PROYECTO FINAL

GOOGLE + YELP – Grupo 8

INFORME DE AVANCE – SEMANA 1

Introducción

La industria Digital está creciendo a pasos agigantados en los últimos años. La competencia es cada vez mayor, incluso para las empresas pequeñas y locales.

Las reviews online hoy en día se han vuelto más importantes que los anuncios locales y la difusión de boca en boca. De hecho, se estima que 9 de cada 10 clientes leen reseñas en línea antes de decidirse a realizar una compra o visitar algún sitio turístico. Por lo tanto, estas críticas (negativas y positivas) pueden tener un impacto significativo en los resultados finales del propietario de una pequeña empresa.

Google My Bussiness y Yelp son las dos plataformas principales que ayudan a las empresas a generar reseñas online y mejorar su reputación.

En el caso de YELP, es considerado el “Facebook” de los sitios de reviews. Es una de las webs más populares para empresas de servicios locales, tales como plomeros, electricistas, etc. Los usuarios de su comunidad, llamados "Yelpers", pueden publicar buenas reseñas, fotos, calificaciones e interactuar con otros clientes potenciales para ayudarlos a encontrar un negocio adecuado para sus necesidades.

Yelp es reconocido por la calidad de sus reseñas y el control de la veracidad de las mismas.

En cuanto a Google, siendo el motor de búsqueda más poderoso del mundo, no sorprende que domine no solamente en la efectividad de sus funciones de búsqueda, sino también en su plataforma de reviews para empresas. Google es uno de los principales sitios de reseñas para empresas de todo tipo y atrae un tráfico mensual de casi 160 millones de usuarios al mes. Además, está conectado con Google Maps.

En este contexto, en el presente proyecto se realiza un análisis exhaustivo de los Reviews de ambas plataformas, focalizadas en el rubro nightlife, hoteles y bares en los estados más populares de Estados Unidos.

Los resultados derivados del análisis serán un input fundamental en la toma de decisiones de negocio, tanto para empresarios ya activos como para nuevos emprendedores.

Necesidad del Cliente

Se solicita diseñar una interfaz que permita al usuario ingresar el nombre de un establecimiento, ya sea un hotel, nightlife o bar y obtener la siguiente información relevante:

1. Numero promedio de estrellas en su calificación.
2. Palabras más frecuentes en sus puntuaciones positivas. (el atributo que más elogian)

3. Palabras más frecuentes en sus puntuaciones negativas. (de lo que más se quejan)
4. Crecimiento de percepción positiva (año a año)
5. Proporción de reseñas positivas.
6. Análisis de Sentimientos basadas en las reviews (si por lo general es positivo, negativo o neutro)
7. Dirección
8. Foto
9. Tres sitios parecidos al emprendimiento ingresado (con cualidades parecidas en la misma zona) ***(sistema de recomendación)***

Si bien esta interfaz está pensada principalmente para los empresarios dueños de los respectivos negocios, también podría ser utilizada por usuarios comunes que quisieran información bien detallada del emprendimiento a analizar.

Alcance del proyecto

El proyecto estará enfocado en brindar información a los empresarios sobre diferentes nichos de su interés: hoteles, nightlife y bars.

El informe se centrará en los estados más visitados de los Estados Unidos.

- California
- Nueva York
- Florida
- Nevada
- Illinois

La decisión anterior se basa en un análisis realizado mediante diferentes páginas web: **Kayak**, **Expedia** y TripAdvisor y páginas de datos estadísticos como **Statista.**

Equipo de trabajo y Roles

Por tratarse de una metodología de trabajo realizada en etapas, los roles se irán intercambiando a lo largo del proyecto, según lo que se requiere en cada sprint. Sin embargo, en la siguiente tabla se presentan de manera general las principales tareas adoptadas por cada integrante del equipo.

Tabla 1. Roles de Equipo.

Miguel Ángel Ramos Cañari	Data Engineer
Juan Manual Valderrama Santos	Data Analitycs
Andrés Olascoaga	Data Analitycs
César Ricardo Moreno Bedoya	Data Science
Camila Criado	Data Science

Metodología de Trabajo

Se adopta la metodología de trabajo Scrum. Esta permitirá abordar el proyecto en un entorno dinámico y cambiante de un modo flexible. Se basa en entregas parciales y regulares del producto final en base a los requerimientos del cliente.

INFORME DE AVANCE – SEMANA 1

Se realizarán reuniones de seguimiento (llamadas dailys) los días Lunes, Miércoles y Viernes, en conjunto con el Product Owner y el Scrum Master.

En estas reuniones de 30 minutos de duración se revisarán las tareas realizadas hasta el momento, se establecerán cuáles se realizarán a continuación y se darán a conocer los problemas o dificultades que se estén teniendo para poder encontrar una solución y avanzar en el proceso hacia el producto final.

El proyecto se ha divido en 3 etapas o sprints, cada uno con una duración aproximada de 2 semanas. En cada una de ellas se presenta un informe de avance con su correspondiente producto (entregable). Al final de cada sprint se revisan los errores o cambios que el cliente (producto owner) solicite para incorporar estos requerimientos en la siguiente instancia.

Tecnologías a utilizar:

A continuación se presentan las tecnologías principales elegidas para elaborar el producto.

Cabe destacar que es posible que se incorporen nuevas tecnologías a medida que se avance en el desarrollo del proyecto.

Análisis Exploratorio Preliminar – Calidad del Dato

En esta sección se realiza un primer análisis de los datos provistos por ***YELP y Google Bussiness***. El objetivo de esta instancia es entender el tipo de datos con los que estaremos trabajando, así también como conocer su calidad y los potenciales procesos de transformación a los que deberá ser sometida la data. En este sentido, no se busca realizar en esta instancia un Análisis Exploratorio Profundo, sino más bien tener un primer panorama de los dataset con su descripción general. .

En cuanto a los datasets provistos por YELP, se identificaron, por un lado datasets orientados al Negocio (Data Bussiness, Data Checkins, Data Tips) y por el otro lado datasets orientados al usuario (Data User, Data Review)

INFORME DE AVANCE – SEMANA 1

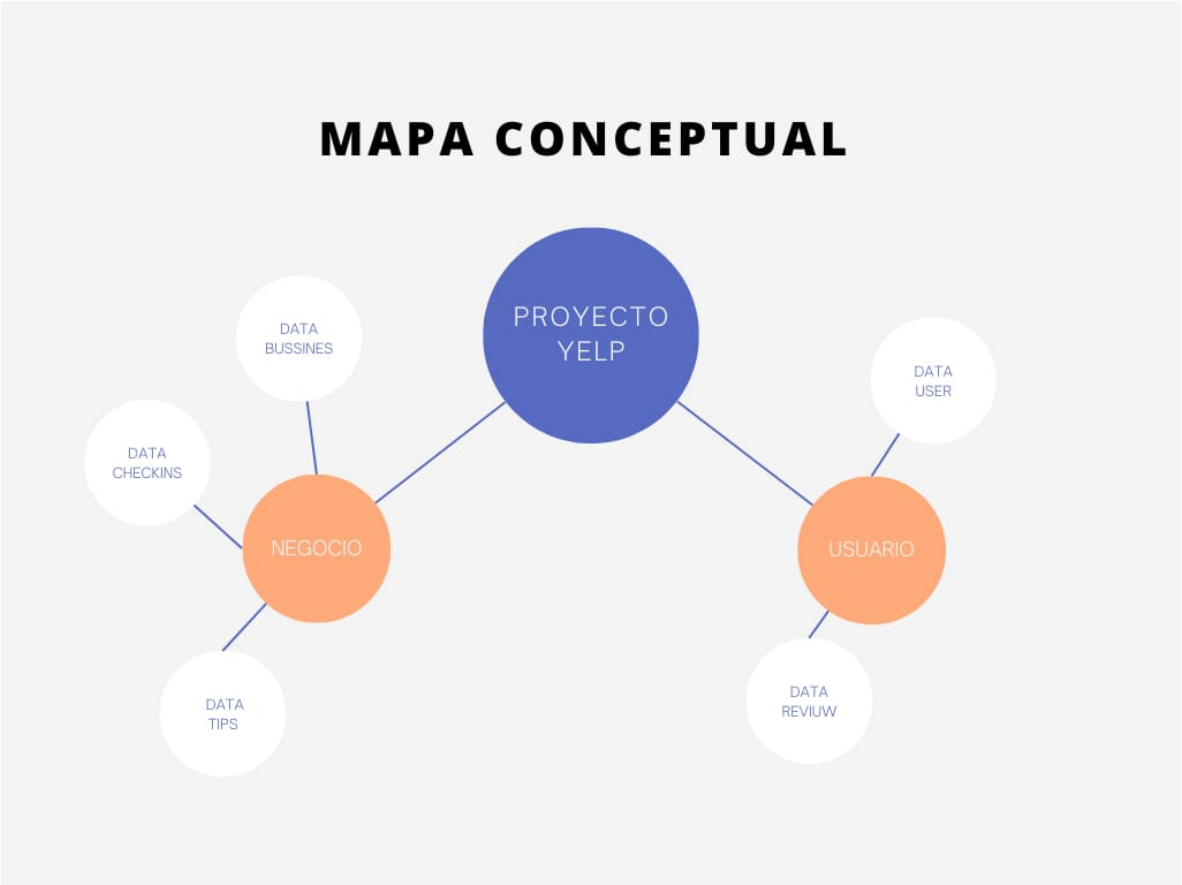


Figura 1. Mapa conceptual de Proyecto YELP.

El **archivo Bussiness** se compone de los siguientes features:

business_id: Un identificador único para cada negocio en la base de datos.

name: El nombre del negocio.

address: La dirección del negocio.

city: La ciudad donde se encuentra el negocio.

state: El estado o provincia donde se encuentra el negocio.

postal_code: El código postal del área donde se encuentra el negocio.

latitude: La latitud geográfica del negocio.

longitude: La longitud geográfica del negocio.

stars: La calificación promedio del negocio en Yelp, en una escala de 1 a 5 estrellas.

review_count: El número total de reseñas que ha recibido el negocio en Yelp.

is_open: Un indicador que muestra si el negocio está abierto o no (1 para abierto, 0 para cerrado).

attributes: Un diccionario o conjunto de atributos específicos del negocio. En el ejemplo proporcionado, un atributo es "ByAppointmentOnly", que indica si el negocio requiere cita previa.

categories: Las categorías o etiquetas que describen el tipo de negocio, separadas por comas.

hours: Un conjunto de horas de operación para el negocio. Puede ser un diccionario o estructura similar, que indica los horarios de apertura y cierre para diferentes días de la semana.

Archivo Yelp Checkins:

business_id: Es un identificador único para cada negocio en Yelp.

date: Representa las fechas en las que se registraron los check-ins para el negocio. Puede haber múltiples fechas separadas por comas para un mismo negocio, lo que sugiere múltiples check-ins en diferentes momentos.

Archivo YELP Academic REVIEW

review_id":cadena " KU_O5udG6zpxOg-VcAEodg "

"id_usuario":cadena " mh_-eMZ6K5RLWhZyISBhwA "

"Identificación del negocio":cadena " XQfwVwDr-v0ZS3_CbbE5Xw "

"estrellas":entero 3

"útil":entero 0

"divertido":entero 0

"Frío":entero 0

"texto":string " Si decide comer aquí, tenga en cuenta que tomará aproximadamente 2 horas desde el principio hasta el final. Lo hemos probado varias veces, ¡porque quiero que me guste! He estado en otros lugares en NJ y nunca Tuve una mala experiencia. La comida es buena, pero tarda mucho en salir. Los camareros son muy jóvenes, pero generalmente agradables. Hemos tenido demasiadas experiencias en las que pasamos demasiado tiempo esperando. Por lo general, optamos por otro comedor o restaurante los fines de semana, para hacerlo más rápido " .
"fecha":cadena " 2018-07-07 22:09:11 "

Archivo Tips 4

user_id: Es un identificador único para cada usuario en Yelp que ha dejado un consejo (tip) en un negocio.

business_id: Es un identificador único para el negocio en el que se dejó el consejo.

text: Es el contenido del consejo que el usuario dejó para el negocio. Es un comentario o sugerencia que el usuario comparte.

date: Es la fecha y hora en la que se dejó el consejo.

compliment_count: Es la cantidad de cumplidos (compliments) que ha recibido el consejo por parte de otros usuarios. Los usuarios pueden dar cumplidos a los consejos que consideran útiles o interesantes.

Archivo Users 5

"user_id": Es un identificador único para el usuario. Es una cadena de texto que identifica al usuario en Yelp.

"nombre": Es el nombre del usuario. Es una cadena de texto que representa el nombre del usuario en Yelp.

"review_count": Es la cantidad total de reseñas que el usuario ha escrito en Yelp. Es un número entero que indica cuántas reseñas ha dejado el usuario.

"yelping_since": Es la fecha y hora en la que el usuario se unió a Yelp. Es una cadena de texto que sigue el formato "YYYY-MM-DD HH:MM:SS" para indicar el año, mes, día, hora, minuto y segundo en que el usuario se unió.

"útil": Representa la cantidad total de votos "útiles" que el usuario ha recibido en sus reseñas por parte de otros usuarios. Es un número entero que indica cuántos usuarios han encontrado útiles las reseñas del usuario.

"gracioso": Representa la cantidad total de votos "graciosos" que el usuario ha recibido en sus reseñas por parte de otros usuarios. Es un número entero que indica cuántos usuarios han encontrado graciosas las reseñas del usuario.

"fresco": Representa la cantidad total de votos "frescos" que el usuario ha recibido en sus reseñas por parte de otros usuarios. Es un número entero que indica cuántos usuarios han encontrado frescas las reseñas del usuario.

"élite": Indica si el usuario pertenece al programa "élite" de Yelp en un año específico. Es una cadena de texto que indica el año en el que el usuario se convirtió en miembro élite. Si no es miembro élite, esta columna puede ser nula.

"amigos": Indica la cantidad de amigos que el usuario tiene en Yelp. Puede estar en blanco (nulo) si el usuario no ha proporcionado esta información.

En cuanto a los datasets de Google Reviews, se presenta la siguiente información:

INFORME DE AVANCE – SEMANA 1

Overview

Overview | Alerts 6 | Reproduction

Dataset statistics

Number of variables	8
Number of observations	150000
Missing cells	328671
Missing cells (%)	27.4%
Duplicate rows	3750
Duplicate rows (%)	2.5%
Total size in memory	63.0 MiB
Average record size in memory	440.7 B

Variable types

Numeric	2
Text	3
Categorical	1
Unsupported	2

Tal como se puede ver en la tabla anterior, el dataset está representado por 8 variables, dentro de las cuales 2 son numéricas, 3 son de texto, 1 es categórica y dos no se han podido descargar inicialmente.
Se han encontrado un 27,4% de celdas vacías.

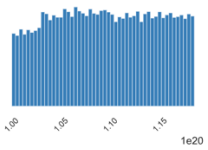
Variables

Select Columns

user_id
Real number (ℝ)

Distinct	74560
Distinct (%)	49.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1.0942786×10^{20}

Minimum	1.0000037×10^{20}
Maximum	1.1844671×10^{20}
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	1.1 MiB



More details

La variable de **user_id** contiene 74560 registros únicos. No se encontraron registros vacíos en esta columna.

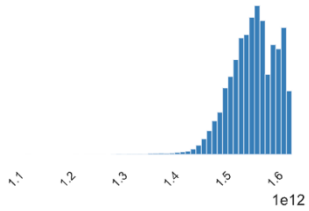
INFORME DE AVANCE – SEMANA 1

time

Real number (ℝ)

Distinct	146062
Distinct (%)	97.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1.5562613×10^{12}

Minimum	1.1183616×10^{12}
Maximum	1.6310094×10^{12}
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	1.1 MiB



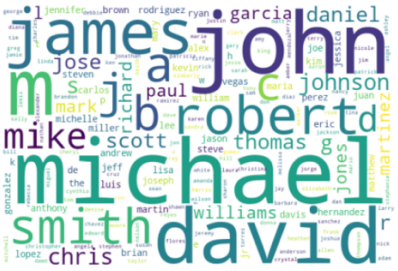
More details

La variable **Time** se presenta en número Real. No se han encontrado registros vacíos en esta columna.

name

Text

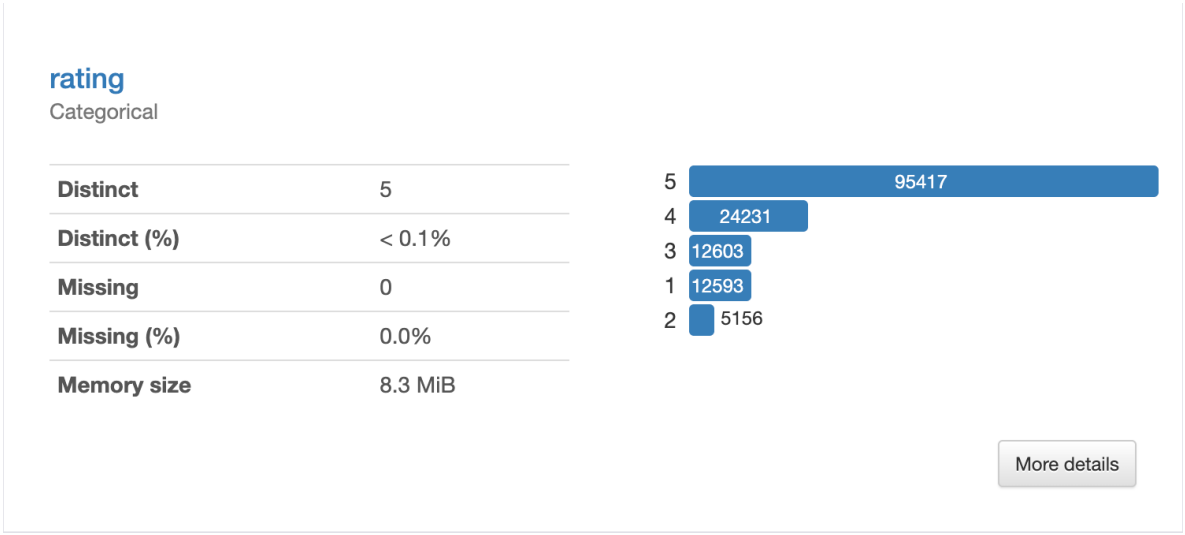
Distinct	71607
Distinct (%)	47.7%
Missing	0
Missing (%)	0.0%
Memory size	10.1 MiB



More details

La variable **name** contiene 71607 registros únicos y no se han registrado filas vacías.

INFORME DE AVANCE – SEMANA 1



La variable **rating** es categórica y tiene 5 valores únicos, del 1 al 5.

Esta columna es de tipo Texto y almacena descripciones sobre el sitio visitado.

gmap_id

Text

Distinct	5116
Distinct (%)	3.4%
Missing	0
Missing (%)	0.0%
Memory size	13.4 MiB

More details

Por último, esta columna se refiere al gmap_id.

A continuación se provee una muestra del dataset analizado.

Sample

First rows | Last rows

	user_id	name	time	rating	text
0	1.028432e+20	CDL CO	1622740504476	5	No frills smaller coffee counter in the c
1	1.149544e+20	So_ Blind	1573159413820	1	Save yourself the frustration and just i
2	1.165234e+20	Greg Hansen	1515251163362	3	I passed two Starbucks in order to ge
3	1.033041e+20	Brett Powell	1506206762095	3	It does the job and is my preference o
4	1.141624e+20	M Aris	1468298771892	3	None
5	1.150063e+20	Javier Banuelos	1462084483429	4	None
6	1.003421e+20	Terri Hughes	1530435882286	3	None
7	1.123530e+20	juiceten711	1550539981604	5	None
8	1.089885e+20	Rori Gilliam	1622512145482	5	Very helpful staff, great prices
9	1.071824e+20	Glenn Mundo	1509485965434	5	Let's see, I came here for over 3 or 4 y

INFORME DE AVANCE – SEMANA 1

Como conclusión del análisis preliminar realizado, se destaca que se trata de una gran cantidad de data, por lo que será necesario tratarla como BIG DATA.
Para esto, se utilizarán herramientas dentro de Google Cloud de modo de trabajar en un entorno que soporte la gran cantidad de datos.

KPIs:

En base a la propuesta y al análisis preliminar realizado, se proponen los siguientes KPIs:

1. Aumentar el porcentaje de las calificaciones positivas respecto al mes anterior.
2. Disminuir el porcentaje de las calificaciones negativas respecto al mes anterior
3. Mejorar en un 10% el tiempo de respuestas a las reviews negativas mensualmente.
4. Mejorar en un 10 % el número de respuestas al total de las reviews, mensualmente.
5. 100% en Proporción de reseñas positivas: Número de reseñas positivas / total de reseñas * 100
6. Crecimiento de percepción positiva: (Promedio actual - promedio del año anterior) / promedio del año anterior * 100

INFORME DE AVANCE – SEMANA 1

Planificación de Trabajo – Estimación de Esfuerzos (Sprint 1)

La planificación del cronograma de trabajo se realizará a través del Diagrama de Gannt. Al momento de la presentación del presente informe, nos encontramos en el día 9 – Entrega Informe de Avance I, terminando el Sprint 1 y comenzando el Sprint 2.

Tabla 2. Diagrama de Gannt

Tareas	Sprint 1								Sprint 2									Sprint 3								
	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.	Ju.	Vi.	Lu.	Ma.	Mi.
	31	1	2	3	4	7	8	9	10	11	14	15	16	17	18	21	22	23	24	25	28	29	30	31	1	4	5	6
PUESTA EN MARCHA																												
Kick OFF																												
Asignación de Roles																												
Alcance y establecimiento de los objetivos																												
Creación Repositorio																												
Definición KPIs																												
EDA Preliminar - Analisis Datos																												
Definición Tecnología a utilizar																												
Elaboración del Informe																												
Implementación del Stack Propuesto																												
Reunión Revisión Pre - Entrega																												
Correciones Finales																												
Entrega Informe de Avance 1																												
DATA ENGINEERING																												
Arquitectura Propuesta y Diagrama																												
Diccionario de Datos																												
Creación del DW																												
Automatizar el DW																												
Carga Incremental de Datos																												
Reunión Revisión Pre - Entrega																												
Correciones Finales																												
Entrega Informe de Avance 2																												
ANALYTICS Y MACHINE LEARNING																												
Diseño de Reportes																												
DashBoards																												
Modelo Machine Learning																												
Modelo Machine Learning en Producción																												
Elaboración del Informe																												
Reunión Revisión Pre - Entrega																												
Correciones Finales																												
Entrega Informe Final																												