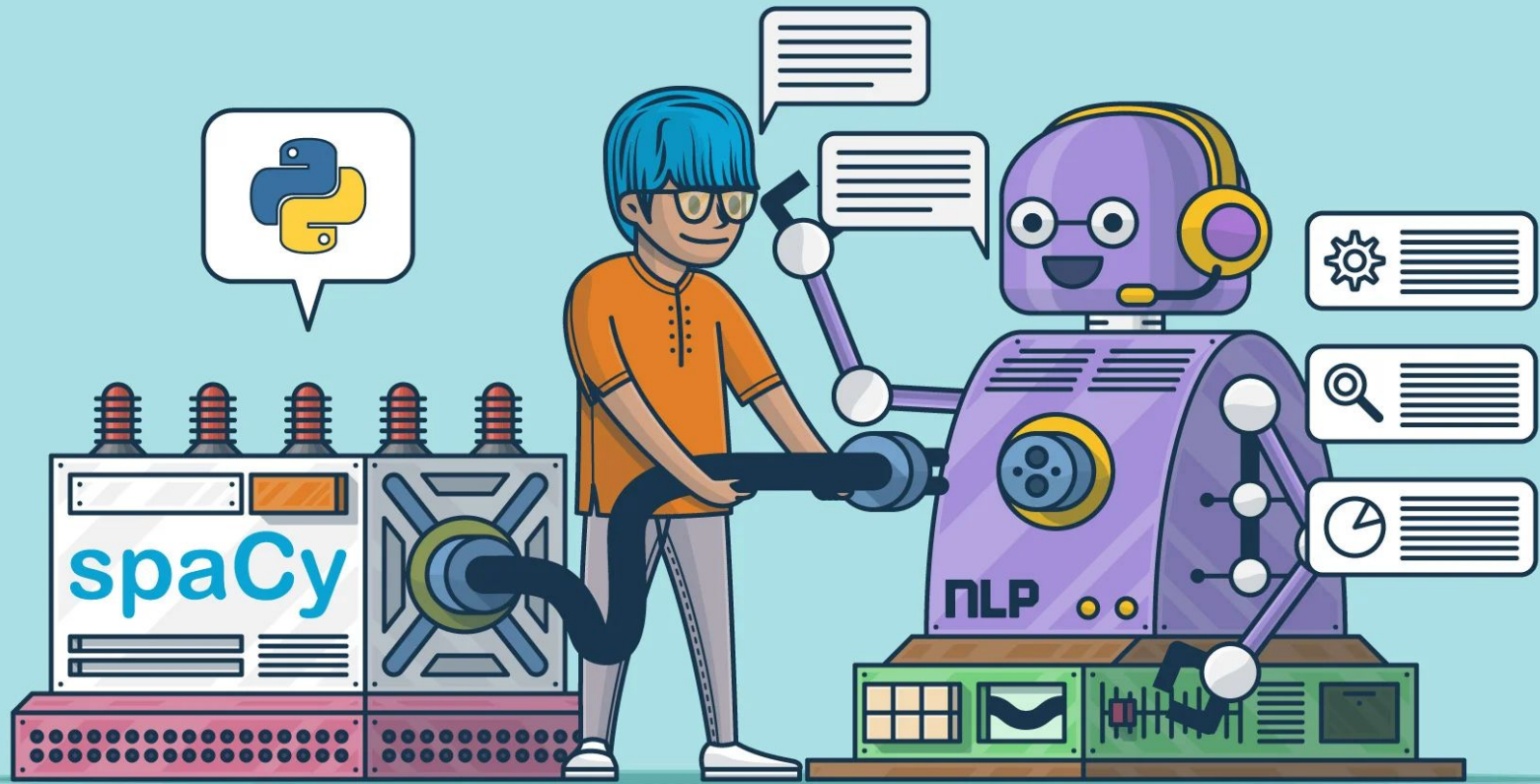


## Spacy en Python Google Colab

Curso de Estadística y  
Lingüística Computacional

Jorge Mauricio Molina Mejía  
Dr. Informática y Ciencias del Lenguaje

# Spacy



Real Python

# Generalidades

SpaCy es una serie de librerías y programas que tienen como finalidad el tratamiento y análisis del lenguaje natural para la lengua inglesa y otras lenguas.

SpaCy no se ejecuta como otros programas en *Windows*, es decir, no es posible abrir una ventana y seleccionar unas opciones. En cambio, es necesario ejecutar el programa desde la consola de *Windows* o mediante un programa ligado a *Python* (tipo de lenguaje de programación) y precisar las opciones con comandos de esta consola o en el programa.

# Ejecución en Google Colab

Pasos para activar Python en Google Colab:

- `pip install --upgrade pip`
- `pip install --upgrade ipython jupyter`

Pasos para ejecutar SpaCy en Google Colab:

- `pip install -U pip setuptools wheel`
- `pip install -U spacy`
- `!python -m spacy download en_core_web_sm`
- `!python -m spacy download es_core_news_sm`



# Funcionamiento I

Escribir el siguiente código para importar las librerías:

```
import spacy  
from spacy import displacy
```

Luego cargar el corpus para el español:

```
nlp = spacy.load('es_core_news_sm')
```



# Funcionamiento II

Escibir un texto cualquiera en una variable:

```
text1 = "Los niños de Martha juegan en la  
escuela. La escuela es muy pequeña. Allí se  
encuentran todos luego de las clases."
```

Cargar el texto:

```
doc1 = nlp(text1)
```

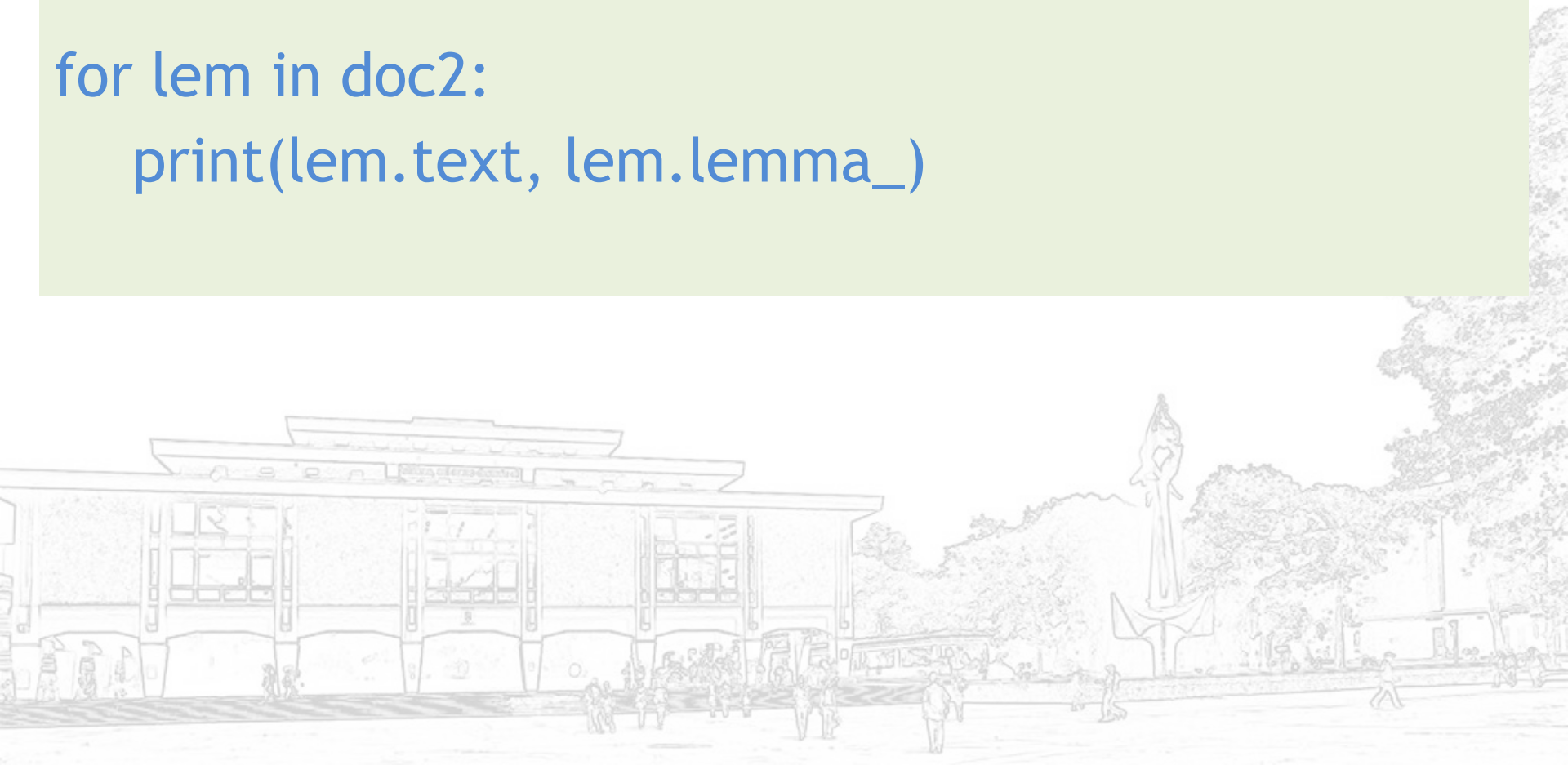
For token in doc1:

```
    print(token)
```

# Funcionamiento III

```
doc2 = nlp("Los niños que estudiaban en la escuela jugaban con varios balones en el patio de la escuela.")
```

```
for lem in doc2:  
    print(lem.text, lem.lemma_)
```



# Funcionamiento IV

```
for token in doc1:  
    print(token.text, token.pos_)
```

```
for token in doc1:  
    print(token.text, token.pos_, token.dep_)
```





# Funcionamiento V

```
import spacy
from spacy import displacy

nlp = spacy.load('es_core_news_sm')
doc3 = nlp(u'Los niños juegan en la calle.')
displacy.render(doc3, style='dep', jupyter=True, options={'distance': 140})
```

```
import spacy
from spacy import displacy

nlp = spacy.load('es_core_news_sm')
displacy.render(doc1, style='dep', jupyter=True, options={'distance': 140})
```

# Funcionamiento VI

```
nlp = spacy.load('es_core_news_sm')  
displacy.render(doc2, style='dep', jupyter=True, options  
={'distance': 140})
```



# Funcionamiento VII

```
import spacy
from spacy import displacy
```

```
text2 = "El señor Jorge Mauricio Molina Mejía nació en la ciudad d  
e Medellín, Colombia el 15 de abril de 1972. Estudió en la Universi  
dad de Antioquia y trabajó en la ONU en el año 2010. Hizo el doct  
orado en Francia en la Universidad Grenoble Alpes allí conoció a A  
gnès Souque."
```

```
nlp = spacy.load('es_core_news_sm')
```

```
doc4 = nlp(text2)
```

```
displacy.render(doc4, style='ent', jupyter=True, options={'distance  
': 140})
```



# Funcionamiento VIII

```
import spacy
from spacy import displacy

nlp = spacy.load('es_core_news_sm')
file_name = '/LComp.txt'
Lcomp_file_text = open(file_name).read()
Lcomp_file_doc = nlp(Lcomp_file_text)
print ([token.text for token in Lcomp_file_doc])
```



# Funcionamiento IX

```
import spacy
from spacy import displacy

nlp = spacy.load('es_core_news_sm')
file_name = '/LComp.txt'
Lcomp_file_text = open(file_name).read()
Lcomp_file_doc = nlp(Lcomp_file_text)
sentences = list(Lcomp_file_doc.sents)
len(sentences)

for sentence in sentences:
    print (sentence)
```



# Funcionamiento X

```
for token in Lcomp_file_doc:  
    print (token, token.idx)
```



# Funcionamiento XII

```
from spacy.lang.es.stop_words import STOP_WORDS
stopwords_spacy = list(STOP_WORDS)
print(stopwords_spacy)
len(stopwords_spacy)
```

```
for token in Lcomp_file_doc:
    if token.is_stop == False:
        print(token)
```

```
displacy.render(Lcomp_file_doc, style='ent', jupyter=True, options={'distance': 140})
```

```
displacy.render(Lcomp_file_doc, style='dep', jupyter=True, options={'distance': 140})
```