Facultad de Ciencias Matemáticas - Informática

Universidad Complutense de Madrid Curso 2014/2015

Práctica 3: Procesamiento de páginas web Fecha entrega: 23/03/2015

En esta práctica vamos a estudiar cómo procesar de manera sencilla cadenas y cómo trabajar con distintos tipos de ficheros. Para ello, es necesario primero saber cómo funcionan las páginas web. Observa, por ejemplo, esta página, mostrada en la figura 1.

Inicio / Categoría / Económica / Bienes inmuebles /

Listado de Bienes Inmuebles de Ministerio de Sanidad, Servicios Sociales e Igualdad

Relación de Bienes Inmuebles de Ministerio de Sanidad, Servicios Sociales e Igualdad, en todo el territorio nacional

Bienes inmuebles

Tipo	Localizacion	Uso	Ministerio	Superficie
Edificio	Madrid CR. POZUELO-MAJADAHONDA	AGENCIA ESPAÑOLA DE CONSUMO, SEGURIDAD ALIMENTARIA Y NUTRICIÓN (AECOSAN)	Ministerio de Sanidad, Servicios Sociales e Igualdad	6202,26 m^2
Edificio	Madrid CALLE HORTALEZA, 104 - 2º IZD.	INSTITUTO DE LA MUJER Y PARA LA IGUALDAD DE OPORTUNIDADES	Ministerio de Sanidad, Servicios Sociales e Igualdad	353 m^2
Edificio	Madrid JESUS Y MARIA, 13/ CABEZA, 4	MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD	Ministerio de Sanidad, Servicios Sociales e Igualdad	749 m^2
Edificio	Madrid PLAZA DE ESPAÑA, 17	MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD	Ministerio de Sanidad, Servicios Sociales e Igualdad	2591,28 m^2
Edificio	Madrid CALLE SERRANO, 150	INSTITUTO DE LA MUJER Y PARA LA IGUALDAD DE OPORTUNIDADES	Ministerio de Sanidad, Servicios Sociales e Igualdad	5565 m^2
Edificio	Madrid CAYETANO PANDO 19	INSTITUTO DE LA MUJER Y PARA LA IGUALDAD DE OPORTUNIDADES	Ministerio de Sanidad, Servicios Sociales e Igualdad	386 m^2
Edificio	Madrid PLAZA COMENDADORAS, 6 / GARAJE, ALMACEN Y LOCAL	INSTITUTO DE LA JUVENTUD	Ministerio de Sanidad, Servicios Sociales e Igualdad	398,62 m^2
Edificio	Madrid PZ GOLETA ,2 (BARAJAS) Y AV. DE CANTABRIA S/N	AGENCIA ESPAÑOLA DE CONSUMO, SEGURIDAD ALIMENTARIA Y NUTRICIÓN (AECOSAN)	Ministerio de Sanidad, Servicios Sociales e Igualdad	4350 m^2
Edificio	Madrid CALLE ALCALÁ, 56	AGENCIA ESPAÑOLA DE CONSUMO, SEGURIDAD ALIMENTARIA Y NUTRICIÓN (AECOSAN)	Ministerio de Sanidad, Servicios Sociales e Igualdad	7891 m^2
Edificio	Madrid P CASTELLANA 27 Y MARQUES RISCAL 16	INSTITUTO DE LA JUVENTUD	Ministerio de Sanidad, Servicios Sociales e Igualdad	1323 m^2
Edificio	Madrid PASEO DE LA CASTELLANA 67 C/V C/ AGUSTIN DE BETHENCOURT 4	MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD	Ministerio de Sanidad, Servicios Sociales e Igualdad	1000 m^2

Figura 1: Página web del Portal de Transparencia

Si haces click en el botón derecho del ratón y seleccionas la opción `Ver código fuente de la página` verás que la representación interna es muy diferente a la representación gráfica, como se puede ver en la figura 2. La información que ves en esta segunda imagen está escrita en `html`, un lenguaje de marcado usado para la creación de páginas web. En este lenguaje, la información se escribe entre etiquetas de la forma `<etiqueta>`, para indicar el inicio del bloque, y `</etiqueta>`, para indicar el final. Por ejemplo, en la imagen de la figura 2 podemos ver que la cabecera de la tabla está dentro de una etiqueta `<thead>`, que empieza en la línea 363 y termina en la línea 371. Dentro de esta etiqueta tenemos un solo campo `tr` (líneas 363-370), que sirve para identificar filas de la tabla. Por último, dentro tenemos 5 campos `<th>`, que delimitan los elementos en cada columna.

De manera similar, el cuerpo de la tabla se encuentra encerrado por etiquetas `<tbody>`. Dentro encontramos de nuevo etiquetas `<tr>` para cada fila, mientras que cada elemento está delimitado por etiquetas `<td>`.

Nuestra intención en esta práctica es transformar esta página web (o cualquier otra similar, como esta o esta) en un fichero más fácil de procesar. En particular, estamos interesados en escribir un fichero de texto en el que las distintas columnas estén separadas por `;`, mientras que las filas se separarán por saltos de línea. Es decir, la página web de la primera imagen quedará guardada como:[1]

[1]Las líneas aparecen partidas para que quepan en la hoja, pero en el fichero habrá una sola línea por cada fila de la tabla.

```
<section class="tr-article--content">
    <table class="mf-table-data mf-table-data__zebra">
<caption>Bienes inmuebles</caption>
<thead>
    <tr>
        <th>Tipo</th>
        <th>Localizacion</th>
        <th>Uso</th>
        <th>Ministerio</th>
        <th>Superficie</th>
    </tr>
</thead>
<tbody>
                    <tr>
        <td>Edificio</td>
        <td>Madrid <span class="tr-data--subtitle">CR. POZUELO-MAJADAHONDA</span></td>
        <td>AGENCIA ESPAÑOLA DE CONSUMO, SEGURIDAD ALIMENTARIA Y NUTRICIÓN (AECOSAN)</td>
        <td>Ministerio de Sanidad, Servicios Sociales e Igualdad</td>
        <td class="tr-cell__measure tr-cell__num">6202,26 <span class="tr-data__unit">m<sup>2</sup></span></td>
    </tr>
                    <tr>
        <td>Edificio</td>
        <td>Madrid <span class="tr-data--subtitle">CALLE HORTALEZA, 104 - 2º IZD.</span></td>
        <td>INSTITUTO DE LA MUJER Y PARA LA IGUALDAD DE OPORTUNIDADES</td>
        <td>Ministerio de Sanidad, Servicios Sociales e Igualdad</td>
        <td class="tr-cell__measure tr-cell__num">353 <span class="tr-data__unit">m<sup>2</sup></span></td>
    </tr>
                    <tr>
        <td>Edificio</td>
        <td>Madrid <span class="tr-data--subtitle">JESUS Y MARIA, 13/ CABEZA, 4</span></td>
        <td>MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD</td>
        <td>Ministerio de Sanidad, Servicios Sociales e Igualdad</td>
        <td class="tr-cell__measure tr-cell__num">749 <span class="tr-data__unit">m<sup>2</sup></span></td>
    </tr>
                    <tr>
        <td>Edificio</td>
        <td>Madrid <span class="tr-data--subtitle">PLAZA DE ESPAÑA, 17</span></td>
        <td>MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD</td>
        <td>Ministerio de Sanidad, Servicios Sociales e Igualdad</td>
        <td class="tr-cell__measure tr-cell__num">2591,28 <span class="tr-data__unit">m<sup>2</sup></span></td>
    </tr>
                    <tr>
```

Figura 2: Código fuente para la página web de la figura 1

```
Tipo;Localizacion;Uso;Ministerio;Superficie
Edificio;Madrid CR. POZUELO-MAJADAHONDA;AGENCIA ESPAÑOLA DE CONSUMO, SEGURIDAD ALIMENTARIA Y
    NUTRICIÓN (AECOSAN);Ministerio de Sanidad, Servicios Sociales e Igualdad;6202,26 m2
Edificio;Madrid CALLE HORTALEZA, 104 - 2Û IZD.;INSTITUTO DE LA MUJER Y PARA LA IGUALDAD DE
    OPORTUNIDADES;Ministerio de Sanidad, Servicios Sociales e Igualdad;353 m2
```

Una vez creado este fichero de texto, es fácil hacer consultas sobre él. Por ejemplo, podemos estar interesados en consultar aquellos inmuebles cuyo `Uso` (que es el nombre de una de las columnas) contenga la palabra `"JUVENTUD"`. En este caso obtendríamos una lista con todas las filas que contienen dicha palabra en la columna `Uso`.

En esta práctica se pide:

Ejercicio 1 Define una función `processPages(urls, file_names)` que, dada una lista con las direcciones de distintas páginas web (`urls`) y una lista con nombres de ficheros (`file_names`), cree un fichero por cada página web en `urls`, usando el nombre indicado por el elemento correspondiente de `file_names`. En estos ficheros debe haber, como hemos explicado arriba, una línea para cada fila de la tabla, y cada columna debe estar separada por un `;`. Los ficheros `sanidad` y `presidencia`, disponibles en el Campus, muestran los resultados esperados para la siguiente ejecución:

```
sanidad = 'http://transparencia.gob.es/es_ES/buscar/contenido/cibi/CIBI_DPTO26'
presidencia = 'http://transparencia.gob.es/es_ES/buscar/contenido/cibi/CIBI_DPTO25'

urls = [sanidad, presidencia]
file_names = ['sanidad', 'presidencia']

processPages(urls, file_names)
```

Ten en cuenta lo siguiente:

(a) La primera fila, indicando los nombres de las columnas, se encuentra dentro de las etiquetas `<th>` de `<thead>`. Las etiquetas `<th>` solo se usan en esta parte de la página.

(b) El resto de filas se encuentran dentro de `<tbody>`. Cada fila corresponde a una entrada con la etiqueta `<tr>`, mientras que cada elemento de la fila está delimitado por la etiqueta `<td>`.

(c) No nos interesan las líneas tal y como se presentan en la página, es necesario "limpiarlas". Para ello, es necesario coger solo el texto que se encuentra entre etiquetas `html`.

Ejercicio 2 Define una función `filterPages(file_names, cat, pal)` que devuelva una lista de lista.

Cada una de estas listas debe contener aquellas líneas del fichero que cumplen que en la columna `cat` contienen la cadena "pal".

Por ejemplo, si ejecutamos la función `filterPages(['sanidad', 'presidencia'], "Uso", "JUVENTUD")` obtenemos el siguiente resultado:

```
[['Edificio;Madrid PLAZA COMENDADORAS, 6 / GARAJE, ALMACEN Y LOCAL;INSTITUTO DE LA JUVENTUD;
    Ministerio de Sanidad, Servicios Sociales e Igualdad;398,62 m2\n',
 'Edificio;Madrid P CASTELLANA 27 Y MARQUES RISCAL 16;INSTITUTO DE LA JUVENTUD;Ministerio
    de Sanidad, Servicios Sociales e Igualdad;1323 m2\n',
 'Edificio;Madrid ORTEGA Y GASSET 71;INSTITUTO DE LA JUVENTUD;Ministerio de Sanidad,
    Servicios Sociales e Igualdad;4425 m2\n',
 'Solar;M\xc3\xa1laga PARTIDO DE LA CAPILLA;INSTITUTO DE LA JUVENTUD;Ministerio de Sanidad,
    Servicios Sociales e Igualdad;100000 m2\n'],
 []]
```

Es decir, en el fichero `'sanidad'` hay 3 edificios y un solar cuyo uso es el Instituto de la Juventud, mientras que en el fichero `'presidencia'` no hay ningún edificio al que se le de un uso que contenga la palabra `JUVENTUD` (y por tanto tenemos la lista vacía).

Ejercicio 3 El problema que tenemos en este caso es que es difícil visualizar la información. Para mejorar este aspecto puedes usar las siguientes funciones:

```
def main(urls, file_names, cat, pal):
    processPages(urls, file_names)
    filtered = filterPages(file_names, cat, pal)
    plot(file_names, filtered, cat, pal)

def plot(file_names, filtered, cat, pal):
    numList = []
    for minist in filtered:
        numList.append(len(minist))
    y_pos = np.arange(len(numList))
    plt.barh(y_pos, numList, align='center', alpha=0.4)
    plt.yticks(y_pos, file_names)
    plt.xlabel('Cantidad')
    title = 'Ministerios con la cadena ' + pal + ' en la categoria ' + cat
    plt.title(title)

    plt.show()
```

Estas funciones permiten generar los gráficos que se muestran en las figuras 3, 4 y 5 y que corresponden a las ejecuciones `main(urls, file_names, 'Uso', 'INSTITUTO')`, `main(urls, file_names, 'Tipo', 'Edificio')` y `main(urls, file_names, 'Localizacion', 'Madrid')`, respectivamente.

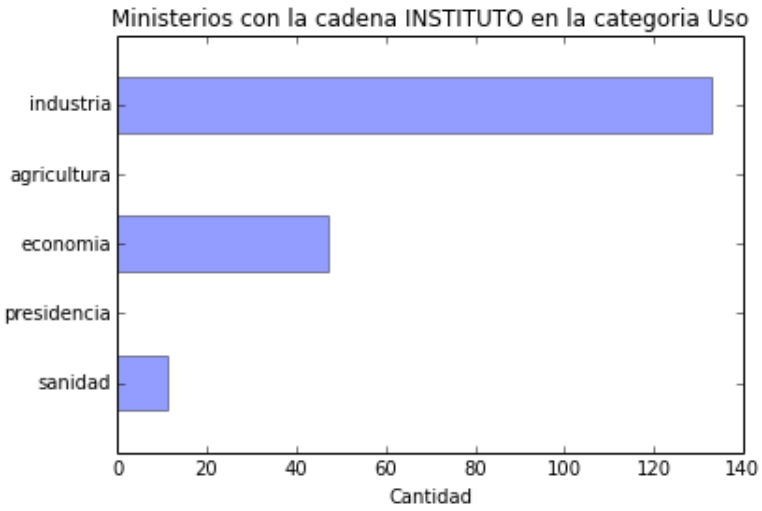

Figura 3: Ejemplo de ejecución I

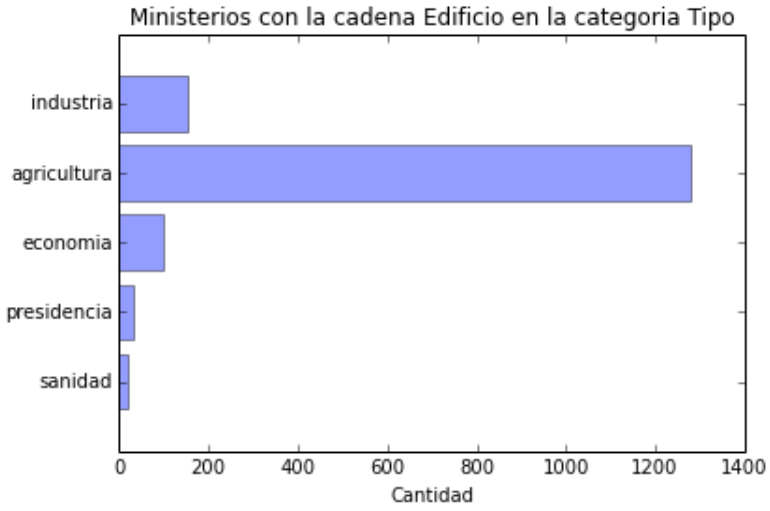

Figura 4: Ejemplo de ejecución II

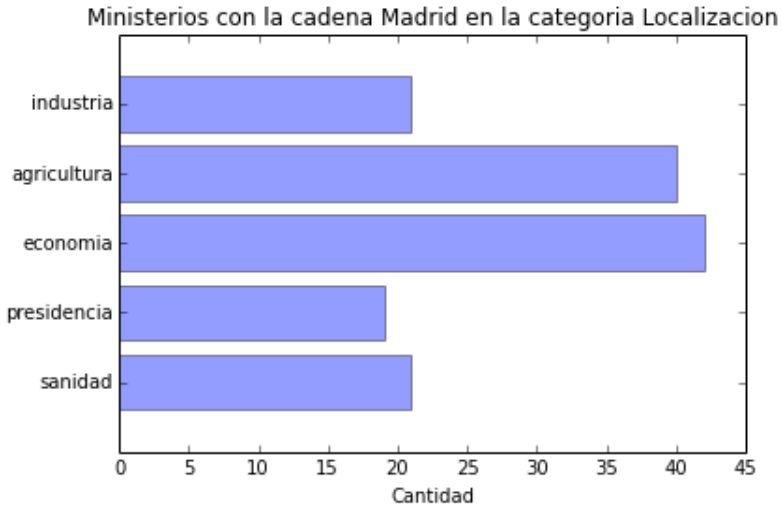

Figura 5: Ejemplo de ejecución III