



Exercise 6:

Web Scrapping

Storage and Data Recovery

Master's Degree in Intelligent Systems

Iván López Muñoz

Steps

01

Choose 5 web pages

Get its URLs to later scrap the TITLE tag and its KEYWORDS

03

Get data from web pages

Access directly to the pages to retrieve data using a php script

02

Create database

Using phpmyadmin create a database to later storage all the data

04

Storage data

With SQL statements in the php script storage all the data retrieved

Choose 5 web pages

The chosen pages URL are:

"https://www.marca.com/"

"https://as.com/"

"https://www.sport.es/es/"

"https://www.mundodeportivo.com/"

"https://www.estadiodeportivo.com/"

Those websites are all related to sports and primarily focus on providing news and information about various sports events, teams, and athletes

Titles and Keywords

Each webpage has a title and the separated words are the keywords

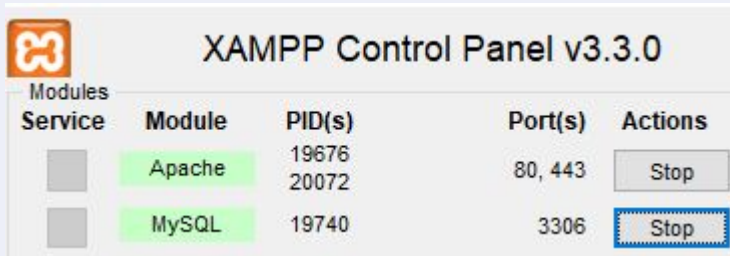


The image shows a vertical list of search results for sports websites. Each result includes a logo, the website name, its URL, and a title. The results are as follows:

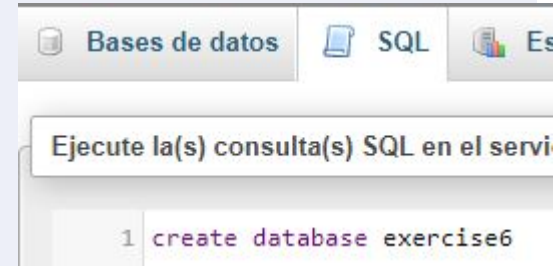
-  **Mundo Deportivo**
<https://www.mundodeportivo.com>
Mundo Deportivo | Noticias deportivas: Fútbol, motor, tenis y ...
-  **Diario AS**
<https://as.com>
AS.com - Diario online deportivo. Fútbol, motor y mucho más
-  **SPORT**
<https://www.sport.es>
SPORT | Noticias del Barça, La Liga, fútbol y otros deportes
-  **Estadio Deportivo**
<https://www.estadiodeportivo.com>
Estadio Deportivo - Diario online de información deportiva
-  **Marca**
<https://www.marca.com>
MARCA - Diario online líder en información deportiva

Create database

1. Turn on 'Apache' and 'MySQL' modules in XAMPP control panel



2. Access <http://localhost/phpmyadmin>
3. Execute the next command and press continue:
4. Select the database created in left panel



Create tables

1. Create different tables to store the data

The database can be like next class diagram:



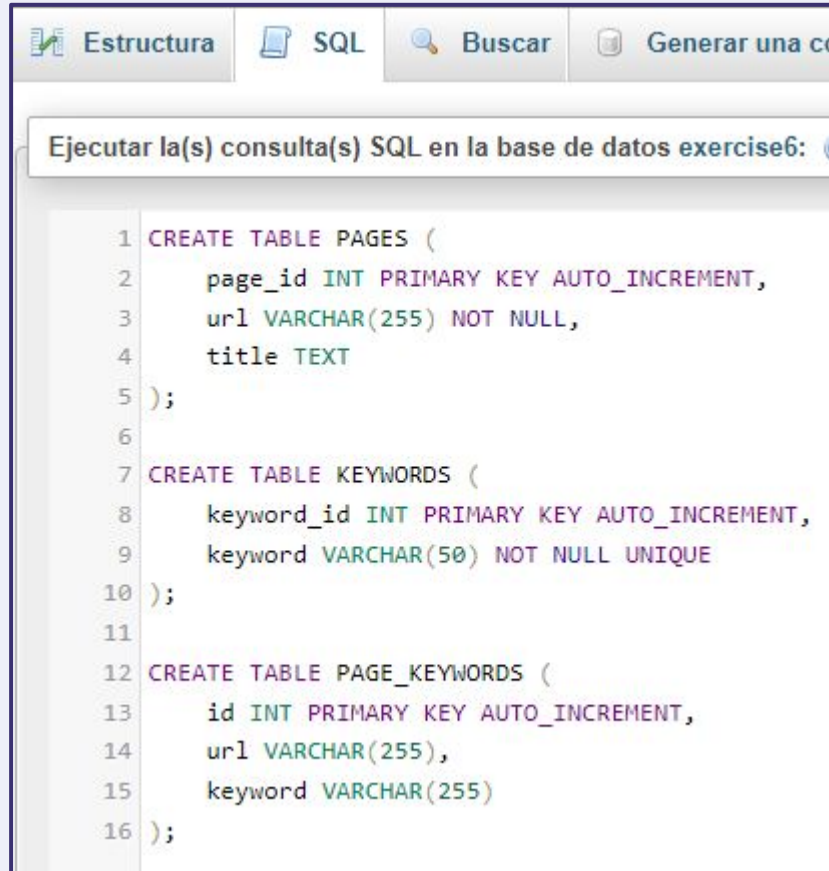
2. As pages is 1 to many and keywords is 0 to many we can add a new table to relate both tables

This way we accomplish both objectives of the exercise:

- Insert the URL in table PAGES
- Split the title in its individual words, insert these words as KEYWORDS in a table of the database and relate them to the page that contains this word into the title tag

Create tables SQL queries

- By setting in table KEYWORDS the keyword UNIQUE, there cannot be repeated words of the titles
- In this way and by writing in the keywords table with INSERT IGNORE we ensure that even if the words are written 1 time in the KEYWORDS table, the words can be related multiple times in the PAGE_KEYWORDS table



The screenshot shows a database management interface with a menu bar containing 'Estructura', 'SQL', 'Buscar', and 'Generar una co'. Below the menu bar is a status bar that reads 'Ejecutar la(s) consulta(s) SQL en la base de datos exercise6:'. The main area displays three SQL queries for creating tables:

```
1 CREATE TABLE PAGES (  
2     page_id INT PRIMARY KEY AUTO_INCREMENT,  
3     url VARCHAR(255) NOT NULL,  
4     title TEXT  
5 );  
6  
7 CREATE TABLE KEYWORDS (  
8     keyword_id INT PRIMARY KEY AUTO_INCREMENT,  
9     keyword VARCHAR(50) NOT NULL UNIQUE  
10 );  
11  
12 CREATE TABLE PAGE_KEYWORDS (  
13     id INT PRIMARY KEY AUTO_INCREMENT,  
14     url VARCHAR(255),  
15     keyword VARCHAR(255)  
16 );
```

Get data from web pages

Function curl:

```
<?php
// FUNCTION curl
function curl($url)
{
    $ch = curl_init($url);
    curl_setopt($ch, CURLOPT_RETURNTRANSFER, TRUE);
    curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, false);
    $info = curl_exec($ch);
    curl_close($ch);
    return $info;
}
// Function to connect to the database
```

This function uses cURL, a library for transferring data with URLs. It retrieves the content from a given URL

- \$info - Returns the fetched content

Storage data in database from web pages

Once we have the curl function with all the fetched content, we have to establish a connection to our MySQL database to be able to store all data:

```
// Function to connect to the database
function connectToDatabase()
{
    $servername = "localhost";
    $username = "root";
    $password = "";
    $dbname = "exercise6";

    $con = mysqli_connect($servername, $username, $password, $dbname);
    if (!$con) {
        die("Connection failed: " . mysqli_connect_error());
    }
    return $con;
}
```

- Be careful to put the correct variable strings otherwise it won't work

Storage data in database from web pages

Function to insert URL and title into PAGES table and split title into keywords and relate them to the page in PAGE_KEYWORDS table:

```
function processWebpage($url, $con)
{
    // Fetch webpage content
    $content = curl($url);

    // Extract title from the webpage
    if (strpos($url, 'marca.com') !== false) {
        preg_match('/<title>(.*)</title>/is', $content, $matches);
    } else {
        preg_match('/<title>(.*)</title>/is', $content, $matches);
    }
    $title = isset($matches[1]) ? $matches[1] : '';

    // Insert URL and title into PAGES table
    $sql = "INSERT INTO PAGES (url, title) VALUES ('$url', '$title')";
    mysqli_query($con, $sql);
    $pageId = mysqli_insert_id($con);

    // Split title into individual words (keywords)
    $keywords = preg_split("/\s+/", $title);

    // Insert keywords into KEYWORDS table and relate them to the page in PAGE_KEYWORDS table
    foreach ($keywords as $keyword) {
        $keyword = mysqli_real_escape_string($con, $keyword);
        $sql = "INSERT IGNORE INTO KEYWORDS (keyword) VALUES ('$keyword')";
        mysqli_query($con, $sql);
        $keywordId = mysqli_insert_id($con);

        // Insert URL and keyword into PAGE_KEYWORDS table
        $sql = "INSERT INTO PAGE_KEYWORDS (url, keyword) VALUES ('$url', '$keyword')";
        mysqli_query($con, $sql);
    }
}
```

Storage data in database from web pages

- Fetches the content of the webpage using the curl() function.
- Extracts the title of the webpage using regular expressions (preg_match). The specific pattern for extracting the title varies depending on whether the URL contains 'marca.com' or not. 'marca.com' uses a different format for the title data
- Inserts the URL and title into a table named PAGES.
- Splits the title into individual words (keywords) using preg_split and inserts them in table KEYWORDS ignoring duplicates.
- Relates the keywords to the page in the PAGE_KEYWORDS table. It also establishes a relationship between the URL and the keyword.

Full PHP code

```
<?php
// FUNCTION curl
function curl($url)
{
    $ch = curl_init($url);
    curl_setopt($ch, CURLOPT_RETURNTRANSFER, TRUE);
    curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, false);
    $info = curl_exec($ch);
    curl_close($ch);
    return $info;
}

// Function to connect to the database
function connectToDatabase()
{
    $servername = "localhost";
    $username = "root";
    $password = "";
    $dbname = "exercise6";

    $con = mysqli_connect($servername, $username, $password, $dbname);
    if (!$con) {
        die("Connection failed: " . mysqli_connect_error());
    }
    return $con;
}

// MAIN PROGRAM
$url = [
    "https://www.marca.com/",
    "https://as.com/",
    "https://www.sport.es/es/",
    "https://www.mundodeportivo.com/",
    "https://www.estadiodeportivo.com/"
];

// Connect to the database
$con = connectToDatabase();
// Process each URL
foreach ($url as $url) {
    processWebpage($url, $con);
}
mysqli_close($con); // Close the database connection
?>
```

1

2

4

```
// Function to insert URL and title into PAGES table and split title into keywords
function processWebpage($url, $con)
{
    // Fetch webpage content
    $content = curl($url);

    // Extract title from the webpage
    if (strpos($url, 'marca.com') !== false) {
        preg_match('/title>(.*?)</title>/is', $content, $matches);
    } else {
        preg_match('/<title>(.*?)</title>/is', $content, $matches);
    }
    $title = isset($matches[1]) ? $matches[1] : '';

    // Insert URL and title into PAGES table
    $sql = "INSERT INTO PAGES (url, title) VALUES ('$url', '$title')";
    mysqli_query($con, $sql);
    $pageId = mysqli_insert_id($con);

    // Split title into individual words (keywords)
    $keywords = preg_split("/\s+/", $title);

    // Insert keywords into KEYWORDS table and relate them to the page in PAGE_KEYWORDS table
    foreach ($keywords as $keyword) {
        $keyword = mysqli_real_escape_string($con, $keyword);
        $sql = "INSERT IGNORE INTO KEYWORDS (keyword) VALUES ('$keyword')";
        mysqli_query($con, $sql);
        $keywordId = mysqli_insert_id($con);

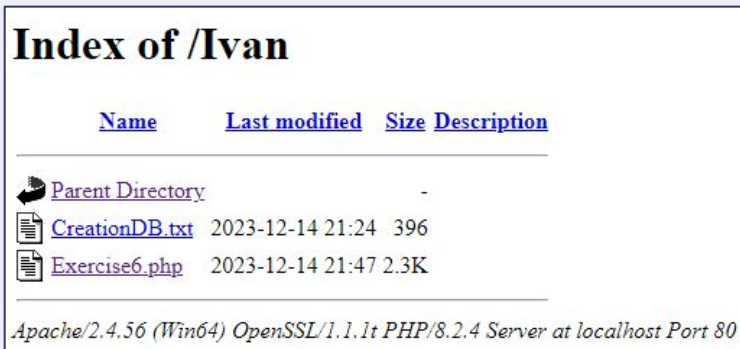
        // Insert URL and keyword into PAGE_KEYWORDS table
        $sql = "INSERT INTO PAGE_KEYWORDS (url, keyword) VALUES ('$url', '$keyword')";
        mysqli_query($con, $sql);
    }
}
```




3

Execute php file

1. Go to `C:\...\xampp\htdocs` where xampp is installed and create a new **folder**
2. Save the PHP file in the created folder `C:\...\xampp\htdocs\folder`
3. Access internet to: `http://localhost/folder/`

In my case **folder = Ivan**



<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 CreationDB.txt	2023-12-14 21:24	396	
 Exercise6.php	2023-12-14 21:47	2.3K	

Apache/2.4.56 (Win64) OpenSSL/1.1.1t PHP/8.2.4 Server at localhost Port 80

4. Click the PHP file with the code ('Exercise6.php') and wait
5. Go to `http://localhost/phpmyadmin` again and check the database tables created before:

Results

Tabla	Acción	Filas
<input type="checkbox"/> keywords	★ Examinar Estructura Buscar Insertar Vaciar Eliminar	33
<input type="checkbox"/> pages	★ Examinar Estructura Buscar Insertar Vaciar Eliminar	5
<input type="checkbox"/> page_keywords	★ Examinar Estructura Buscar Insertar Vaciar Eliminar	47
3 tablas	Número de filas	85

- 5 pages
- 33 keywords
- 47 related keywords to pages

In total there are 47 keywords and 14 of them are duplicated

Results

Table PAGES

page_id	url	title
1	https://www.marca.com/	MARCA - Diario online l?der en informaci?n deporti...
2	https://as.com/	AS.com - Diario online deportivo. Fútbol, motor y ...
3	https://www.sport.es/es/	SPORT Noticias del Barça, La Liga, fútbol y otro...
4	https://www.mundodeportivo.com/	Mundo Deportivo Noticias deportivas: Fútbol, mot...
5	https://www.estadiodeportivo.com/	Estadio Deportivo - Diario online de información d...

keyword_id	keyword
1	MARCA
2	-
3	Diario
4	online
5	l?der
6	en
7	informaci?n
8	deportiva
9	AS.com
13	deportivo.
14	Fútbol,
15	motor
16	y
17	mucho
18	más
19	SPORT
20	
21	Noticias
22	del
23	Barça,
24	La
25	Liga,
26	fútbol
28	otros
29	deportes
30	Mundo
31	Deportivo
34	deportivas:
36	motor,
37	tenis
40	Estadio
45	de
46	información

Table
KEYWORDS

Results

Table PAGE_KEYWORDS

id	url	keyword
1	https://www.marca.com/	MARCA
2	https://www.marca.com/	-
3	https://www.marca.com/	Diario
4	https://www.marca.com/	online
5	https://www.marca.com/	l?der
6	https://www.marca.com/	en
7	https://www.marca.com/	informaci?n
8	https://www.marca.com/	deportiva
9	https://as.com/	AS.com
10	https://as.com/	-
11	https://as.com/	Diario
12	https://as.com/	online
13	https://as.com/	deportivo.
14	https://as.com/	Fútbol,
15	https://as.com/	motor
16	https://as.com/	y
17	https://as.com/	mucho
18	https://as.com/	más
19	https://www.sport.es/es/	SPORT
20	https://www.sport.es/es/	
21	https://www.sport.es/es/	Noticias
22	https://www.sport.es/es/	del
23	https://www.sport.es/es/	Barça,
24	https://www.sport.es/es/	La
25	https://www.sport.es/es/	Liga,
26	https://www.sport.es/es/	fútbol
27	https://www.sport.es/es/	y
28	https://www.sport.es/es/	otros
29	https://www.sport.es/es/	deportes
30	https://www.mundodeportivo.com/	Mundo
31	https://www.mundodeportivo.com/	Deportivo
32	https://www.mundodeportivo.com/	
33	https://www.mundodeportivo.com/	Noticias
34	https://www.mundodeportivo.com/	deportivas:
35	https://www.mundodeportivo.com/	Fútbol,
36	https://www.mundodeportivo.com/	motor,
37	https://www.mundodeportivo.com/	tenis
38	https://www.mundodeportivo.com/	y
39	https://www.mundodeportivo.com/	más
40	https://www.estadiodeportivo.com/	Estadio
41	https://www.estadiodeportivo.com/	Deportivo
42	https://www.estadiodeportivo.com/	-
43	https://www.estadiodeportivo.com/	Diario
44	https://www.estadiodeportivo.com/	online
45	https://www.estadiodeportivo.com/	de
46	https://www.estadiodeportivo.com/	información
47	https://www.estadiodeportivo.com/	deportiva