

**UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR**



TRABAJO FIN DE GRADO


DESARROLLO DE UN SISTEMA DE RECOGIDA Y CATEGORIZACIÓN DE DATOS DE RECETAS

GONZALO GALLASTEGUI PEÑALBA
Tutor: ALEJANDRO BELLOGÍN KOUKI
Ponente: PABLO CASTELLS AZPILICUETA

JUNIO 2015



Tabla de contenidos

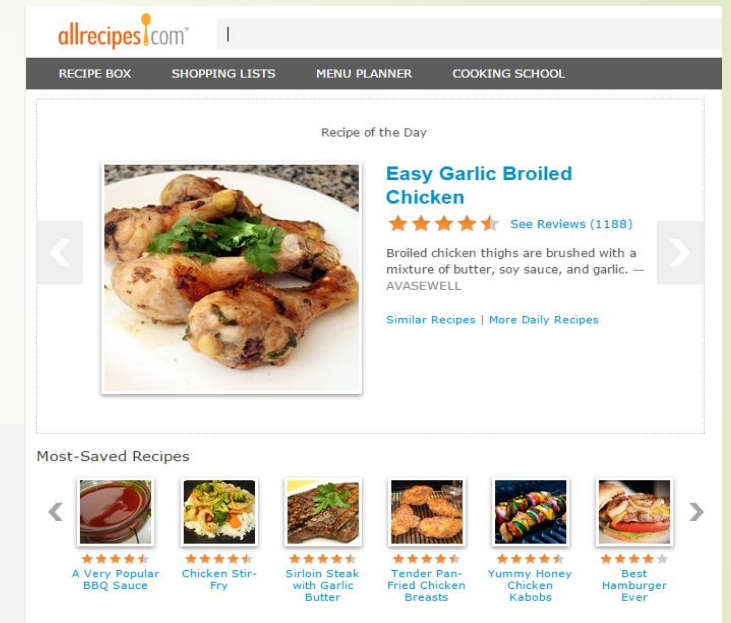
- Introducción
 - Objetivos
 - Análisis de requisitos
 - Diseño
 - Desarrollo
 - Pruebas y resultados
 - Conclusiones y Trabajo futuro
- 

Introducción

Páginas web de recetas



The screenshot shows the homepage of Simply Recipes. At the top left is the logo, a stylized flower with the text "Simply Recipes". To the right is a search bar with the placeholder text "Search Simply Recipes...". Below the search bar are navigation links: "Home", "About", "Index", and "Subscribe". On the left side, there is a "Featured" menu with categories like "Chicken", "Beef", "Fish and Seafood", "Pasta", "Gluten-Free", "Vegetarian", "Quick", "Budget", "How To", "Paleo", and "Low Carb". The main content area features a large image of a pitcher of lemonade and a glass. To the right of the image is a "Welcome to Simply Recipes!" message from a woman named "Elise", with social media icons for Facebook, Pinterest, Twitter, and Instagram, and a "Follow me on Pinterest" link. At the bottom of the welcome message is the "BlogHer" logo.



The screenshot shows the homepage of allrecipes.com. At the top is the logo "allrecipes.com" and a search bar. Below the logo are navigation links: "RECIPE BOX", "SHOPPING LISTS", "MENU PLANNER", and "COOKING SCHOOL". The main content area features a "Recipe of the Day" section with a large image of "Easy Garlic Broiled Chicken" and a description: "Broiled chicken thighs are brushed with a mixture of butter, soy sauce, and garlic. — AVASEWELL". Below the image are "Similar Recipes" and "More Daily Recipes" links. At the bottom, there is a "Most-Saved Recipes" section with a row of recipe thumbnails: "A Very Popular BBQ Sauce", "Chicken Stir-Fry", "Sirloin Steak with Garlic Butter", "Tender Pan-Fried Chicken Breasts", "Yummy Honey Chicken Kabobs", and "Best Hamburger Ever".



wikicocina.com
recetas para cocineros y cocinillas

RECETAS +

GASTRONOMÍA +

ESCUELA DE COCINA +

NOSOTROS +

ENVÍANOS TUS RECETAS

RECETAS

Aquí encontrarás todas las recetas que hemos probado, y que estaban tan buenas que queremos que las cocines tu también.

Introducción

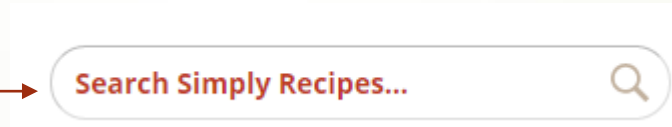
Páginas web de recetas

- Problema: Buscadores de contenido de búsquedas por literal.

➤ Wikicocina →



➤ Simplyrecipes →



➤ Allrecipes →



Introducción

Páginas web de recetas

➤ Ejemplo de consulta:

vegetarian pizza with tomatoes and mozzarella and without onion



About 836 results



How to Slice an Onion

Step by step instructions on how to safely slice onions both lengthwise, and crosswise.



How to Chop an Onion

How to chop an onion, safely, easily, and with minimum tears! Video plus step-by-step photos.



How to Caramelize Onions

How to slowly caramelize onions to bring out deep, rich, sweet flavor as the natural sugars in the onions caramelize. Video included.



How to Grill Pizza

Step-by-step instructions for using your grill, gas or charcoal, to make homemade pizza.



Introducción

Propuesta

Desarrollar un sistema avanzado de búsqueda
que mejore la búsqueda por literal.

Introducción

Allrecipes

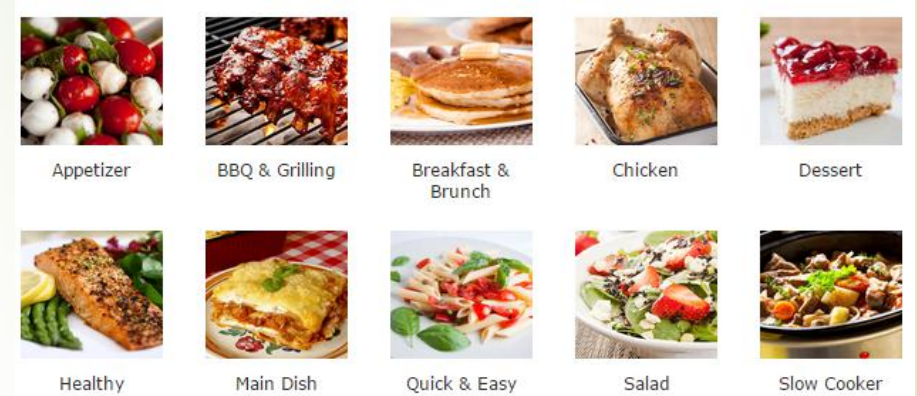
Recipes 56,324

Ofrece una gran cantidad de información.



★★★★★ (797)
Chocolate Banana Bread

El sitio web es muy vivo y con gran número de valoraciones de usuarios.



Las categorías de las recetas establecen el patrón de distribución de la información.

Nutrition

Calories	448 kcal	■ ■ ■ ■ ■	22%	Carbohydrates	36.5 g	■ ■ ■ ■ ■	12%
Cholesterol	82 mg	■ ■ ■ ■ ■	27%	Fat	21.3 g	■ ■ ■ ■ ■	33%
Fiber	4 g	■ ■ ■ ■ ■	16%	Protein	29.7 g	■ ■ ■ ■ ■	59%
Sodium	1788 mg	■ ■ ■ ■ ■	72%				

* Percent Daily Values are based on a 2,000 calorie diet.

La información que ofrece es muy variada y completa.

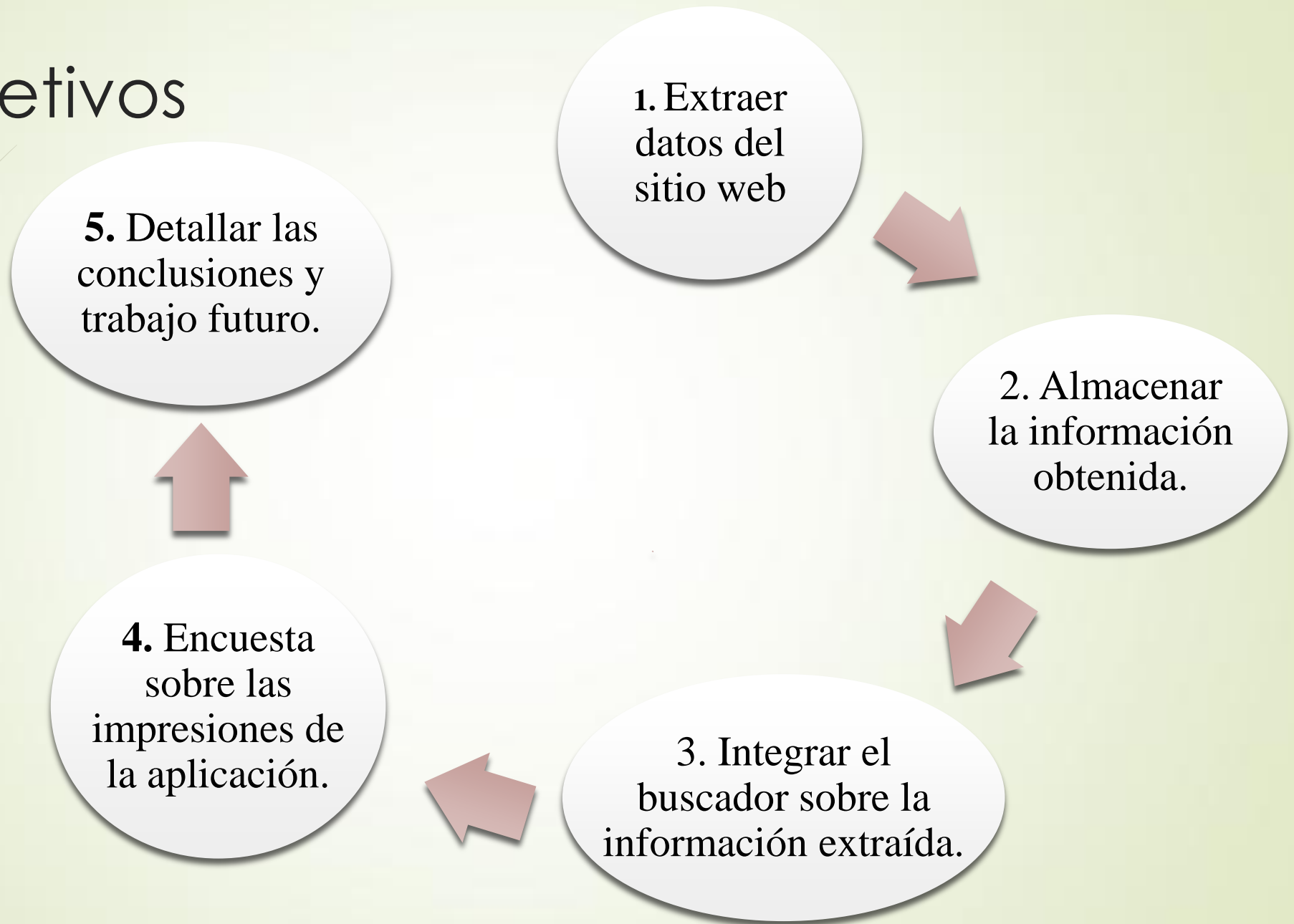


Tabla de contenidos

- Introducción
- Objetivos
- Análisis de requisitos
- Diseño
- Desarrollo
- Pruebas y resultados
- Conclusiones y Trabajo futuro



Objetivos



Esquema

Sistema general

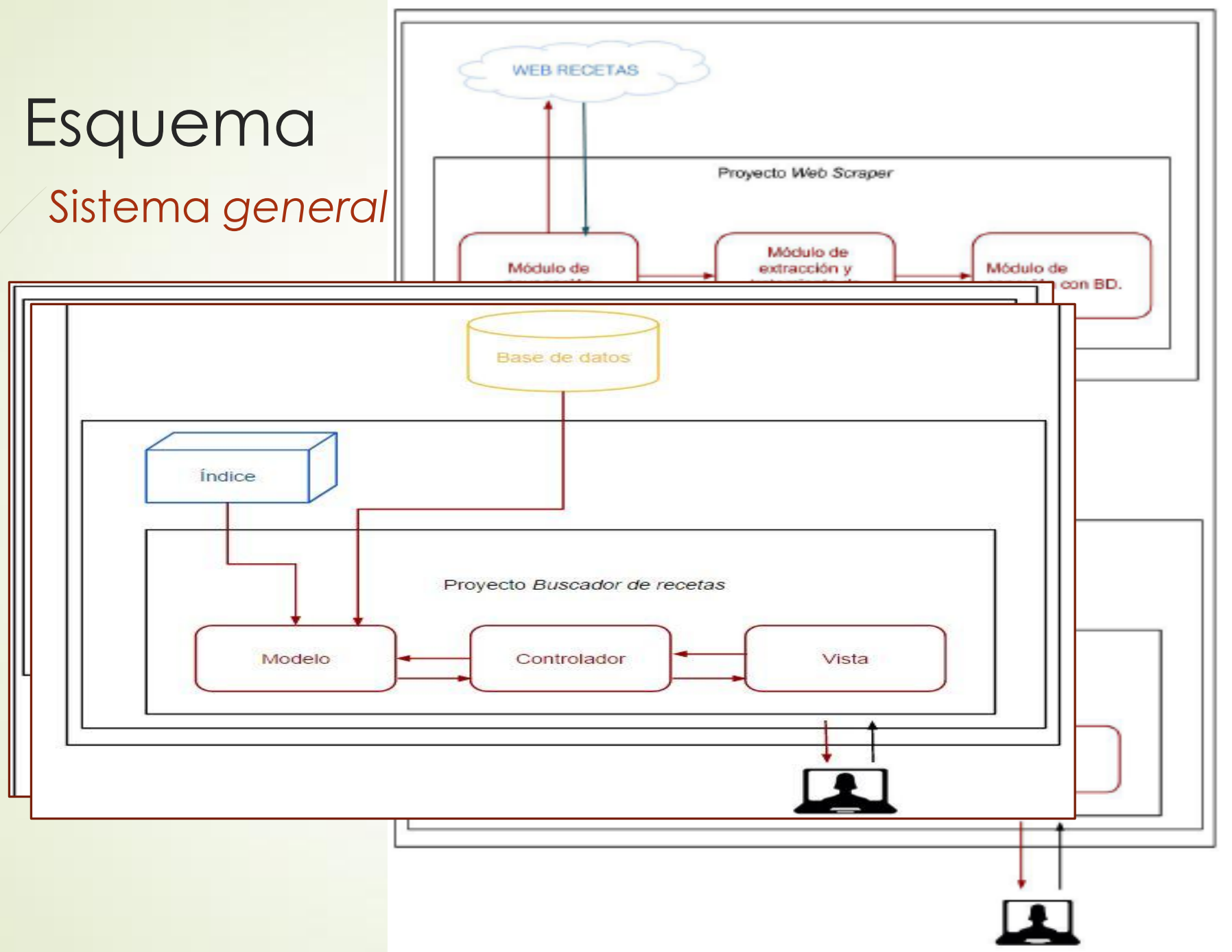


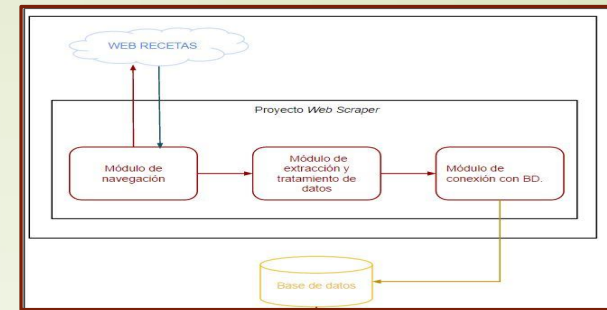


Tabla de contenidos

- Introducción
- Objetivos
- Análisis de requisitos
- Diseño
- Desarrollo
- Pruebas y resultados
- Conclusiones y Trabajo futuro

Análisis de requisitos

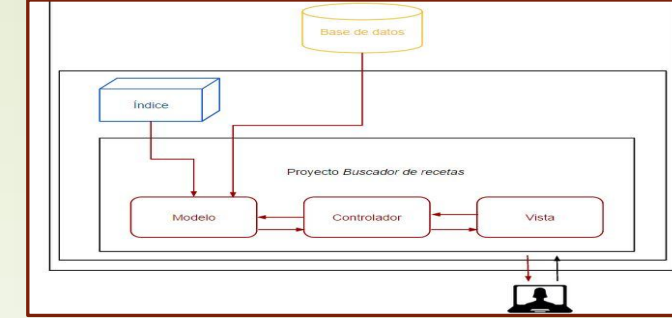
Requisitos funcionales: Sistema *Web Scraper*



- Capacidad de conexión con la página web.
- Capacidad de navegación por el sitio web de cocina.
- Capacidad de extracción de las direcciones de las recetas.
- Capacidad de extracción de los datos a partir de las direcciones obtenidas.
- Capacidad de tratamiento de datos.
- Capacidad de integración de los datos en una base de datos.

Análisis de requisitos

Requisitos funcionales: Sistema *Buscador*

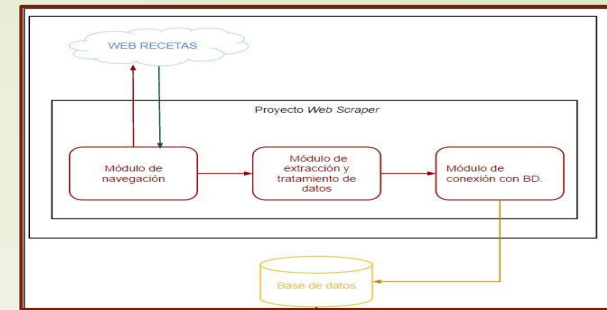


- Crear un índice basado en el sistema *Lucene* a partir de la base de datos.
- Posibilidad de utilización de dicho índice.
- Devolución de los resultados ordenados por puntuación de mayor a menor.
- Posibilidad de lanzar consultas sobre la base de datos.
- Mostrar listado de los resultados obtenidos.
- Visualización detallada de las recetas.

Análisis de requisitos

Requisitos no funcionales: Sistema Web Scraper

- Estabilidad del sistema.
- Rendimiento del sistema.
- Escalabilidad del sistema.



Análisis de requisitos

Requisitos no funcionales: Sistema *Buscador*

- Estabilidad del sistema.
- Rendimiento del sistema.
- Usabilidad del sistema.
- Portabilidad del sistema.
- Escalabilidad del sistema.

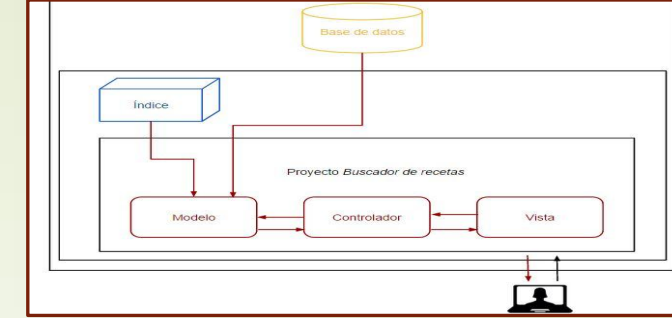


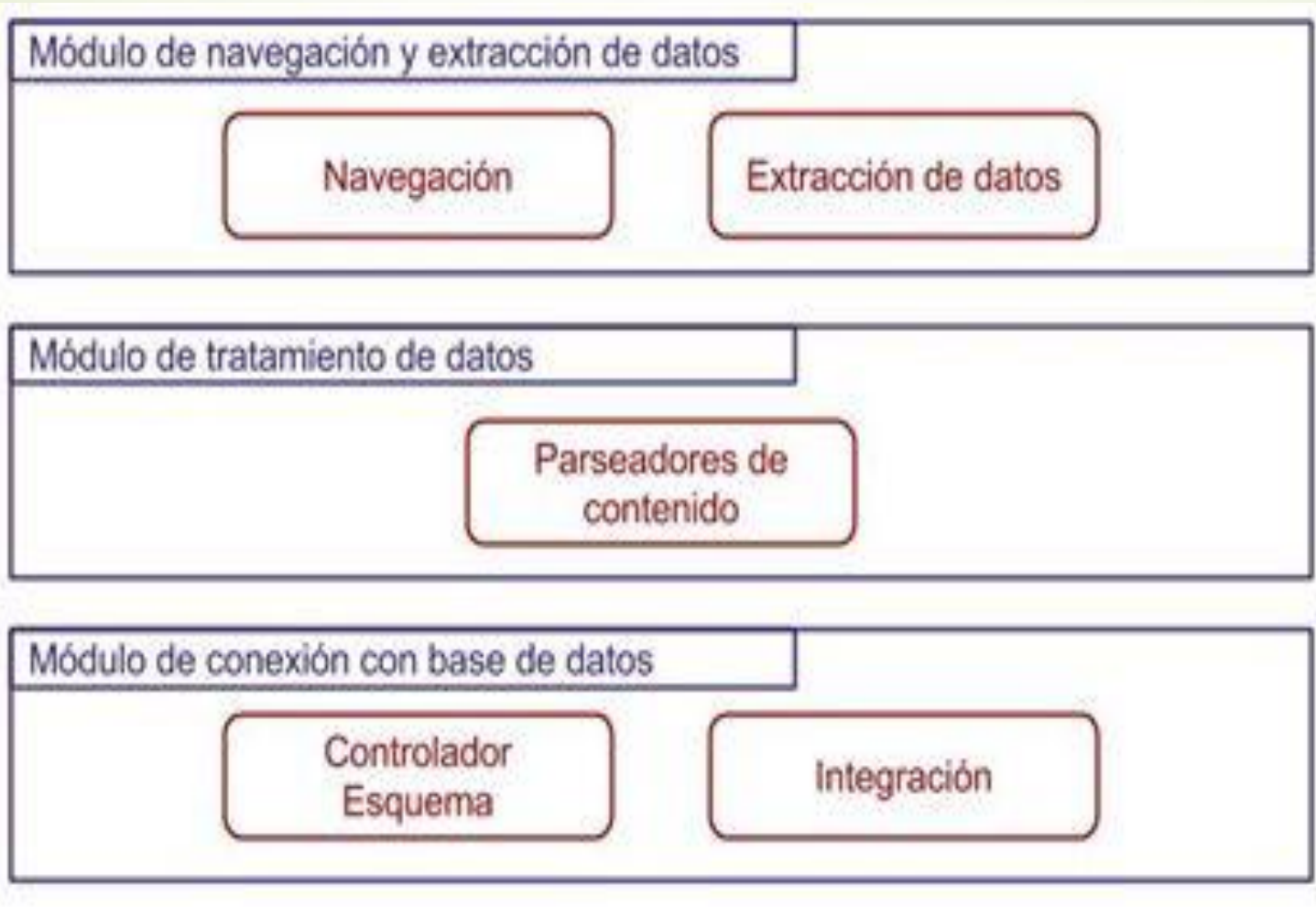
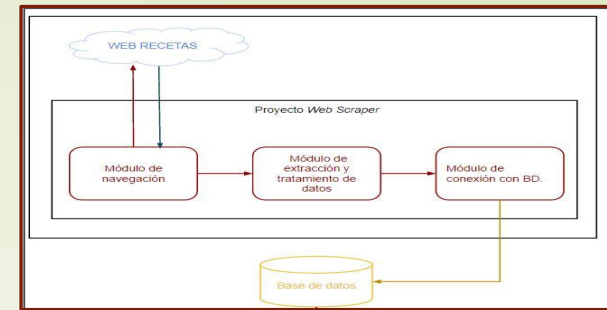


Tabla de contenidos

- Introducción
- Objetivos
- Análisis de requisitos
- Diseño
- Desarrollo
- Pruebas y resultados
- Conclusiones y Trabajo futuro

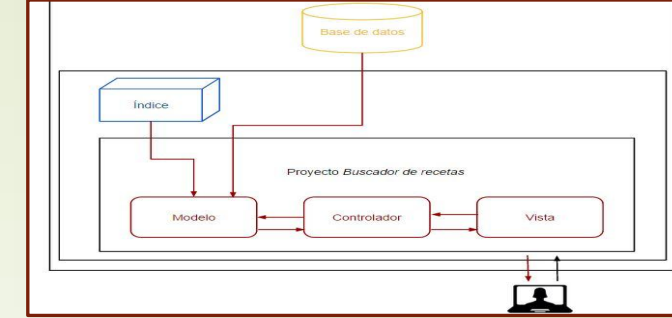
Diseño

Sistema *Web Scraper*



Diseño

Sistema *Buscador*



Módulo de presentación

Vistas de la aplicación

Módulo de lógica de negocio

Índice

Indexación

Parseo

Búsqueda

Conector con base de datos

Entidades del modelo

Módulo controlador

Controladores sobre el índice

Controladores sobre la base de datos



Tabla de contenidos

- Introducción
- Objetivos
- Análisis de requisitos
- Diseño
- Desarrollo
- Pruebas y resultados
- Conclusiones y Trabajo futuro

Implementación

Sistema Web Scraper: librería JSoup

- ✓ Codificación similar a *Jquery*.
- ✓ Permite manipular los elementos.
- ✓ Optimiza el código de recogida de datos.
- ✓ Utilización del método *UserAgent*.

1. Conexión con la web a través del método `connect`.



2. Obtención de las direcciones de las recetas a través del método `select`.



3. Recuperación de la información utilizando los métodos `css` de la librería.



4. Inserción de los elementos procesados en la base de datos relacional.

Implementación

Sistema Buscador: Librería Lucene

- ✓ Acceso eficiente a los datos.
- ✓ Flexibilidad en los tipos de búsqueda.
- ✓ Resultados relevantes y ranking.
- ✓ Actualización y búsqueda en el índice.

1. Tratamiento de la información suprimiendo *stopwords*.



2. Creación del índice de lucene.



3. Consulta de la información sobre el índice.



4. Devolución de los resultados relevantes.

Implementación

Sistema *Buscador*

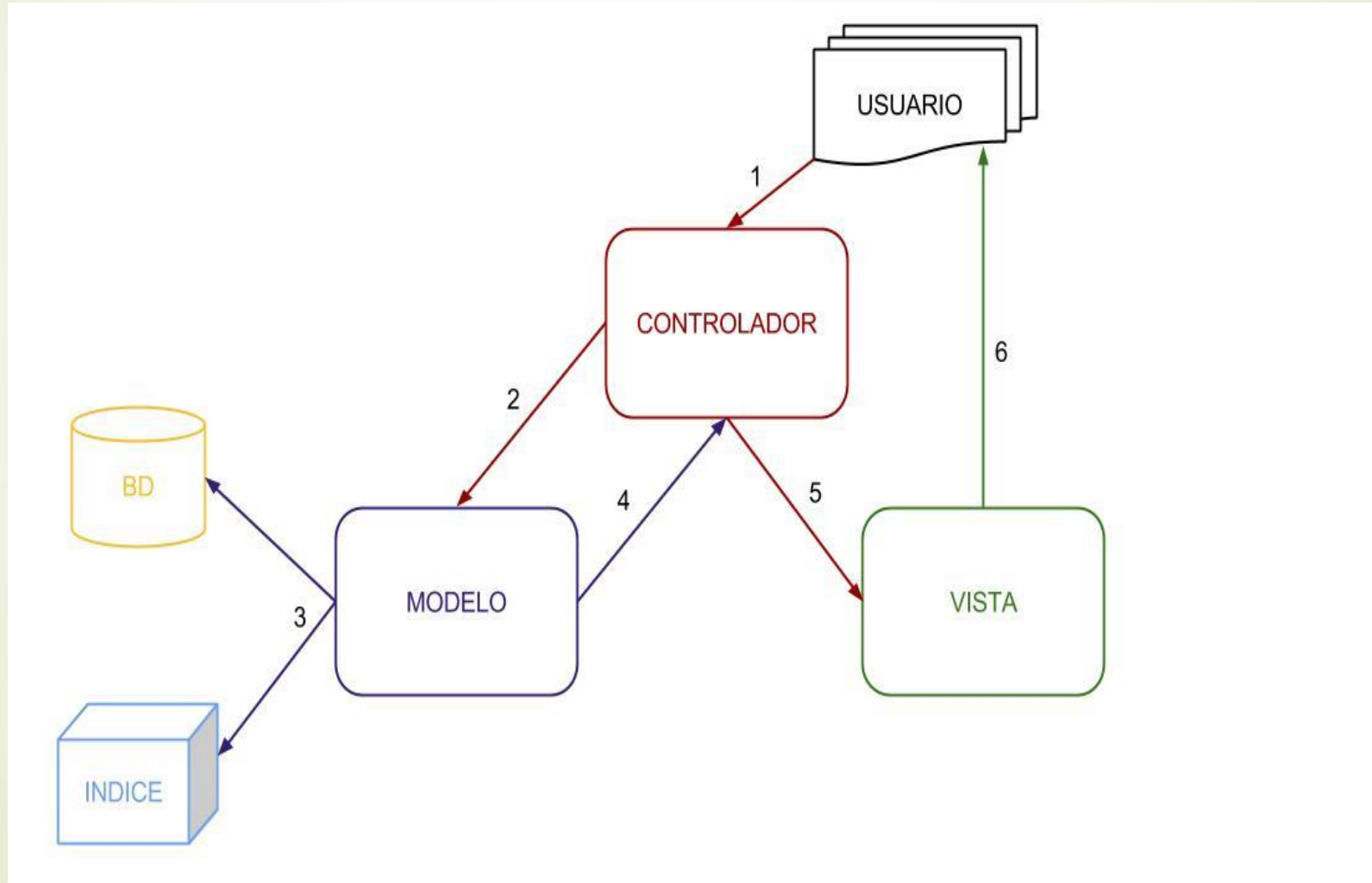




Tabla de contenidos

- Introducción
- Objetivos
- Análisis de requisitos
- Diseño
- Desarrollo
- Pruebas y resultados
- Conclusiones y Trabajo futuro



Pruebas

- ✓ Pruebas unitarias.
- ✓ Pruebas de integración.
- ✓ Pruebas de sistema.
- ✓ Pruebas de validación.
- Pruebas de aceptación.

Pruebas aceptación

Sistema Web Scraper

Table: RECIPES

New Record Delete Record

	recipeld	name	description	timePrep	timeCook	timeTotal	rating	category
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	0	Easy Rumaki wi...	Pineapple cube...	20 mins	20 mins	40 mins	4.5	appetizers and snacks
2	1	Brown Sugar S...	Mini smoked sa...	10 mins	20 mins	30 mins	4.754885	appetizers and snacks
3	2	Mouth-Wa						
4	3	Baked Buff						
5	2	Mouth-Waterin...	A restaurant-w...	25 mins	20 mins	45 mins	4.605215	appetizers and snacks
6								
7	6	Garlicky Ap						
8	7	Artichoke Heart...	Artichoke heart...	10 mins	10 mins	35 mins	4.210526	appetizers and snacks
9	8	Guacamole	Cilantro and ca...	10 mins	10 mins		4.807159	appetizers and snacks
10	9	Annie's Fruit Sal...	This delicious s...	15 mins	10 mins	45 mins	4.776792	appetizers and snacks
11	10	Buffalo Chicke...	Five simple ingr...	5 mins	40 mins	45 mins	4.704667	appetizers and snacks
12	11	Jalapeno Poppe...	This is an easy ...	10 mins	3 mins	13 mins	4.739302	appetizers and snacks
13	12	Mouth-Waterin...	A restaurant-w...	25 mins	20 mins	45 mins	4.605215	appetizers and snacks
14	13	Restaurant-Styl...	Restaurant-styl...	15 mins	15 mins	2 hrs	4.784651	appetizers and snacks
15	14	Double Tomato...	Bruschetta is a t...	15 mins	7 mins	35 mins	4.821642	appetizers and snacks

< < 1 - 16 of 52145 > >

Go to: 1

Pruebas aceptación

Sistema Buscador

Rating	Idfp--SV-Ni08-----	1.0	4.754885
description	Idfp--SV-Ni08-----	0.25	Mini smoked sausages disappear extra quickly blanketed crisp bacon topped brown-sugar glaze.
direction	Idfp--SV-Ni08-----	0.15625	4;Preheat oven 350 degrees F (175 degrees C).
direction	Idfp--SV-Ni08-----	0.15625	5;Cut bacon thirds wrap strip little sausage. Place wrapped sausages wooden skewers, skewer.
direction	Idfp--SV-Ni08-----	0.15625	6;Bake bacon crisp brown sugar melted.
ingredient	Idfp--SV-Ni08-----	0.25	bacon;1 pound
ingredient	Idfp--SV-Ni08-----	0.25	little smokie sausages;1 (16 ounce) package
ingredient	Idfp--SV-Ni08-----	0.25	brown sugar, taste;1 cup
name	Idfp--SV-Ni08-----	0.5	Brown Sugar Smokies
nutrient	Idfp--SV-Ni08-----	0.1875	Calories;356 kcal;18%
nutrient	Idfp--SV-Ni08-----	0.1875	Carbohydrates;18.9 g;6%
nutrient	Idfp--SV-Ni08-----	0.1875	Cholesterol;49 mg;16%
nutrient	Idfp--SV-Ni08-----	0.1875	Fat;27.2 g;42%
nutrient	Idfp--SV-Ni08-----	0.1875	Fiber;0 g;0%
nutrient	Idfp--SV-Ni08-----	0.1875	Protein;9 g;18%
nutrient	Idfp--SV-Ni08-----	0.1875	Sodium;696 mg;28%

Index name: C:\Users\g...ipeSearcher\resources\index

Pruebas aceptación

Sistema Buscador: Demostración



The screenshot shows a web application window titled "Recipe searcher for cooking". Below the title, the text "select the search type to start:" is displayed. Three orange buttons are arranged on the page: "Quick Search" on the left, "Advanced Search" on the right, and "Predefined Search" at the bottom center. A yellow circle highlights a mouse cursor hovering over the "Predefined Search" button.

Recipe searcher for cooking

select the search type to start:

Quick Search

Advanced Search

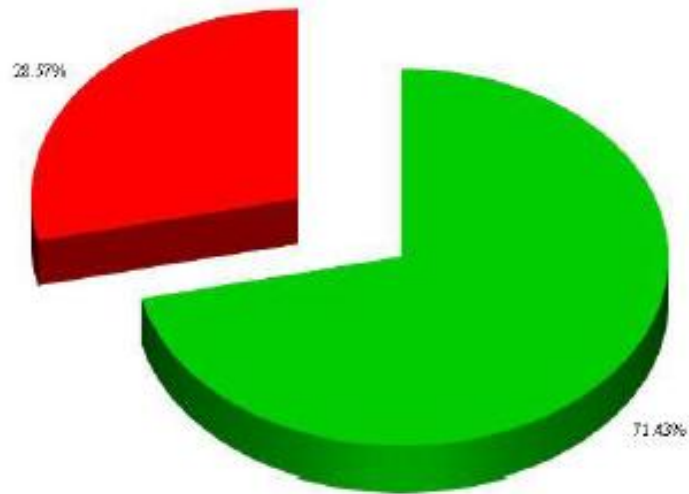
Predefined Search

Recorded with **SCREENCAST MATIC**

Pruebas aceptación

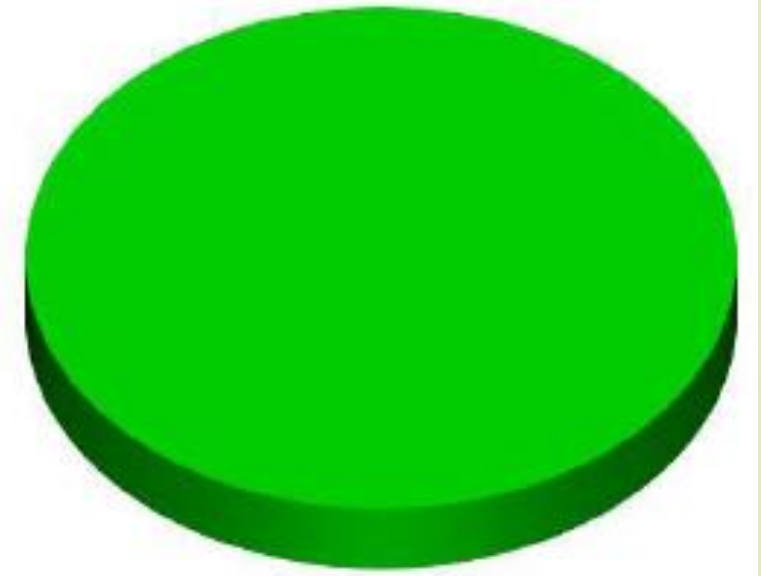
Sistema *Buscador*: Encuesta

¿Los resultados devueltos satisfacen la receta que andaba buscando?



■ Si ■ No

¿Le parece interesante la posibilidad de realizar filtros más avanzados?



■ Si



Tabla de contenidos

- Introducción
- Objetivos
- Análisis de requisitos
- Diseño
- Desarrollo
- Pruebas y resultados
- Conclusiones y Trabajo futuro

Conclusiones y trabajo futuro

Conclusiones

- Ciclo de vida del software.
- Nuevas técnicas tecnológicamente novedosos.
- Metodología de implementación modular en ambos proyectos.
- Patrón utilizado para las pruebas.
- Estudio de satisfacción del producto desarrollado.

Conclusiones y trabajo futuro

Trabajo futuro

- Adaptar una aplicación web.
- Sistema de recomendación.
- Extracción paralela de la información.
- Desarrollar un sistema de personalización de filtros.
- Sistema de gustos y preferencias.



Muchas gracias