# Archivos en Internet

Esto no es web-scrapping, pero también podemos descargar archivos que están en Internet desde un programa y procesarlos como los archivos de disco.

Hay varias librerías: `urllib` , `urllib2` , `urllib3` , etc.

In [1]:
```
# La librería para acceder a archivos en Internet:

import requests

# Elijo el Quijote, cómo no, disponible en la librería Gutenberg:

url_quijote = "https://www.gutenberg.org/cache/epub/2000/pg2000.txt"
f = requests.get(url_quijote)

# Veamos la codificación en que está este texto:
print(f.encoding)

# Y veamos ahora unas pocas líneas:

texto = f.text.split('\n')
for line in texto[613:620]:
    print(line)
```

```
utf-8
En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha muc
ho
tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigu
a,
rocín flaco y galgo corredor. Una olla de algo más vaca que carnero,
salpicón las más noches, duelos y quebrantos los sábados, lantejas los
viernes, algún palomino de añadidura los domingos, consumían las tres
partes de su hacienda. El resto della concluían sayo de velarte, calzas
de
velludo para las fiestas, con sus pantuflos de lo mesmo, y los días de
```

**Observaciones**:

Ya puestos, vamos a seleccionar las palabras y a limpiarlas, por el principio y por el final, de los caracteres de puntuación:

In [2]:
```
palabras = "Dijo Don Quijote: ¿Qué tal Sancho? Bien señor, pero rendido;

print([palabra for palabra in palabras.split()])
print([palabra.strip(" ,;:.¡!¿?") for palabra in palabras.split()])
```

```
['Dijo', 'Don', 'Quijote:', '¿Qué', 'tal', 'Sancho?', 'Bien', 'señor,',
'pero', 'rendido;', 'y', 'dolorido', 'por', 'los', 'lances', 'del', 'dí
a...']
['Dijo', 'Don', 'Quijote', 'Qué', 'tal', 'Sancho', 'Bien', 'señor', 'pe
ro', 'rendido', 'y', 'dolorido', 'por', 'los', 'lances', 'del', 'día']
```

In [3]:

```
# Hagámoslo ahora con un fragmento de nuestro Quijote:

for linea in texto[613:625]:
    palabras = linea.strip().split()
    print(palabras)
    pal_limpias = [palabra.strip(" ,;:.¡!¿?") for palabra in palabras]
    print(pal_limpias)
    print()
```

```
['En', 'un', 'lugar', 'de', 'la', 'Mancha,', 'de', 'cuyo', 'nombre', 'n
o', 'quiero', 'acordarme,', 'no', 'ha', 'mucho']
['En', 'un', 'lugar', 'de', 'la', 'Mancha', 'de', 'cuyo', 'nombre', 'n
o', 'quiero', 'acordarme', 'no', 'ha', 'mucho']

['tiempo', 'que', 'vivía', 'un', 'hidalgo', 'de', 'los', 'de', 'lanza',
'en', 'astillero,', 'adarga', 'antigua,']
['tiempo', 'que', 'vivía', 'un', 'hidalgo', 'de', 'los', 'de', 'lanza',
'en', 'astillero', 'adarga', 'antigua']

['rocín', 'flaco', 'y', 'galgo', 'corredor.', 'Una', 'olla', 'de', 'alg
o', 'más', 'vaca', 'que', 'carnero,']
['rocín', 'flaco', 'y', 'galgo', 'corredor', 'Una', 'olla', 'de', 'alg
o', 'más', 'vaca', 'que', 'carnero']

['salpicón', 'las', 'más', 'noches,', 'duelos', 'y', 'quebrantos', 'lo
s', 'sábados,', 'lantejas', 'los']
['salpicón', 'las', 'más', 'noches', 'duelos', 'y', 'quebrantos', 'lo
s', 'sábados', 'lantejas', 'los']

['viernes,', 'algún', 'palomino', 'de', 'añadidura', 'los', 'domingo
s,', 'consumían', 'las', 'tres']
['viernes', 'algún', 'palomino', 'de', 'añadidura', 'los', 'domingos',
'consumían', 'las', 'tres']

['partes', 'de', 'su', 'hacienda.', 'El', 'resto', 'della', 'concluía
n', 'sayo', 'de', 'velarte,', 'calzas', 'de']
['partes', 'de', 'su', 'hacienda', 'El', 'resto', 'della', 'concluían',
'sayo', 'de', 'velarte', 'calzas', 'de']

['velludo', 'para', 'las', 'fiestas,', 'con', 'sus', 'pantuflos', 'de',
'lo', 'mesmo,', 'y', 'los', 'días', 'de']
['velludo', 'para', 'las', 'fiestas', 'con', 'sus', 'pantuflos', 'de',
'lo', 'mesmo', 'y', 'los', 'días', 'de']

['entresemana', 'se', 'honraba', 'con', 'su', 'vellorí', 'de', 'lo', 'm
ás', 'fino.', 'Tenía', 'en', 'su', 'casa', 'una']
['entresemana', 'se', 'honraba', 'con', 'su', 'vellorí', 'de', 'lo', 'm
ás', 'fino', 'Tenía', 'en', 'su', 'casa', 'una']

['ama', 'que', 'pasaba', 'de', 'los', 'cuarenta,', 'y', 'una', 'sobrin
a', 'que', 'no', 'llegaba', 'a', 'los', 'veinte,']
['ama', 'que', 'pasaba', 'de', 'los', 'cuarenta', 'y', 'una', 'sobrin
a', 'que', 'no', 'llegaba', 'a', 'los', 'veinte']

['y', 'un', 'mozo', 'de', 'campo', 'y', 'plaza,', 'que', 'así', 'ensill
aba', 'el', 'rocín', 'como', 'tomaba', 'la']
['y', 'un', 'mozo', 'de', 'campo', 'y', 'plaza', 'que', 'así', 'ensilla
ba', 'el', 'rocín', 'como', 'tomaba', 'la']

['podadera.', 'Frisaba', 'la', 'edad', 'de', 'nuestro', 'hidalgo', 'co
n', 'los', 'cincuenta', 'años;', 'era', 'de']
['podadera', 'Frisaba', 'la', 'edad', 'de', 'nuestro', 'hidalgo', 'co
n', 'los', 'cincuenta', 'años', 'era', 'de']

['complexión', 'recia,', 'seco', 'de', 'carnes,', 'enjuto', 'de', 'rost
ro,', 'gran', 'madrugador', 'y', 'amigo']
['complexión', 'recia', 'seco', 'de', 'carnes', 'enjuto', 'de', 'rostr
o', 'gran', 'madrugador', 'y', 'amigo']
```

# Referencias

Ya decíamos que esto no es *web scrapping*. Si esto te ha gustado, apriétate el cinturón y sigue el camino con Beautiful Soup, o amplía con el manejo de expresiones regulares, o si estás interesado en el procesamiento de texto escrito en español o inglés, adeante con nltk...