Tokenizacion

Es el segmentar contenido en unidades mas pequeñas. Un libro en capitulos a parrafos a frases a palabras y caracteres.

Para este ejemplo, usare la sinopsis de la pelicula de "Memorias de un Caracol" del portal Filmaffinity

```
text = "Australia, años 70. Grace Pudel es una niña solitaria e inadaptada, aficionada a coleccionar figuras decorativas de caracoles y con una devoción profunda por las r
```

```
# Tokenizacion por frases
sentences = text.split('.')

for idx, word in enumerate(sentences):
    print('Frase {0:5}{1:5}'.format(str(idx), word))
    if idx==5:  # 5 Primeras frases
        break
```

```
Frase 0    Australia, años 70
Frase 1     Grace Pudel es una niña solitaria e inadaptada, aficionada a coleccionar figuras decorativas de caracoles y con una devoción profunda por las novelas romár
Frase 2     La muerte de su padre cuando tan solo es una niña, la lleva a tener que separarse de su hermano mellizo, Gilbert, lo que la aboca a una espiral de ansiedac
Frase 3     Sin embargo, la esperanza vuelve a su vida cuando conoce a una excéntrica anciana llena de determinación y amor por la vida llamada Pinky, con la que entat
Frase 4
```

A continuacion se tokenizara pero por palabras, siendo nuestro delimitador el espacio al final de cada secuencia de caracteres. (" ")

```
# Tokenizacion por palabras
sentences = text.split(' ')

for idx, word in enumerate(sentences):
    print('Palabra {0:5}{1:5}'.format(str(idx), word))
    if idx==5:  # mostramos solo las 5 primeras palabras
        break
```

```
Palabra 0    Australia,
Palabra 1    años
Palabra 2    70.
Palabra 3    Grace
Palabra 4    Pudel
Palabra 5    es
```

A continuacion usaremos la libreria NLTK , la cual podemos usar como alternativa en lugar de los arreglos y slices tradicionales de python

```
# Instalacion de liberias

import nltk
nltk.download('punkt_tab')
```

```
from nltk.tokenize import sent_tokenize, word_tokenize
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
```

```
# Mostramos la lista de frases usando el tokenizdor de nltk
print(sent_tokenize(text))
```

```
['Australia, años 70.', 'Grace Pudel es una niña solitaria e inadaptada, aficionada a coleccionar figuras decorativas de caracoles y con una devoción profunda por las
```

```
# Mostramos la lista de palabras usando el tokenizdor de nltk
print(word_tokenize(text))
```

```
['Australia', ',', 'años', '70', '.', 'Grace', 'Pudel', 'es', 'una', 'niña', 'solitaria', 'e', 'inadaptada', ',', 'aficionada', 'a', 'coleccionar', 'figuras', 'decorat
```

CONCLUSION

Con este ejemplo se ve lo básico pero clave del preprocesamiento: dividir un texto largo en partes más manejables. Primero se hace a lo simple con split() de Python, separando por puntos o espacios. Después se mete NLTK, que ya es una librería más dedicada para esto, y permite dividir el texto de forma más precisa en frases y palabras.