Resumo do Traballo de Fin de Grao

Introducción

Github es hoy en día una de las plataformas más populares entre la comunidad de desarrolladores, sirviendo de base para la realización de numerosos estudios y análisis gracias a la gran cantidad de datos relacionados con el desarrollo de software que aloja.

La minería de datos es una rama de las ciencias de la computación que permite encontrar patrones en grandes volúmenes de datos haciendo posible el desarrollo de modelos descriptivos, prescriptivos y clasificadores capaces de generar nuevo conocimiento a partir de la información almacenada. Con ayuda de la minería se podría llegar a conocer cómo es la actividad de los desarrolladores dentro de Github: qué proyectos están en auge, dónde está el foco en el desarrollo de software, cuáles son las categorías emergentes, qué lenguaje de programación es el más utilizado, qué tipo de proyectos gusta más a los usuarios, quiénes son los desarrolladores de referencia

En este contexto, el objetivo general del trabajo es realizar un análisis masivo de datos que pasa por la creación y gestión de una base de datos (extraídos de la plataforma Github) para llevar a cabo la identificación de patrones y desarrollo de modelos con el fin de dotar de valor a los datos almacenados y poder recabar información sobre la plataforma. La finalidad es realizar un estudio cualitativo, descriptivo y predictivo de la plataforma para conocer su evolución y cuál es la actualidad en el desarrollo de software.

Objetivos

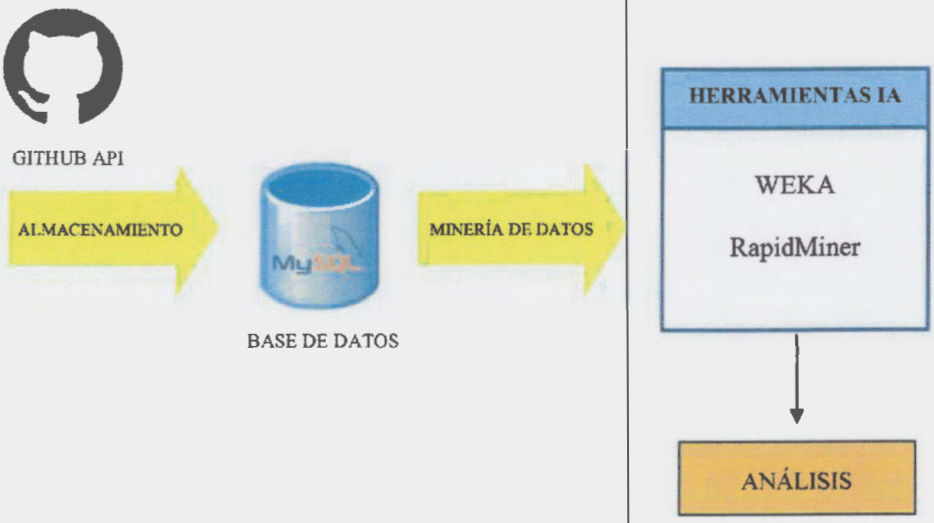
Para la consecución del objetivo general anteriormente planteado, se identifican los siguientes objetivos concretos a lograr con el presente trabajo:

- Creación y gestión de una base de datos, empezando por la descarga de los datos de la plataforma Github y su posterior mantenimiento, con inserciones y actualizaciones automatizadas mediante código.
- Transformación y depuración de los datos con el fin de prepararlos para su análisis: selección, limpieza, enriquecimiento, reducción y adaptación.
- Creación de modelos analíticos a partir de patrones identificados usando minería de datos. Los modelos serán evaluados para conocer su grado de idoneidad y serán usados para la obtención de nueva información a partir de la base de datos inicial.
- Realización del estudio del dominio y obtención de respuestas a preguntas como: ¿cuál es el lenguaje de programación más utilizado?, ¿qué tipo de licencia es la más frecuente?, ¿cuáles son las categorías con un mayor número de proyectos? o ¿qué proyectos son los más populares?

Arquitectura

A continuación se presenta un esquema resumido de la arquitectura.

Resumo do Traballo de Fin de Grao



Los datos serán descargados gracias a la API de Github hacia una base de datos. Posteriormente se aplicarán técnicas de IA mediante software de aprendizaje automático como Weka y/o RapidMiner. Con ayuda de estas plataformas se podrá obtener respuesta a las preguntas anteriormente mencionadas y con ello concluir el estudio.

Tecnologías

- REST API de Github para la conexión con la plataforma y descarga de datos.
- Uso de librerías para la manipulación de datos en formato JSON.
- MySQL como gestor de base de datos y SQL como lenguaje de bases de datos relacionales.
- JAVA, como lenguaje de programación orientado a objetos.
- RapidMiner y/o Weka para la minería de datos y el apoyo en la identificación de patrones y creación de modelos.

Proceso de desarrollo

El proceso de desarrollo elegido es el Proceso Unificado. Caracterizado por estar dirigido por casos de uso, centrado en la arquitectura y por ser iterativo e incremental. Es iterativo por estar cada etapa dotada de una serie de repeticiones. Es incremental porque cada iteración añade o mejora las funcionalidades del sistema.