Instituto Tecnológico de Costa Rica

Unidad de Computación

Introducción al desarrollo de aplicaciones Web

**“Web Scrapping Investigación 2”**

Responsable:

Paola Fabiana Chaves Alvarado

Sede San Carlos

03 de abril del 2017

# **RESUMEN:**

En el presente documento se presenta una, de las tantas formas, de aplicar la técnica para la obtención de información de sitios web por medio del HTML, conocida como web scrapping. Se presenta la forma en la que se desarrolló el web scrapping, el código y también la página de donde se obtuvo la información, así como los datos extraídos.

**TABLA DE CONTENIDO**

[RESUMEN: 2](#_Toc479017385)

[INTRODUCCIÓN 4](#_Toc479017386)

[DESARROLLO 5](#_Toc479017387)

[LINK DEL GIT: 6](#_Toc479017388)

[CONCLUSIONES: 7](#_Toc479017389)

[RECOMENDACIONES 8](#_Toc479017390)

[REFERENCIAS: 9](#_Toc479017391)

# **INTRODUCCIÓN**

Web Scrapping es una necesidad, la obtención de información de diversos sitios para su visualización desde uno solo es necesario. Le facilita al usuario la comprensión de información que se encuentra desplegada en miles de páginas. Imaginar a una persona buscando un hotel y que como resultado tenga miles de páginas separadas como resultado, cuando una página que aplique el web scrapping puede tener toda esa información en un solo lugar es algo completamente diferente. Definitivamente copiar y pegar no sería un buen resultado en este caso, web scrapping es la solución.

Existen varias técnicas para aplicar el web scrapping, y hay muchas muy sencillas de utilizar. Todo desarrollador web debe saber manejar al menos una técnica para realizar el web scrapping.

# **DESARROLLO**

Para el ejemplo realizado se escogió la página web del juego League of Legends conocido como LOL. En la página se encuentra la lista de todos los campeones que forman parte del juego.

Aquí se encuentra el link de la página:

[**http://gameinfo.lan.leagueoflegends.com/es/game-info/champions/**](http://gameinfo.lan.leagueoflegends.com/es/game-info/champions/)

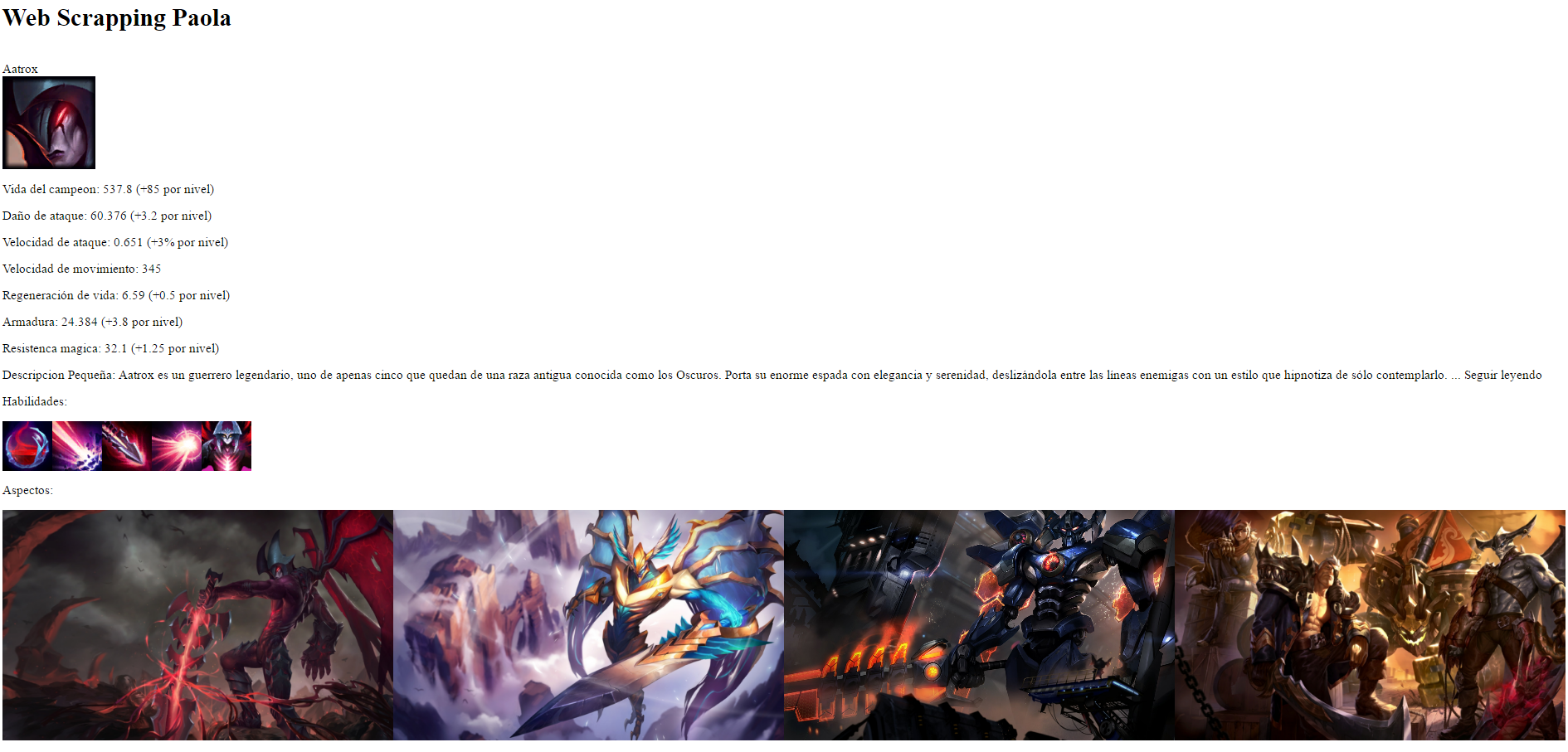
Además, para cada campeón se ingresó a una extensión de la página que especificaba más detalles. Entonces el link del sitio de cada campeón sería de la siguiente forma:

[**http://gameinfo.lan.leagueoflegends.com/es/game-info/champions/nombreDelCampeon**](http://gameinfo.lan.leagueoflegends.com/es/game-info/champions/nombreDelCampeon)

Los atributos tomados del sitio fueron:

* Nombre del campeón: Lo toma de la página principal y con este atributo accede a los demás del campeón.
* Imagen del campeón: Estaba contenida en un div y dentro de este un img, se toma el src del img para poder acceder a la imagen.
* Vida del campeón: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 0.
* Daño de ataque: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 1.
* Velocidad de ataque: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 2.
* Velocidad de movimiento: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 3.
* Regeneración de vida: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 4.
* Armadura: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 5.
* Resistencia mágica: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 6.
* Descripción: Contenido dentro de un span, había varios con la misma identificación por lo que se tomó según el orden. Es decir, para acceder a este atributo se busca los span que están en el índice 7.
* Habilidades: Estas están representadas con imágenes.
* Aspectos: De igual forma representadas con imágenes

Para que se muestre la información es necesario abrir el archivo index.php en XAMPP o WAMP, así se debería de ver el resultado final:



# **LINK DEL GIT:**

# **CONCLUSIONES:**

A forma de aclaración, en la primera investigación se planteó la metodología utilizando Python por medio de beautifulSoup, debido a algunos inconvenientes en la instalación de la librería se cambió esto a una solución en PHP. PHP facilita el trabajo el web scrapping, se requieren conocimientos básicos para poder realizar esta técnica de obtención de información.

Se ven muchos sitios web en los que el html es algo desorganizado y puede llegar a complicarse la obtención de información, eso complica un poco realizar el web scrapping de manera organizada pero sin duda es la mejor opción cuando se piensa en facilitar la visualización de información para los usuarios.

# RECOMENDACIONES

* Realizar la solución en PHP fue algo sencillo, además se encuentra bastante documentación para llegar a realizar varios ejemplos fácilmente, pero puede que hayan otras formas de realizar el mismo ejemplo y de una mejor manera, incluso más organizada. Como lo pudo haber sido la opción en Python o talvez rvest

# **REFERENCIAS:**

Reyes, J. J. (s.f.). *DevCode*. Obtenido de Hacer web scraping con PHP: https://devcode.la/tutoriales/hacer-web-scraping-con-php/