



ANTEPROYECTO DEL TRABAJO DE FIN DE GRADO

INFORMACIÓN GENERAL

Alumno/a	Enrique López Encinas				
Titulación:	Grado en Ingeniería Informática				
Tutor/es:	Eduardo Guzmán De los Riscos				
Título	Herramienta de crawling de publicaciones científicas				
Subtítulo <i>(solo si en grupo)</i>					
Título en inglés	Scientific publications crawling tool				
Subtítulo en inglés <i>(solo si en grupo)</i>					
Trabajo en grupo:	<input checked="" type="checkbox"/> Sí	<input type="checkbox"/>	<input type="checkbox"/> No	<input checked="" type="checkbox"/> X	
Otros integrantes del grupo:					

INTRODUCCIÓN

Contextualización del problema a resolver. Describir claramente de dónde surge la necesidad de este TFG y el dominio de aplicación. En caso de que el TFG se base en trabajos previos, debe aclararse cuáles son las aportaciones del TFG.

Después de consultar diferentes sitios web y servicios de consulta sobre la producción científica de los investigadores, es decir, artículos de investigación, etc. queda patente que cada uno de estos sitios muestra la información de forma diferente, dificultando la lectura y el análisis de la información que se busca conseguir, en primer lugar, por tener que acudir al uso de diferentes herramientas, y también por la propia disposición de la información que queda a criterio del autor o autores de cada una de estas herramientas.

El objetivo de este TFG es desarrollar una herramienta, con una arquitectura basada en microservicios, que contenga servicios de extracción de la información en Python, procesamiento y normalización de estos datos, y una pequeña aplicación front-end web encargada de integrar los datos obtenidos en un modelo que limpie los datos de entrada y los estructure como paso previo al análisis.

OBJETIVOS

Descripción detallada de en qué consistirá el TFG. En caso de que el objeto principal del TFG sea el desarrollo de software, además de los objetivos generales deben describirse sus funcionalidades a alto nivel.

Con este TFG se busca desarrollar una aplicación web que recopile información de varios sitios web dedicadas a la divulgación de artículos académicos (por ejemplo, Google Scholar) así como de sus autores, a partir de las búsquedas realizadas por los usuarios de la herramienta. De esta forma, una vez que un usuario realice una búsqueda de una línea de investigación, los servicios de extracción se encargarán de realizar las búsquedas en las fuentes de datos. Posteriormente, toda esa información se integrará, normalizará y almacenará en la base de datos para que el usuario pueda consultarla directamente, así como realizar ciertos análisis sobre los resultados.

El sistema tendrá por tanto una base de datos relacional donde se guardará toda la información que se obtenga de las búsquedas de los usuarios, que se realizará a través de una aplicación web basada en Angular.

Esta información se obtendrá usando el método de “web scraping”, a través de scripts capaces de recoger y organizar todos los datos de una o más fuentes web.

En resumen, se busca desarrollar una herramienta dirigida a investigadores que facilite la tarea de búsqueda de artículos académicos de diferentes fuentes filtrados por los parámetros que indique el investigador para ser tratados, organizados y guardados en una base de datos, que se podrá consultar a través de la aplicación web, mostrando por pantalla los datos más relevantes para su posterior análisis.

ENTREGABLES

Listado de resultados que generará el TFG (aplicaciones, estudios, manuales, etc.)

Aplicación web

Memoria del TFG

MÉTODOS Y FASES DE TRABAJO

METODOLOGÍA:

Descripción de la metodología empleada en el desarrollo del TFG. Especificar cómo se va a desarrollar. Concretar si se trata de alguna metodología existente y, en caso contrario, describir y justificar adecuadamente los métodos que se aplicarán.

Se usará una metodología ágil, que permite gran flexibilidad a la hora de realizar las diferentes tareas del proyecto según se vayan necesitando.

A pesar de existir diferentes fases de trabajo en el proyecto, esta metodología facilita saltar de una a otra si se encuentran errores que necesitan ser arreglados con urgencia o se necesita rehacer o cambiar alguna parte anterior.

FASES DE TRABAJO:

Enumeración y breve descripción de las fases de trabajo en las que consistirá el TFG.

Contextualización del proyecto: Esta fase está enfocada en entender el alcance del proyecto. Tienen que quedar claras las fuentes de información (páginas web) de la que obtener los datos, así como las tecnologías necesarias para realizar el proyecto según los requisitos que se definan.

Fase de aprendizaje: En esta fase el alumno investigará las tecnologías que requiera para la realización del TFG, así como aprender lo básico para manejarse con ellas.

Fase de desarrollo del crawler: Se desarrollará el script en Python que se encarga de realizar el crawling a Google Scholar y de la limpieza y estructuración de los datos para su uso posterior.

Fase de desarrollo del Back-end y Front-end: En estas dos etapas se montará la aplicación web donde se guardarán y mostrarán los datos obtenidos al usuario final. Consiste en la instalación de la base de datos y el desarrollo de la aplicación web con Angular.

Documentación y pruebas: Durante todo el proyecto se llevarán a cabo periodos de pruebas del código para verificar que todo funcione como debe, y se documentará conforme se vaya trabajando en el proyecto.

Memoria final del proyecto: Finalmente se elaborará la memoria del proyecto.

TEMPORIZACIÓN:

La siguiente tabla deberá contener una fila por cada una de las fases enumeradas en la sección anterior. En caso de tratarse de un trabajo en grupo, se añadirá una columna HORAS por cada miembro del equipo. Debe especificarse claramente el número de horas dedicado por cada alumno/a y la suma de horas individual deberá ser también de 296.

FASE	HORAS
	Enrique López Encinas
Contextualización del proyecto	24
Investigar páginas web que se usarán para extraer datos	8
Definir tecnologías y lenguajes que se usarán	4
Estructurar trabajo a realizar durante el proyecto	4
Rellenar plantilla anteproyecto	5
Crear repositorio git para guardar el proyecto	3
Aprendizaje de los lenguajes y tecnologías a usar.	43
Aprendizaje Python	4
Aprendizaje HTML y CSS	4
Aprendizaje Typescript	8
Aprendizaje Git	3
Aprendizaje Angular	16
Aprendizaje de técnicas de scraping de páginas web	8
Desarrollo de los Crawlers	64
Realizar crawler para Google Scholar	24
Realizar crawler para otra fuente de datos	24
Script para estructurar y exportar los datos obtenidos a la BBDD	16
Desarrollo del Back-end	36
Creación de la base de datos relacional	8
Crear modelo lógico y tablas a usar	24
Introducir los datos obtenidos en la base de datos	4
Desarrollo del Front-end	72
Crear proyecto Angular	8
Creación página Inicio	8
Creación de página para introducir filtros de búsqueda	24
Creación de página para visualización de los datos obtenidos	16
Dar estilo a las páginas anteriores	16
Documentación y pruebas	36
Pruebas y Test sobre el crawler	8
Pruebas y Test sobre Back-end	8
Pruebas y Test sobre el Front-end	8
Documentación de los scripts del crawler	4
Documentación del Backend	4
Documentación del Frontend	4
Realización de la memoria final del proyecto	21
	296



ENTORNO TECNOLÓGICO

TECNOLOGÍAS EMPLEADAS:

Enumeración de las tecnologías utilizadas (lenguajes de programación, frameworks, sistemas gestores de bases de datos, etc.) en el desarrollo del TFG.

Python, Angular, Typescript, HTML, CSS, Git.

Gestor de base de datos relacional (MySQL, MariaDB, PostgreSQL, ...)

RECURSOS SOFTWARE Y HARDWARE:

Listado de dispositivos (placas de desarrollo, microcontroladores, procesadores, sensores, robots, etc.) o software (IDE, editores, etc.) empleados en el desarrollo del TFG.

Visual Studio Code.

REFERENCIAS

Listado de referencias (libros, páginas web, etc.)

[Google Scholar](#)

[Web of Science](#)

[Exaly](#)

[ORCID](#)

[DBPL](#)

[ResearchGate](#)

Buelta, J. (2018). *Python Automation Cookbook: Explore the world of automation using Python recipes that will enhance your skills*. Packt Publishing Ltd.

Vanden Broucke, S., & Baesens, B. (2018). *Practical Web scraping for data science* (pp. 3-5). New York, NY: Apress.

[AngularJS](#)

Málaga, _____ de _____ de _____

Firma tutor/tutora:

Firma cotutor/a:

Firma tutor/a coordinador/a: