SPRINT #1

QuanTile Analytics

Documentación primer semana

Integrantes:

- Albariño Damián N.
- Castillo Jonathan
- Paniceres Lucio
- Tonetto Jeferson
- Zenobio Bruno

Contexto

La opinión de los usuarios es un dato muy valioso. Su análisis puede ser determinante para la planificación de estrategias. Nos brindan información de Yelp, una plataforma de reseñas de todo tipo de negocios, restaurantes, hoteles, servicios entre otros. Además, Google posee una plataforma de reseñas de todo tipo de negocios, integrada en su servicio de localización y mapas, Google Maps. Los usuarios utilizan el servicio y luego suben su reseña según la experiencia que han recibido.

Pertenecemos a la consultora de data, QuanTile Analytics, nos han contratado para poder realizar un análisis del mercado estadounidense. Nuestro cliente es parte de un conglomerado de empresas de restaurantes y afines, y desean tener un análisis detallado de la opinión de los usuarios en Yelp y cruzarlos con los de Google Maps. Además, nos solicitan saber dónde es conveniente emplazar los nuevos locales de restaurantes y afines, y desean poder tener un sistema de recomendación de restaurantes para los usuarios de ambas plataformas para darle, al usuario por ejemplo la posibilidad de poder conocer nuevos sabores basados en sus experiencias previas.

KPIs:

- **Enfoque técnico: Optimización del alcance y calidad del servicio**
    - Mejorar un 10% la accesibilidad de los restaurantes en Florida
        - Cantidad de restaurantes por estado
    - Evaluar calidad media de productos por categoría de restaurantes
        - Correlación entre las estrellas y las 20 categorías con mayor frecuencia
- **Enfoque orgánico: Networking**
    - Mejorar la tasa de diferencia entre reviews y respuestas
        - Ratio: respuestas / reviews
    - Promover la tasa de retención de cliente
        - Cantidad de meses distintos con review para usuario
    - Aumentar 10% reviews de usuarios influyentes
        - Cantidad de amigos por usuario

Análisis Exploratorio de Datos:

Nos encontramos con la información sobre las reseñas de Yelp y Google Maps, las reseñas están en un rango temporal de 2017 hasta 2022, en lo provisto además de las reseñas podemos ver que tenemos información sobre negocios, una diferencia que pudimos observar es que en el dataset de Google Maps tenemos datos de los negocios de todos los 51 estados y en Yelp solo de 15 estados, por esa razón nos vemos obligados a reducir la cantidad de estados a analizar.

Además de tener la información del estado al que pertenece el negocio también podemos ver las categorías a la cual pertenece el negocio, es decir el tipo de negocio, ejemplo: restaurante, shopping, estación de servicio, etc. Haciendo el análisis de qué tipo de negocios tenemos más información, podemos observar que en ambos datasets el negocio con mayor cantidad de locales son los restaurantes, por ende nos vamos a centrar en este tipo de negocio.

Uniendo esta información, optamos por la utilización de los 4 estados con más restaurantes registrados en ambas plataformas para su análisis, que nos da como resultado California, New Jersey, Florida e Illinois.

Producto:

Sistema de recomendación para empresas: Análisis de ubicación

- Por cantidad competencia en la zona
- Por calidad competencia en la zona
- Por horario por atributo

Sistema de recomendación para empresas: Publicidad

- Promoción por gustos del usuario
- Por gustos de usuarios similares
- Por gustos de amigos del usuario

Stack elegido:

Python, con las librerías: Pandas, JSON, Pyarrow, Numpy, NLTK, PySpark, Matplotlib, Seaborn.

Pandas, Numpy: Librería de Python para el tratado tabular de los datos.

PySpark: Proporciona interfaz en Python para trabajar con Spark, el cual está diseñado para el procesamiento de grandes volúmenes de datos.

Pyarrow:

Matplotlib, Seaborn: Utilizamos estas librerías para realizar análisis exploratorio, y visualización de los datos en forma gráfica.

NLTK:(Natural Language Toolkit) Librería de Python para trabajar con lenguaje humano, aporta herramientas para analizar el tono emocional, permitiendo un análisis de sentimiento sin tener que realizar el entrenamiento de un modelo.

boto3: Librería de Python que permite interactuar con servicios de AWS como Amazon s3.

Docker: es nuestra elección como sistema para la gestión de contenedores, lo que nos permite trabajar dinámicamente en varias plataformas sin la complicación de lidiar directamente con el control de versiones. Esta herramienta nos proporciona un esquema eficiente para implementar distintas imágenes, lo que simplifica enormemente la manipulación de grandes volúmenes de datos.

Amazon S3 AWS: Se emplea para almacenar y recuperar datos a través de internet. Es altamente escalable, lo que implica su capacidad para manejar desde pequeñas cantidades hasta grandes volúmenes de datos sin dificultad. Está diseñado para ser altamente disponible y duradero. Los datos se almacenan en "buckets" (cubos), que son contenedores de almacenamiento donde es posible organizar y configurar los permisos de acceso. Esta característica facilita el trabajo en equipo sobre una misma base de datos.

GitHub: Optamos por esta herramienta como nuestra plataforma de control de versiones, esencial para la colaboración en equipo. Facilita la disponibilidad y sincronización de procesos, manteniendo el orden y gestionando las distintas versiones del proyecto de manera eficiente.

Mega, Google Cloud Storage y Google Drive: Optamos por estas opciones de almacenamiento para mantener un control centralizado de nuestros datos y data sets optimizados para su uso en la nube.

Flujo de Trabajo:

Data sources:

Google Maps, Yelp

ETL-EDA:

Python, Pandas, JSON, Pyarrow, Numpy, NLTK, Docker, PySpark, Matplotlib, Seaborn.

Data Store:

GitHub, Mega, Google Cloud Storage, Google Drive, Amazon S3 AWS

Visualization:

Google Sheet