16002653 Alejandra Sierra
Product Development
Postgrado en Inteligencia de Negocios
Universidad Galileo

Investigación de Tecnologías para MLOps

Investigación de Tecnologías de Docker, PyCaret, MLFlow, DagsHub, DVC, CokieCutter y Flask:

¿Qué es Docker?

El término "Docker" se aplica a diferentes conceptos, entre los que se incluyen un proyecto de la comunidad open source y sus herramientas; Docker Inc., la principal empresa promotora del proyecto; y las herramientas que la empresa respalda formalmente. El hecho de que las tecnologías y la compañía compartan el mismo nombre puede ser confuso.

- El sistema de software de TI llamado "Docker" es la tecnología de organización en contenedores que posibilita la creación y el uso de los contenedores de Linux®.
- La comunidad open source Docker se encarga de mejorar estas tecnologías para beneficiar a todos los usuarios.
- La empresa, Docker Inc., se basa en el trabajo de la comunidad Docker para aumentar la seguridad de las herramientas y comparte los avances con el resto de la comunidad. Entonces, brinda soporte a las tecnologías mejoradas y reforzadas para los clientes empresariales.

¿Qué es PyCaret?

PyCaret es una poderosa biblioteca de código abierto para el procesamiento del lenguaje natural (NLP). Está diseñado para hacer que el proceso de creación, capacitación e implementación de modelos de aprendizaje automático sea más fácil y rápido. PyCaret proporciona un conjunto completo de herramientas para el preprocesamiento de texto, la ingeniería de características, el entrenamiento de modelos y la implementación.

PyCaret tiene varias ventajas sobre otras bibliotecas NLP. Primero, es fácil de usar y requiere una codificación mínima. Esto lo hace ideal para principiantes que recién comienzan con la PNL. En segundo lugar, es altamente personalizable y permite a los usuarios modificar fácilmente los modelos existentes para adaptarlos a sus necesidades. En tercer lugar, es rápido y eficiente, lo que permite a los usuarios entrenar e implementar modelos rápidamente. Finalmente, es de código abierto, por lo que los usuarios pueden modificar y compartir libremente el código.

16002653 Alejandra Sierra
Product Development
Postgrado en Inteligencia de Negocios
Universidad Galileo

¿Qué es MLflow?

MLflow es una herramienta increíblemente útil para gestionar el ciclo de vida de tus proyectos de machine learning. No solo te permite llevar un seguimiento de tus experimentos, sino que también facilita el proceso de guardar y cargar modelos, y de implementarlos en producción.

Mlflow sirve para poner orden al trabajo de ensayo y error de los *data scientists.* La plataforma permite hacer un seguimiento del entrenamiento de modelos, de las pruebas, con los datos utilizados, los parámetros y cambios realizados.

Así pues, los usuarios pueden guardar todos los experimentos realizados y recuperarlos en cualquier momento, algo especialmente útil cuando el modelo finalmente funciona. Además, se puedes empaquetar todo el proceso de entrenamiento de forma que sea reproducible.

De esta forma, se pueden crear diferentes experimentos, usuarios, trabajar en grupos, saber quién ha hecho qué. Y además del seguimiento y de poder recuperar los ensayos, mlflow también permite **servir fácilmente los modelos** para que otras personas, un cliente por ejemplo, lo puedan probar.

¿Qué es DAGsHub?

DAGsHub es una plataforma web para el control de versiones de datos y la colaboración para científicos de datos e ingenieros de aprendizaje automático. La plataforma se basa en DVC, un sistema de control de versiones de código abierto para proyectos de aprendizaje automático que funciona perfectamente.

¿Qué es DVC?

Se refiere a los sistemas de codificación de video digital utilizados para comprimir y almacenar archivos de video de alta calidad en formatos digitales. Estos sistemas permiten reducir el tamaño de los archivos de video para facilitar su transferencia y almacenamiento, sin comprometer la calidad de la imagen.

La tecnología de DVC incluye diferentes algoritmos de codificación de video, como MPEG (Moving Picture Experts Group), que utiliza técnicas de compresión basadas en la eliminación de redundancias espaciales y temporales, y H.264/AVC (Advanced Video Coding), que utiliza métodos de codificación más avanzados para lograr una mayor eficiencia de compresión.

16002653 Alejandra Sierra
Product Development
Postgrado en Inteligencia de Negocios
Universidad Galileo

¿Qué es CokieCutter?

CokieCutter es una herramienta de tecnología que se utiliza para crear rápidamente nuevos proyectos o componentes basados en plantillas predefinidas. Esta tecnología permite a los desarrolladores automatizar la configuración inicial y generar automáticamente la estructura de archivos y código necesarios para comenzar un nuevo proyecto. La tecnología de CokieCutter puede ser utilizada en diferentes áreas, como desarrollo de software, desarrollo web y desarrollo de aplicaciones móviles. Al usar plantillas predefinidas, los desarrolladores pueden ahorrar tiempo y esfuerzo al evitar tener que crear desde cero los mismos tipos de proyectos una y otra vez.

¿Qué es Flask?

La tecnología de Flask es un framework ligero de desarrollo web para Python. Es una herramienta que permite construir aplicaciones web de forma rápida y sencilla, siguiendo el patrón de diseño MVC (Model-View-Controller).

Flask incluye funcionalidades básicas como enrutamiento de URLs, conexiones a bases de datos, manejo de cookies y sesiones, así como la gestión de peticiones y respuestas HTTP.

Una de las ventajas de Flask es su simplicidad y flexibilidad. No impone una estructura rígida en el desarrollo de aplicaciones, permitiendo al desarrollador tomar decisiones sobre la organización del código y la arquitectura de la aplicación.

Adicionalmente para cada tecnología deberá responder las siguientes preguntas:

1) ¿Qué hace la Tecnología de Docker, PyCaret, MLFlow, DagsHub, DVC, CokieCutter y Flask en particular?

- La tecnología de Docker es una plataforma de contenedores que permite empaquetar y distribuir aplicaciones junto con todas sus dependencias en un entorno aislado. Proporciona una forma eficiente y consistente de implementar aplicaciones en diferentes entornos, lo que facilita la portabilidad y la escalabilidad.

- PyCaret es una biblioteca de aprendizaje automático de código abierto que se utiliza para simplificar el flujo de trabajo de creación de modelos. PyCaret proporciona una interfaz fácil de usar que permite a los investigadores y científicos de datos realizar tareas comunes de aprendizaje automático,

como selección de características, ajuste de modelos y evaluación, de manera eficiente.

- MLFlow es una plataforma de código abierto para el ciclo de vida de Machine Learning. Proporciona herramientas y funcionalidades para gestionar, rastrear, visualizar y comparar experimentos de Machine Learning.

- MLFlow proporciona una solución integral para el ciclo de vida de Machine Learning, desde la gestión de experimentos hasta el registro, seguimiento y despliegue de modelos entrenados.

- DagsHub es una plataforma de colaboración y gestión de proyectos de ciencia de datos y aprendizaje automático. Ofrece una serie de herramientas y características que permiten a los equipos de investigación y desarrollo trabajar de manera más eficiente y colaborativa.

  Mejora la colaboración, la gestión de datos, la reproducibilidad y la eficiencia en proyectos de ciencia de datos y aprendizaje automático, facilitando el trabajo en equipo y la obtención de resultados confiables y reproducibles.

- DVC, que significa Data Version Control (Control de Versión de Datos), es una tecnología que se utiliza para el seguimiento y control de versiones de datos en proyectos de ciencia de datos y aprendizaje automático.

  DVC proporciona un sistema de control de versiones para los datos necesarios en un proyecto, al igual que Git proporciona un sistema de control de versiones para el código fuente. Con DVC, los científicos de datos pueden manejar los cambios en los datos, realizar un seguimiento de las versiones anteriores de los datos y colaborar con otros miembros del equipo en la gestión y control de los datos.

  Además, DVC permite a los científicos de datos y desarrolladores de aprendizaje automático gestionar los flujos de trabajo de procesamiento de datos y entrenamiento de modelos, lo que facilita la reproducción de resultados y la iteración en el desarrollo de modelos.

- La tecnología de CookieCutter, en particular, es una herramienta que se utiliza para generar automáticamente estructuras o proyectos repetitivos en el desarrollo de software. Permite crear plantillas personalizables y reutilizables para la generación de código, evitando tener que escribirlo desde cero en cada proyecto.

  CookieCutter se basa en el concepto de "cortadores de galletas" (cookie cutters), que son moldes utilizados para dar forma a las galletas de manera rápida y consistente. De manera similar, la tecnología de CookieCutter

permite definir una plantilla con una estructura predeterminada y configuración específica, que se puede utilizar para generar y personalizar proyectos similares de forma ágil.

- Flask es un marco de desarrollo web en Python que se utiliza para crear aplicaciones web y servicios RESTful. Proporciona herramientas y bibliotecas para facilitar el desarrollo web, como enrutamiento de URL, manejo de solicitudes y respuestas, manejo de cookies y sesiones, generación de HTML dinámico y mucho más.

  La tecnología de Flask se centra en la simplicidad y la flexibilidad. A diferencia de otros marcos web más completos, Flask tiene una estructura mínima y no impone una arquitectura específica, lo que permite a los desarrolladores crear aplicaciones web de manera más libre según sus necesidades.

  La tecnología de Flask permite a los desarrolladores crear aplicaciones web en Python de manera rápida y sencilla, proporcionando herramientas y funcionalidades para facilitar el desarrollo de diferentes aspectos de una aplicación web.

2) ¿Según su experiencia, para qué podría servirnos esta Tecnología de Docker, PyCaret, MLFlow, DagsHub, DVC, CokieCutter y Flask en la implementación de un proyecto de ML?

- La tecnología de Docker puede ser utilizada en un proyecto de Machine Learning (ML) para crear y gestionar entornos aislados y reproducibles. Docker permite empaquetar el código, dependencias y configuraciones del proyecto en un contenedor, lo que facilita la portabilidad y distribución del proyecto entre diferentes sistemas.

- PyCaret es una biblioteca de Python que proporciona una interfaz de alto nivel para el desarrollo de modelos de ML. PyCaret simplifica el flujo de trabajo de ML al automatizar tareas comunes como la preparación de datos, selección de modelos, entrenamiento y evaluación. Esto permite acelerar el desarrollo y prototipado de modelos de ML.

- MLFlow es una plataforma de código abierto para el ciclo de vida de ML. Permite gestionar y rastrear experimentos de ML, realizar el seguimiento de métricas, almacenar y comparar modelos, y facilitar la implementación en diferentes entornos. MLFlow facilita la colaboración y gestión de versiones en proyectos de ML.

- DagsHub es una plataforma de control de versiones basada en Git, diseñada específicamente para proyectos de ciencia de datos y ML. DagsHub

proporciona características adicionales orientadas a las necesidades de los proyectos de ML, como la visualización de Notebooks Jupyter, seguimiento de código y datos, y la colaboración con otros investigadores.

- DVC (Data Versión Control) es una herramienta para controlar la versión de los conjuntos de datos utilizados en los proyectos de ML. DVC permite gestionar y rastrear cambios en los datos, compartirlos entre diferentes miembros del equipo y sincronizar los datos con el control de versiones del código.

- Cookiecutter es una herramienta de generación de proyectos que permite automatizar la estructura y configuración inicial de un proyecto de ML. Cookiecutter proporciona plantillas predefinidas que pueden adaptarse al proyecto específico, lo que facilita la configuración inicial del proyecto y asegura una estructura coherente y organizada.

- Flask es un framework de desarrollo web ligero en Python que puede ser utilizado para implementar un servicio de API para el despliegue de modelos de ML. Flask permite crear una interfaz accesible a través de endpoints HTTP para enviar solicitudes al modelo entrenado y recibir las predicciones correspondientes.

3) ¿Hay alguna otra tecnología sustituta para este caso? Mencione y describa brevemente.

Sí, existen varias tecnologías sustitutas a las mencionadas entre ellas:

Kubernetes: es una plataforma de orquestación de contenedores de código abierto que permite administrar y escalar aplicaciones en contenedores de manera eficiente. Puede considerarse como una alternativa a Docker, ya que Docker es uno de los contenedores más utilizados junto con Kubernetes para implementaciones escalables.

Anaconda: Anaconda es una plataforma completa para la ciencia de datos y el aprendizaje automático. Proporciona una distribución de Python que incluye una amplia gama de bibliotecas y herramientas populares para el análisis de datos y el aprendizaje automático.

TensorFlow: es una biblioteca de código abierto para el aprendizaje automático desarrollada por Google. Es muy utilizado en aplicaciones de inteligencia artificial y aprendizaje automático, y puede considerarse como una alternativa a PyCaret y MLFlow.

GitLab: es una plataforma de gestión de repositorios de código fuente y colaboración, similar a DagsHub y DVC. Ofrece características como integración continua, seguimiento de problemas y edición colaborativa, lo que lo convierte en una alternativa completa para el desarrollo de proyectos de aprendizaje automático.

FastAPI: FastAPI es un marco web de Python de alto rendimiento y fácil de usar para crear API RESTful. Es una alternativa a Flask, ya que proporciona una mayor rapidez y rendimiento para el desarrollo de servicios web.