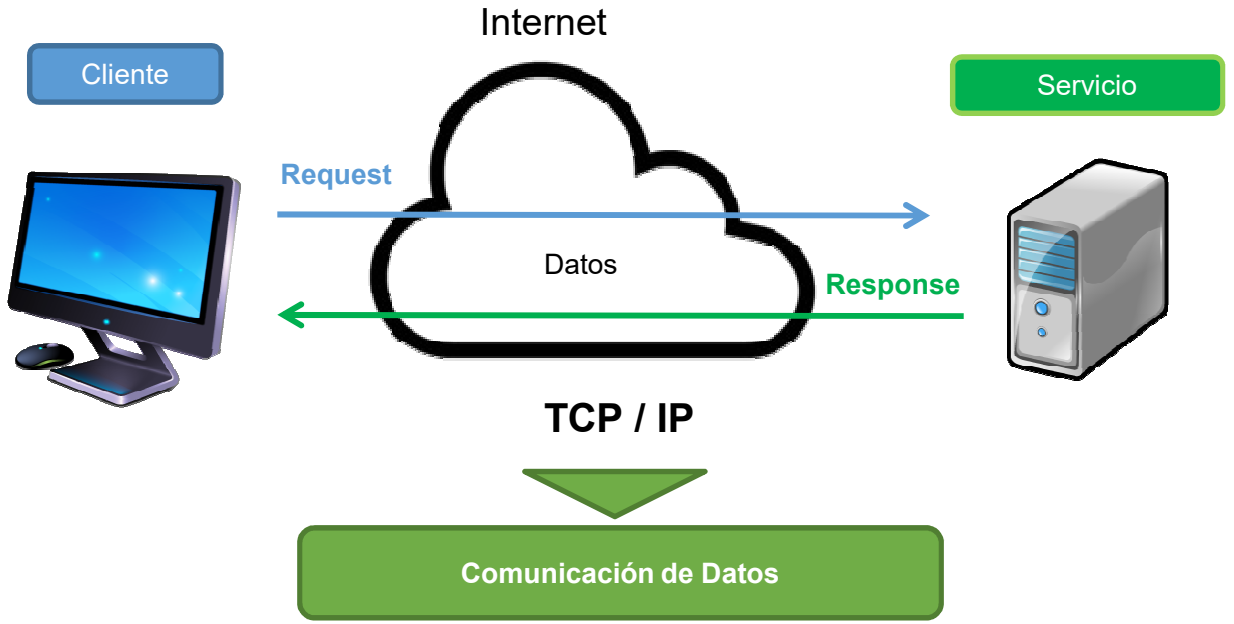
4.
Datos en la Nube #1

-Extracción de datos en la nube
-Peticiones HTTP
-Utilizando repositorios externos
-Trabajando con HTML
-Web Scraping
-Trabajando con Beautiful Soup
-Parseando archivos XML

¿Cómo funciona un portal Web?

Web Tech

Arquitectura Cliente/Servicio



Cómo funciona un portal Web?

El cliente

Cliente

Navegador/Browser

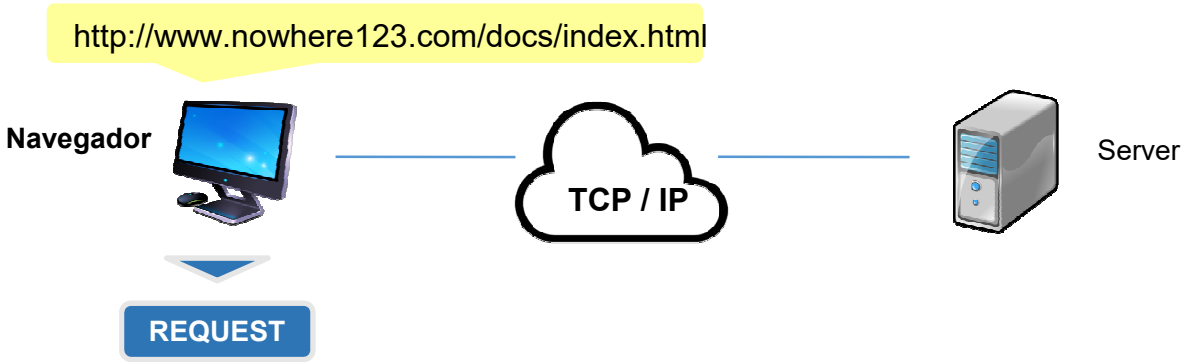
Los Navegadores son Aplicaciones de Software Cliente que se especializan en 2 cosas:

1) la construcción y direccionamiento de peticiones web, y

2) la posterior representación gráfica (**renderización**) de las respuestas enviadas por un Software de Servicio

Cómo funciona un portal Web?

El cliente



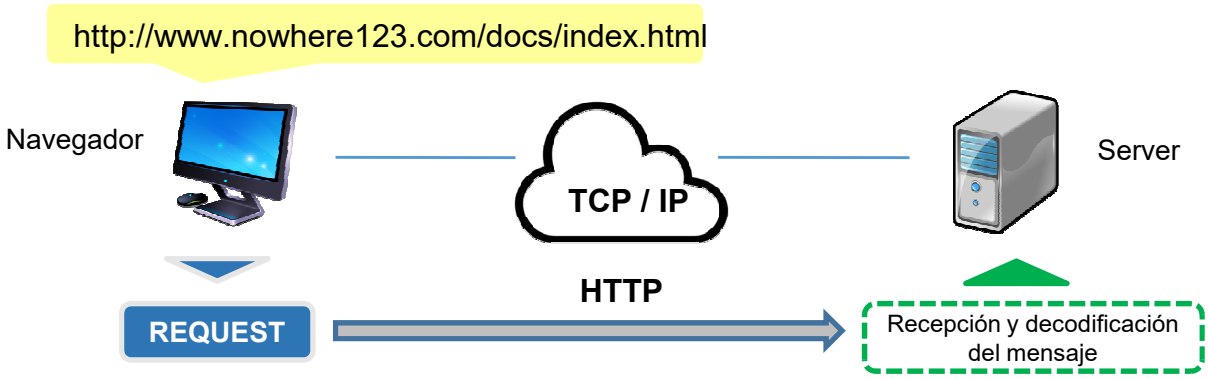
1) Localización del recurso en la web

URL **protocol**://**hostname**:**port**/**path-and-file-name**

- protocol: **http**
- hostname: **www.nowhere123.com** (DNS) ó 181.168.2.102 (IP)
- port: **80**
- path-and-file-name: **/docs/index.html**

Cómo funciona un portal Web?

El cliente

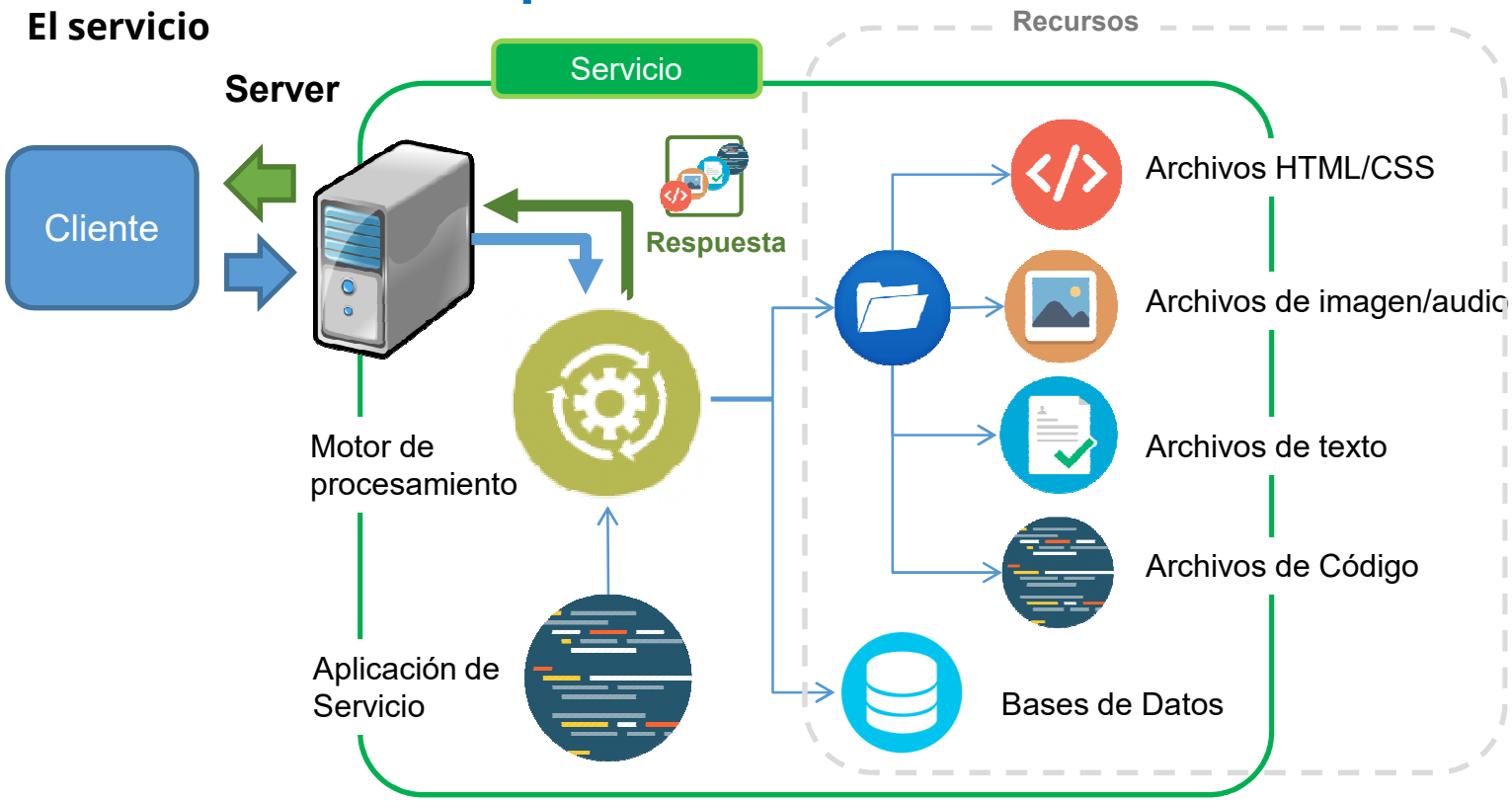


2) Confección y envío del Request en protocolo HTTP

```
GET /docs/index.html HTTP/1.1
Host: www.nowhere123.com
Accept: image/gif, image/jpeg, */*
Accept-Language: en-us Accept-Encoding: gzip, deflate
User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
(blank line)
```

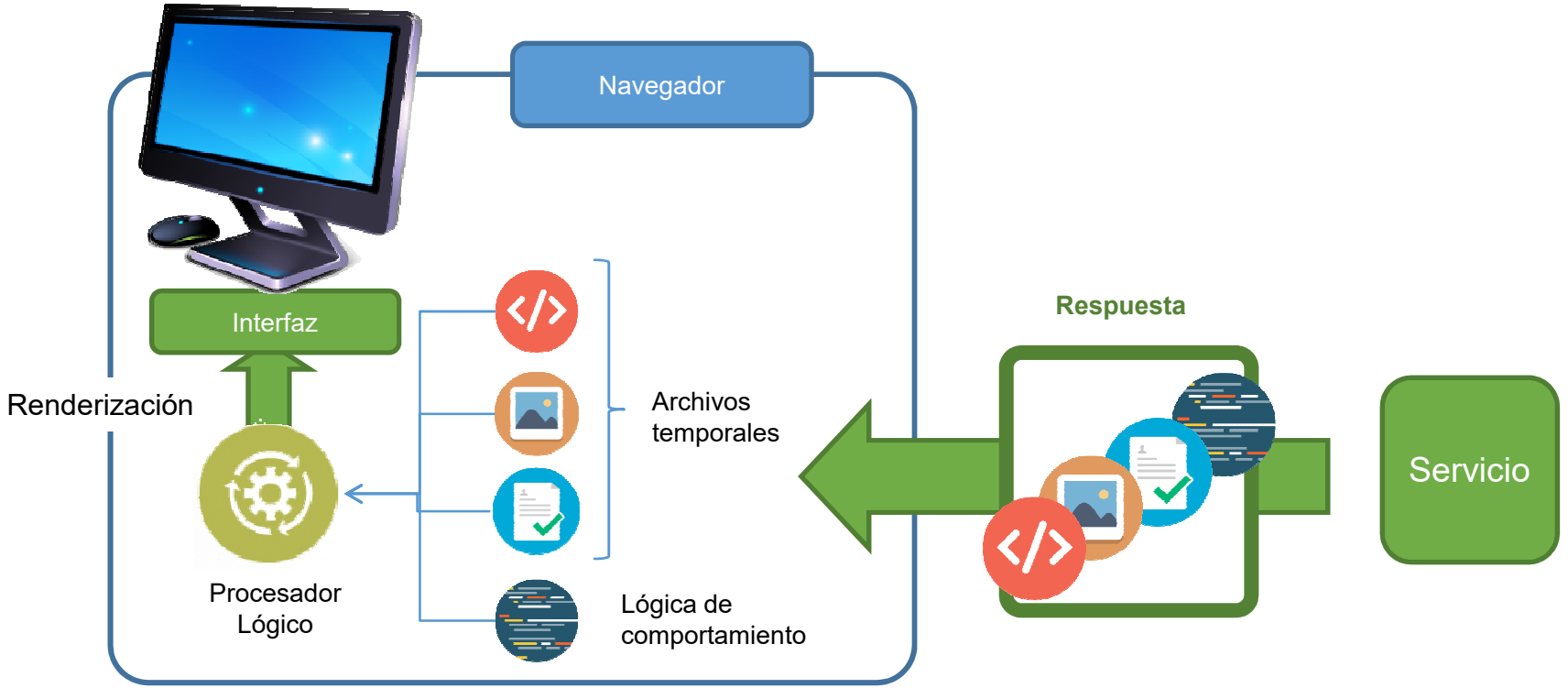
Cómo funciona un portal Web?

El servicio



Cómo funciona un portal Web?

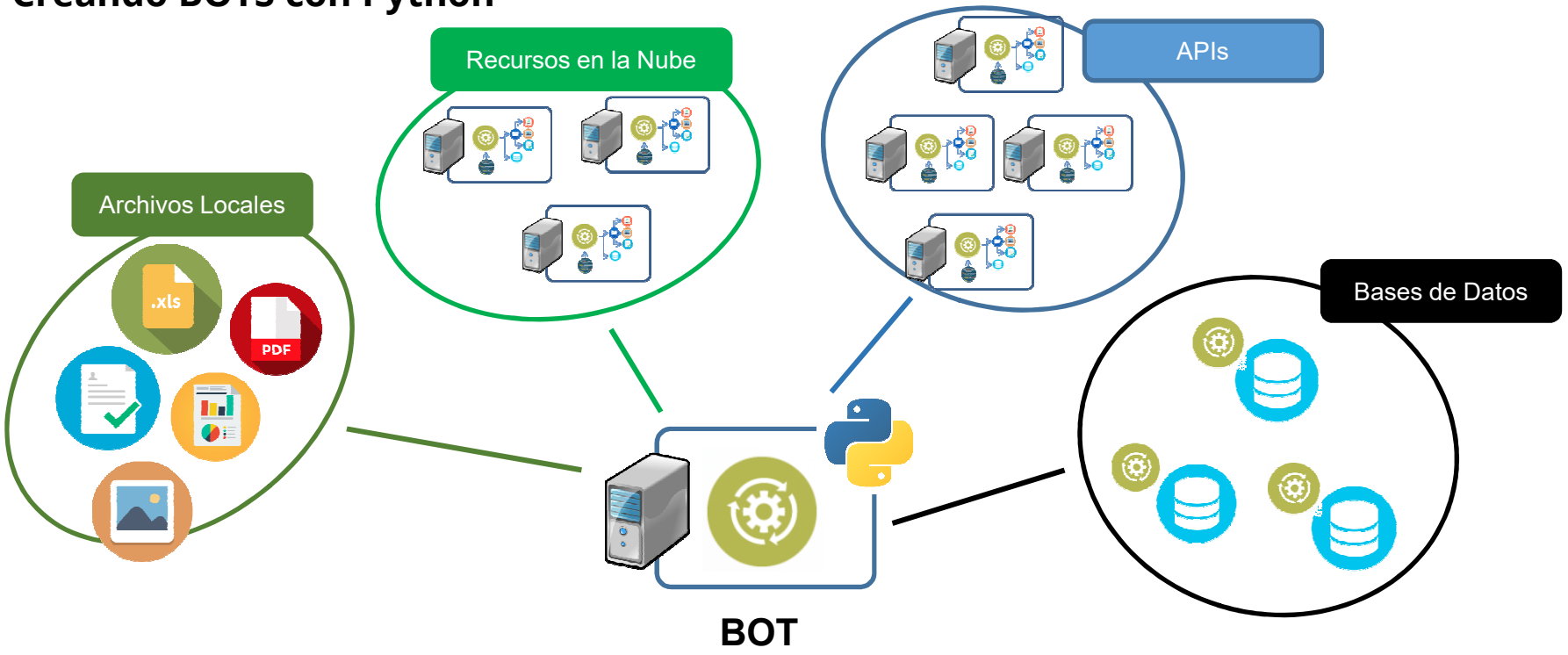
El servicio



Creando Clientes con Python

Integración de Fuentes

Creando BOTS con Python



Práctica
Archivos en la Nube

-Creando peticiones HTTP desde Python con Librería Requests
-Bajando archivos de manera automática
-Procesando CSV de manera directa
-Descarga de repositorios Googlesheets

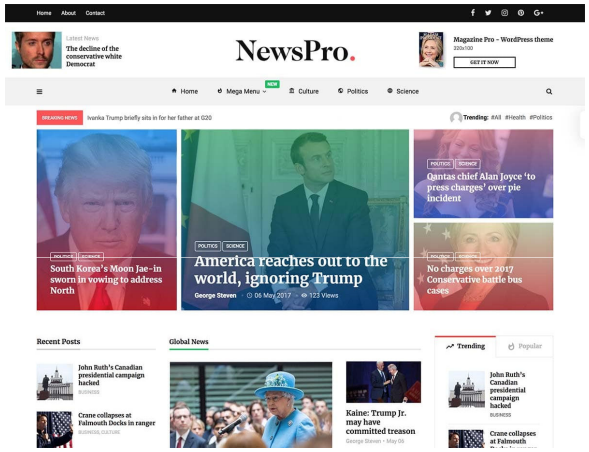
Web Scraping

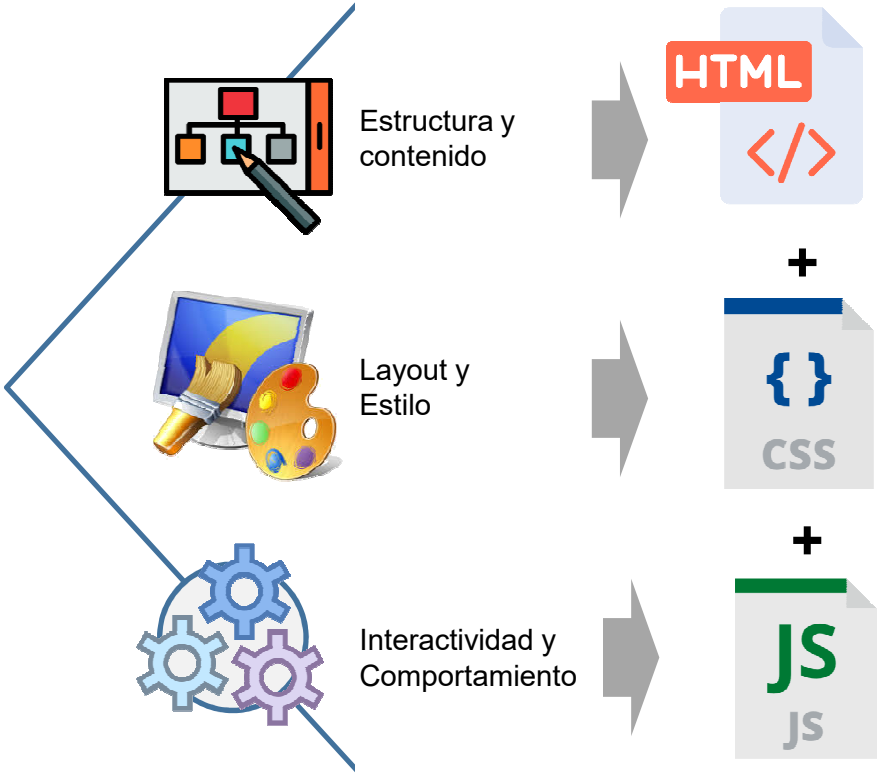
Descargando la Web

Web Scraping

Estructura de una página Web

Página Web





Web Scraping

Estructuras de contenido con HTML

HTML es un lenguaje de **“Marcado por etiquetas”** en donde cada elemento se identifica mediante una etiqueta que “abre” y otra etiqueta que “cierra”

<elemento> → Abre / Inicia el elemento

…..

…..

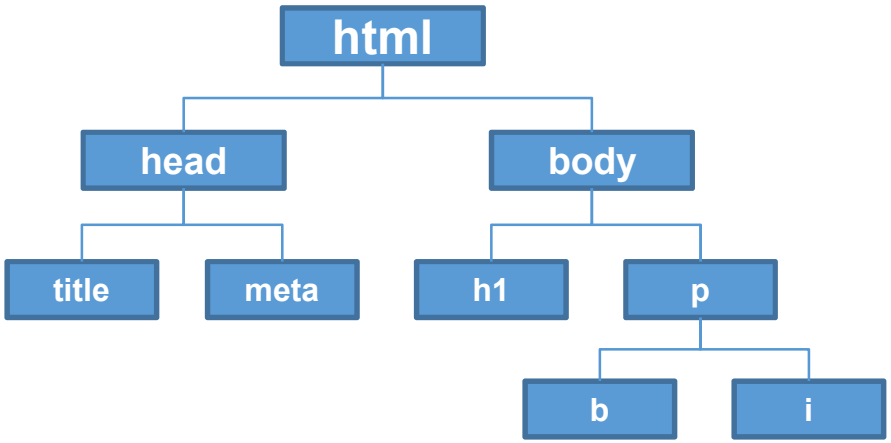
</elemento> → Cierra / Fin del elemento

```
<!DOCTYPE html>
<html>
    <head>
        <title>Recetas!</title>
        <meta charset="utf-8">
    </head>
    <body>
        <h1>Mis recetas de cocina</h1>
        <p>Hace tiempo que quería escribir sobre las
        recetas que he logrado conocer a través de
        los años.<b>Son fantásticas!</b> y quiero
        compartirlas con aquellos que también se
        entusiasman con <i>un buen plato</i>.</p>
    </body>
</html>
```

Web Scraping

Estructuras de contenido con HTML

DOM – Document Object Model




```
<!DOCTYPE html>
<html>
    <head>
        <title>Recetas!</title>
        <meta charset="utf-8">
    </head>
    <body>
        <h1>Mis recetas de cocina</h1>
        <p>Hace tiempo que quería escribir sobre las
        recetas que he logrado conocer a través de
        los años.<b>Son fantásticas!</b> y quiero
        compartirlas con aquellos que también se
        entusiasman con <i>un buen plato</i>.</p>
    </body>
</html>
```