Semana N°3

DATA ANALYTICS - MACHINE LEARNING

Yelp & Google Maps
Reseñas y Recomendaciones

CRONOGRAMA

1. PROPUESTA DE PROYECTO
2. DATA ENGINEERING
3. DATA ANALYTICS - MACHINE LEARNING
4. PRESENTACIÓN FINAL

DIAGRAMA DE GANT

TERCERA SEMANA

Tareas	Lunes	Martes	Miércoles	Jueves	Viérnes
Diseño del Dashboard					
Visualización de KPI´s					
Creación del modelo de Machine Learning					
Documentación					

Critico

No Critico

Entrega

DATA ANALYTICS

TAREAS

1. ANALISIS EXPLORATORIO DE DATOS (EDA)
   - BIG QUERY
2. ELABORACIÓN DEL DASHBOARD
   - LOOKER STUDIO

CONTEXTO

MAPA DE E.E.U.U

- Pacifico
- Rocosas
- Suroeste
- No Contiguo
- Noreste
- Sureste
- Medioste

Reseñas con imágenes (Google Maps)

KPI: Aumento del 20% de reseñas con imágenes, por estado, anualmente.

% R.c.Im.
4,12
24.1%

Año
2.000 2.021

Año | Región | estado

Estados con mayor porcentaje de reseñas con imágenes

	Estados	Regiones	% R.c.Im.	% Δ
1.	hawaii	no_contiguous	9,45	26.7%
2.	district_of_columbia	southeast	7,22	42.6%
3.	nevada	rocky_mountains	5,38	20.8%
4.	washington	pacific	5,26	28.0%
5.	california	pacific	5,26	35.9%

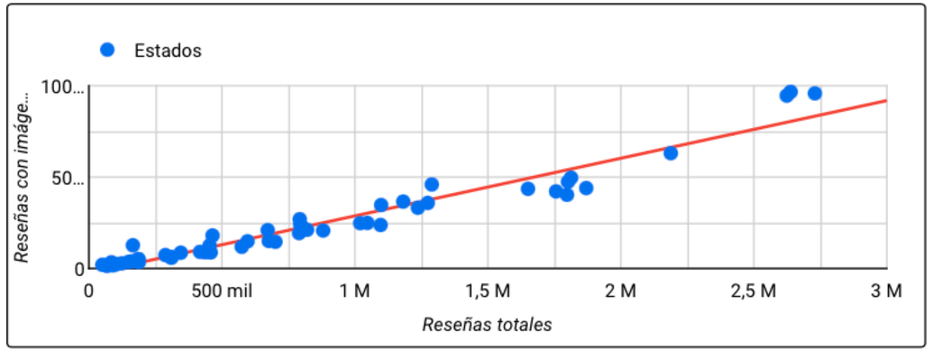
1 - 51 / 51

Estados con menor porcentaje de reseñas con imágenes

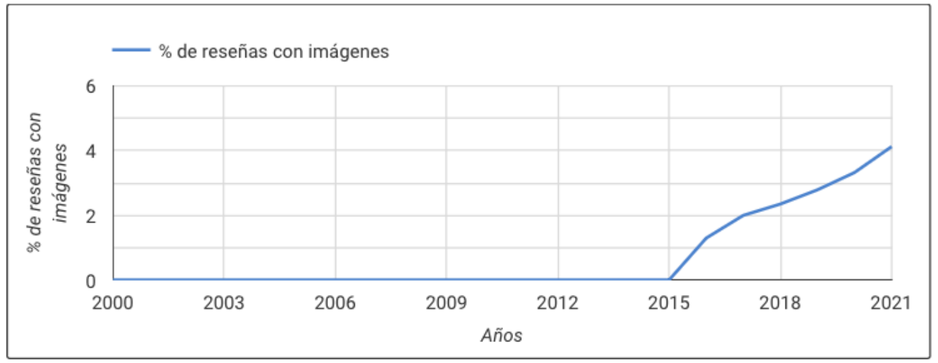
	Estados	Regiones	% R.c.Im.	% Δ
1.	wyoming	rocky_mountains	2,64	3.7%
2.	west_virginia	southeast	2,74	8.9%
3.	mississippi	southeast	2,82	23.4%
4.	arkansas	southeast	2,89	20.2%
5.	iowa	midwest	2,91	23.9%

1 - 51 / 51

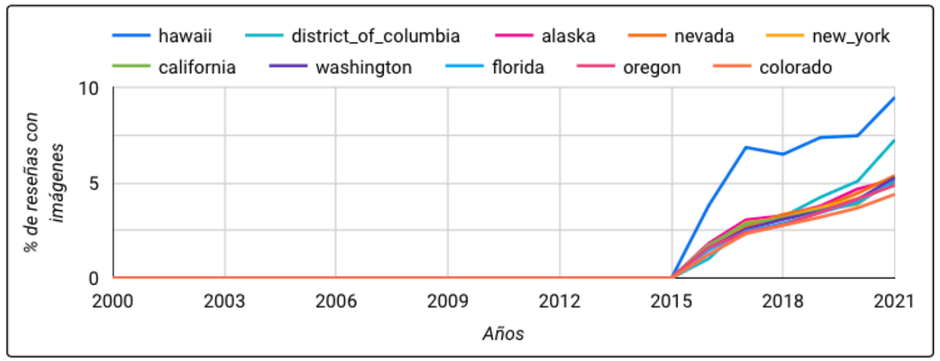
Reseñas con imágenes respecto a reseñas totales



Evolución de las reseñas con imágenes



Evolución de las reseñas con imágenes porcentualmente



Reseñas con Respuestas (Google Maps)

KPI: Aumento del 10% de reseñas con respuesta. por parte del local destinatario, por estado, anualmente.

% R.c.Resp.
15,9
5.2%

Año
2.000 2.021

Año | Región | estado

Estados con mayor porcentaje de reseñas con respuesta

	Estados	Regiones	% R.c.Resp.	% Δ
1.	colorado	rocky_mountains	20,9	4.8%
2.	utah	rocky_mountains	19,95	3.8%
3.	idaho	rocky_mountains	19,79	4.1%
4.	arizona	southwest	19,14	1.2%
5.	texas	southwest	17,95	2.4%

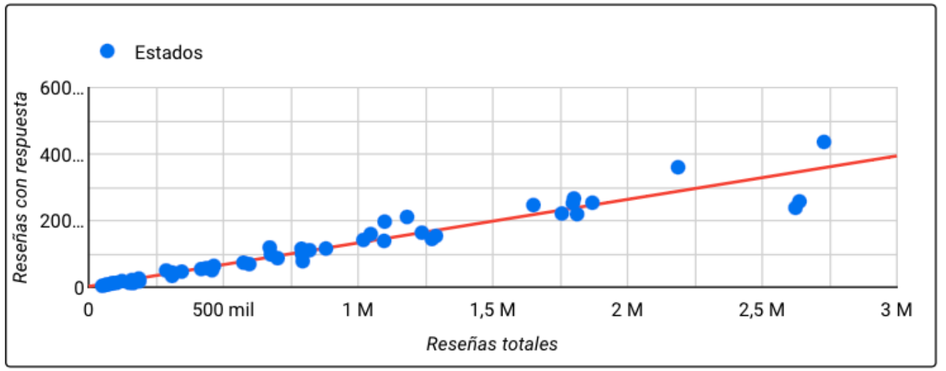
1 - 51 / 51

Estados con menor porcentaje de reseñas con respuesta

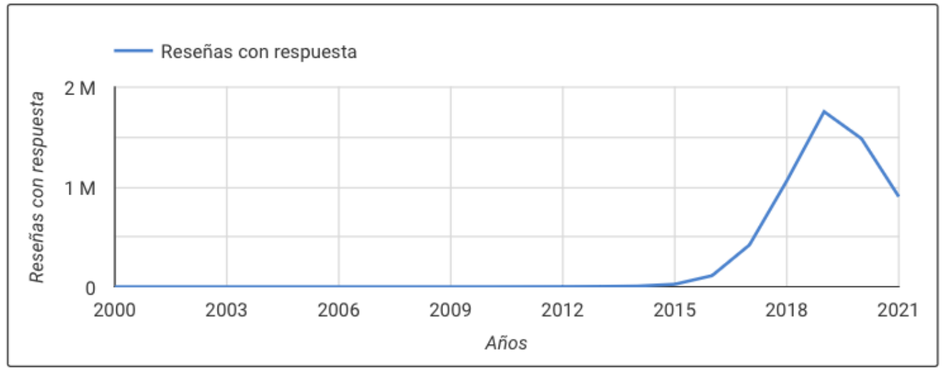
	Estados	Regiones	%R.c.Resp.	% Δ
1.	district_of_columbia	southeast	9,77	-5.5%
2.	hawaii	no_contiguous	10,37	11.4%
3.	rhode_island	northeast	11,11	9.4%
4.	west_virginia	southeast	11,29	3.8%
5.	california	pacific	12,06	8.1%

1 - 51 / 51

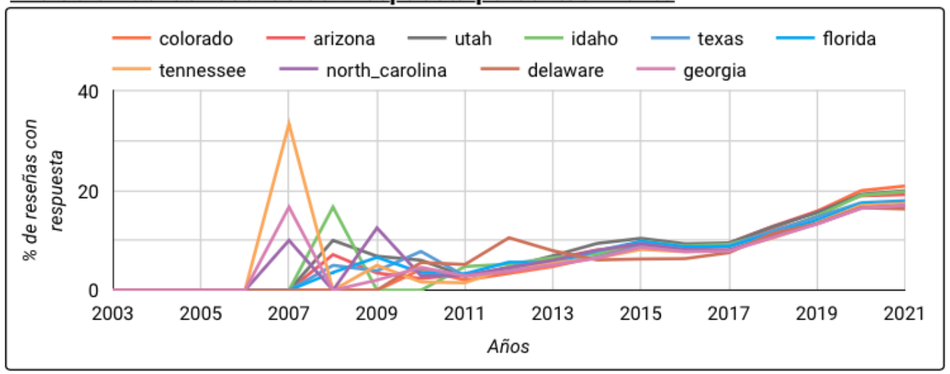
Reseñas con respuesta respecto a reseñas totales



Evolución de las reseñas con respuesta



Evolución de las reseñas con respuesta porcentualmente



Reseñas con texto (Google Maps)

KPI: Aumento del 10% de las reseñas con texto, por estado, anualmente.

% R.c.Text.
37,38
2.8%

Año
2.000 2.021

Año | Región | estado

Estados con mayor porcentaje de reseñas con texto

	Estados	Regiones	% R.c.Text.	% Δ
1.	utah	rocky_mountains	39,45	2.6%
2.	colorado	rocky_mountains	38,93	1.7%
3.	oregon	pacific	38,79	2.7%
4.	arizona	southwest	38,71	1.5%
5.	nevada	rocky_mountains	38,61	2.5%

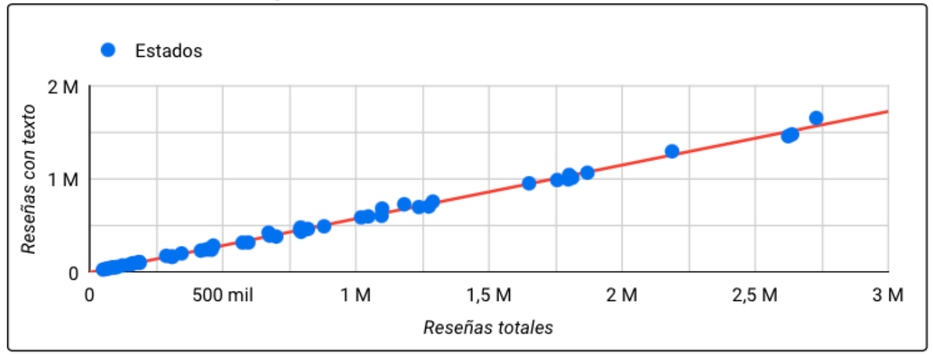
1 - 51 / 51

Estados con menor porcentaje de reseñas con texto

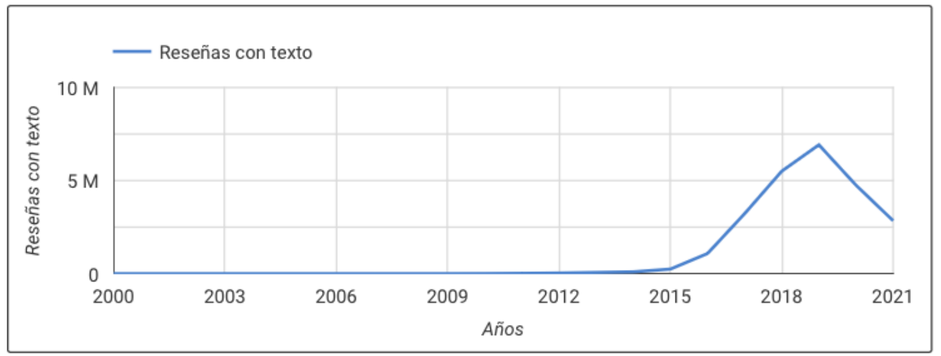
	Estados	Regiones	% R.c.Text.	% Δ
1.	mississippi	southeast	35,25	3.4%
2.	west_virginia	southeast	35,42	2.0%
3.	louisiana	southeast	35,51	3.6%
4.	iowa	midwest	35,57	3.5%
5.	alabama	southeast	35,57	1.3%

1 - 51 / 51

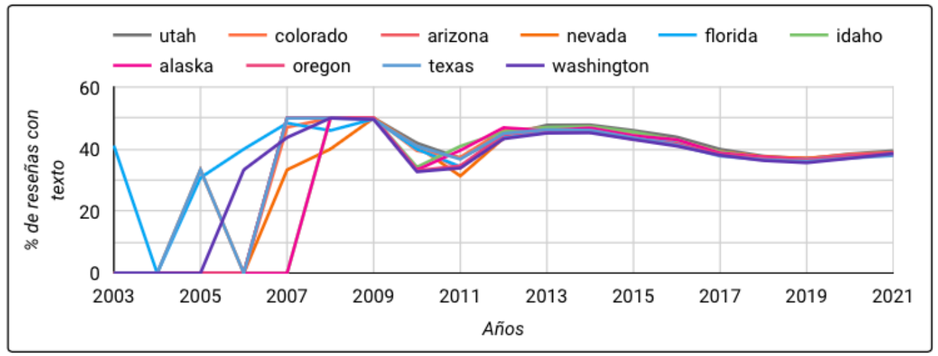
Reseñas con texto respecto a reseñas totales



Evolución de las reseñas con texto



Evolución de las reseñas con texto porcentualmente



KPI: rating mayores a 4.32 por región, anualmente

Año
1.990 2.021

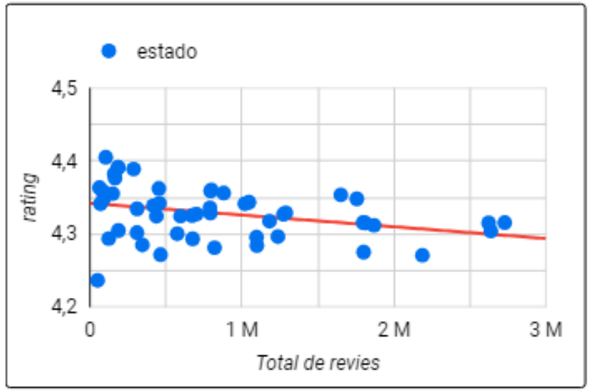
Año

Región

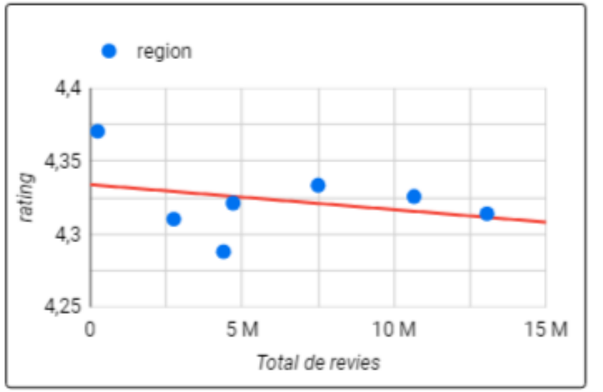
category

Reviews con rating por año

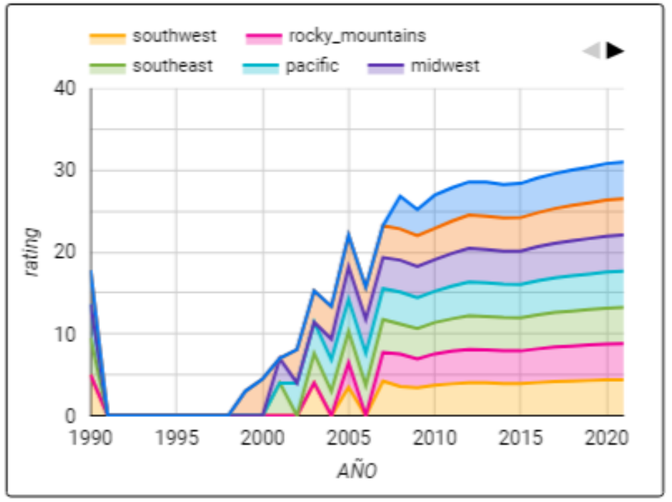
Relación del rating y el Total de reviews por estado



Relación del rating y el Total de reviews por Región



Promedio de rating por región anualmente



	estado	region	rating
1.	hawaii	no_contiguous	4,38
2.	district_of_c...	southeast	4,24
3.	alaska	no_contiguous	4,36
4.	nevada	rocky_mountains	4,27
5.	new_york	northeast	4,3
6.	california	pacific	4,32
7.	washington	pacific	4,33
8.	florida	southeast	4,32
9.	oregon	pacific	4,33

1 - 51 / 51

	region	rating
1.	no_contiguous	4,37
2.	pacific	4,32
3.	rocky_mountains	4,31
4.	northeast	4,33
5.	southwest	4,29
6.	southeast	4,31
7.	midwest	4,33

1 - 7 / 7

INFLUENCERS EN YELP

KPI:7% de las reseñas de cada ciudad por año son hechas por influencers de más de 100 seguidores (Yelp)

Usuario con mas seguidores
12.497

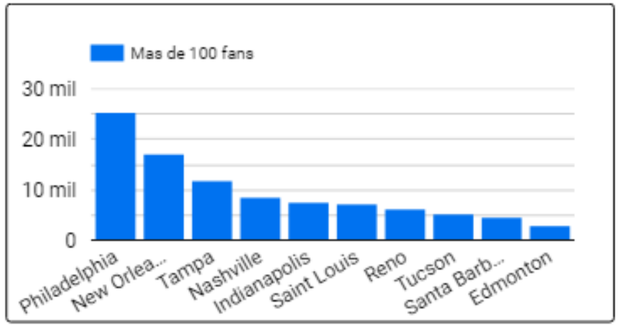
% de Review totales de influencers
6,96 %

city | Año | categories

Conteo de usuarios por seguidores

Fila	grupo	cantidad
1	Grupo 0	1618303
2	Grupo 1000-5000	125
3	Grupo 10-100	59430
4	Grupo 100-1000	5200
5	Grupo 1-10	422536
6	Grupo 5000+	3

Reviews de Influencers por ciudad



% de Reviews de influencers por ciudad

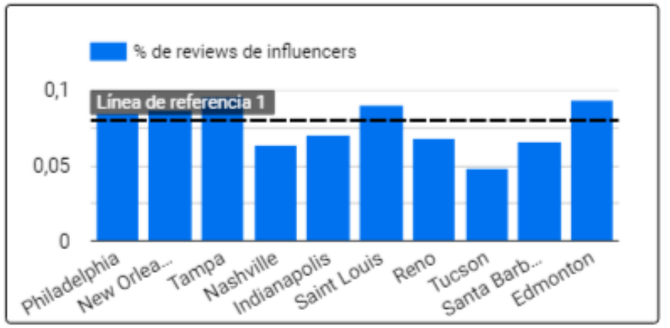
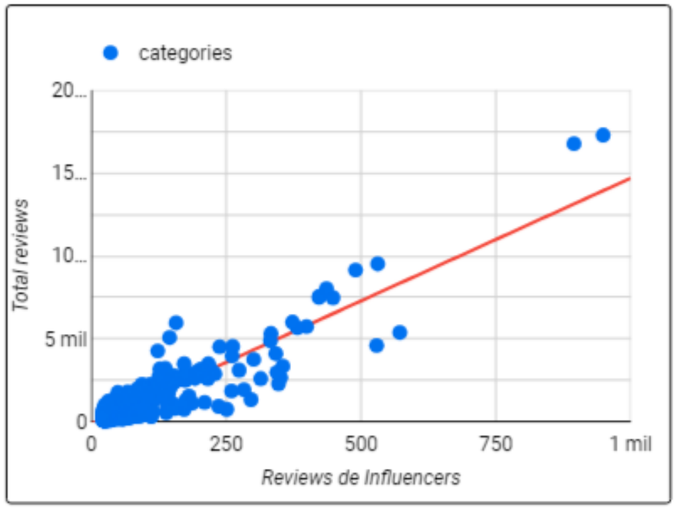


Grafico de dispersión de reviews por Influencers vs el total de reviews, por categoría



	city	Reviews de influencers	Reviews totales	% de reviews de influenc...
1.	Philadelphia	25.390	301.985	8,41 %
2.	New Orleans	17.203	198.895	8,65 %
3.	Nashville	8.723	137.175	6,36 %
4.	Tampa	11.807	122.998	9,6 %
5.	Tucson	5.370	110.426	4,86 %
6.	Indianapolis	7.723	109.117	7,08 %
7.	Reno	6.355	93.007	6,83 %
8.	Saint Louis	7.213	80.050	9,01 %
9.	Santa Barbara	4.625	70.138	6,59 %

1 - 100 / 1416

MACHINE LEARNING

ESQUEMA DE LA PRESENTACION

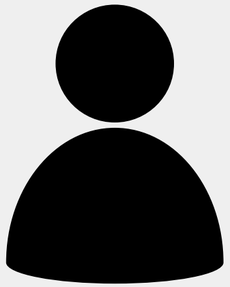
1) SISTEMA DE RECOMENDACION
   - DISEÑO
   - FUNCIONAMIENTO

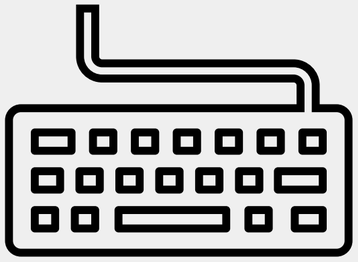
2) RESUMEN INFORMATIVO
   - ENFOQUES CONSIDERADOS
   - PROBLEMAS Y SOLUCIONES
   - ARQUITECTURA
   - FUNCIONAMIENTO

3) CIERRE
   - ESTRUCTURA FINAL

SISTEMA DE RECOMENDACION

PRIMERA PARTE
FILTRO NAIVE



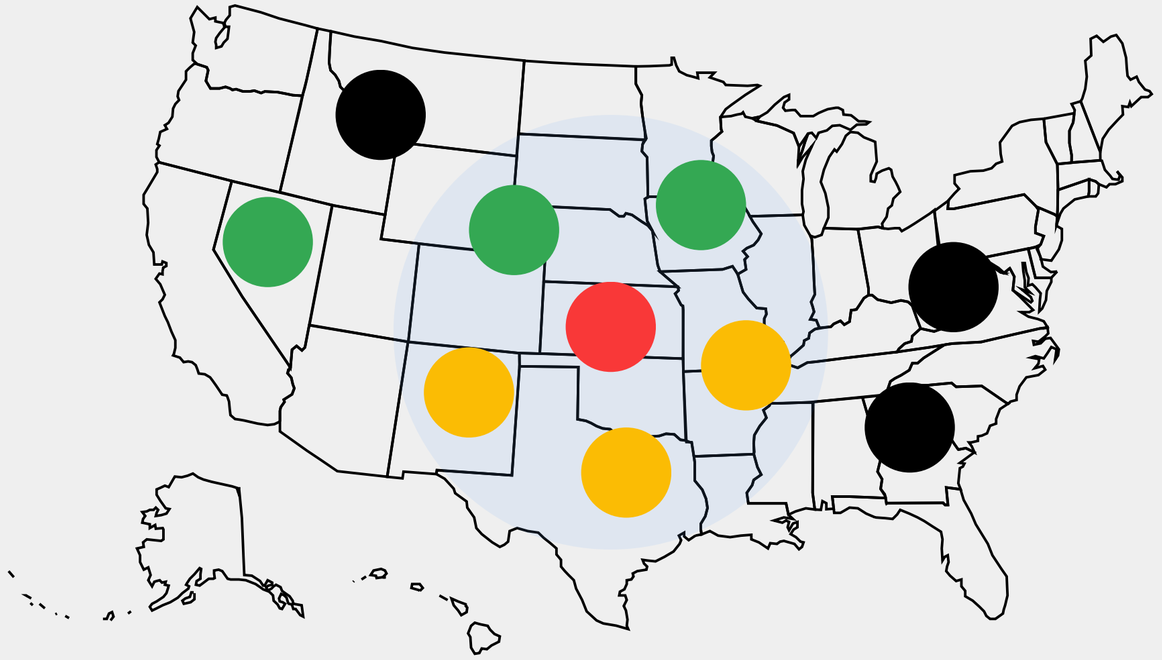


ID DEL USUARIO

LATITUD | LONGITUD

CATEGORIA

DISTANCIA



- **USUARIO**
- **INSTITUCIONES VISITADAS**
- **INSTITUCIONES NO VISITADAS**
- **INSTITUCIONES NO CONSIDERADAS**

SISTEMA DE RECOMENDACION

SEGUNDA PARTE
FILTRO COLABORATIVO

MATRIZ DE USUARIOS Y RATINGS

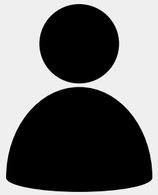
X1	1	5	3
USER	NA	NA	NA
X3	4	4	2
	X	Y	Z

USUARIOS

CONJUNTO DE IDS

VISITADOS

NO VISITADOS



RATINGS	4	3	5	1
IDS	X	Y	Z	W

EMPRESAS

X	Z

SISTEMA DE RECOMENDACION

TERCERA PARTE
FILTRO POR CONTENIDO

MATRIZ DE CONTENIDO

X	0.6	0.1	0	0.3
Z	0.4	0.3	0	0.1
X1	0	0.3	0.5	0.2
X2	0.1	0	0	0.9
X3	0	0.3	0.5	0.2

AGRUPAMOS

X	0.6	0.1	0	0.3
Z	0.4	0.3	0	0.1
XN	0.1	0.2	0.3	0.4

MATRIZ DE SIMILITUD

	XN	Z	X
XN	1	0.6	0
Z	0.6	1	0.2
X	0	0.2	1

RESUMEN INFORMATIVO

DESDE 0

Necesario cuando no existe el modelo.

Libertad y control del dataset

Requiere mucho trabajo y tiempo.

Numerosos enfoques.

Alto coste energetico, en infraestructura, y ambiental.

OPENAI

Interfaz sencilla.

Altamente flexible.

No requiere infraestructura.

OGB ocupa de memoria

Es pago y lento.

Nuevo enfoque.

Requiere conexión a internet.

HUGGING FACE

API sencilla

Enorme cantidad de opciones.

Se le puede realizar fine-tunned.

Comunidad establecida.

Modelos OpenSource.

Ocupa mucho espacio.

Requiere infraestructura

GENSIM
SPACY

API OPENAI
LANGCHAIN

TRANSFORMERS
DATASET

RESUMEN INFORMATIVO

PROBLEMAS

SOLUCIONES

OVERFLOW DE MEMORIA → ZERO-COPY MEMORY MAPPING

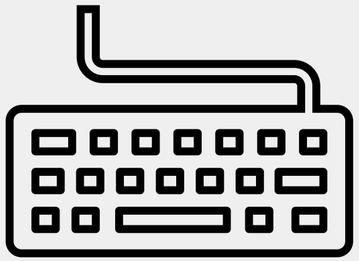
APACHE ARROW

FALTA DE GPU → GOOGLE COLAB Y KAGGLE

POCAS EPOCHS

SMALL DATASET EN EL FINE-TUNNED

RESUMEN INFORMATIVO



ID DE LA EMPRESA

RESEÑAS

.....

.....

.....

MODELOS PRE-ENTRENADOS

knkarthick/
MEETING_SUMMARY

FINE-TUNNED MODELOS

bert-base-uncased

roberta-base

facebook/bart-large

RESUMEN INFORMATIVO

NOMBRE DEL MODELO: knkarthick/MEETING_SUMMARY
MODELO BASE: facebook/bart-large-xsum
FINALIDAD: Realizar un resumen sobre las reseñas.
METRICA: ROUGE-1 on samsum 53.188
DATASET: AMI Meeting Corpus, SAMSUM Dataset, DIALOGSUM Dataset, XSUM Dataset

---

MODELO BASE: bert-base-uncased
FINALIDAD: Predice el rating de una reseña entre 1 a 5 puntos.
METRICA: ACCURACY 0.88
DATASET: GOOGLE REVIEWS

MODELO BASE: roberta-base
FINALIDAD: Clasifica las reseñas entre cool | funny | useful.
METRICA: ACCURACY 0.43
DATASET: YELP REVIEWS

MODELO BASE: facebook/bart-large
FINALIDAD: Responder a las reviews realizadas teniendo en cuenta el contexto de la misma.
METRICA: CrossEntropyLoss 2.01
DATASET: GOOGLE REVIEWS

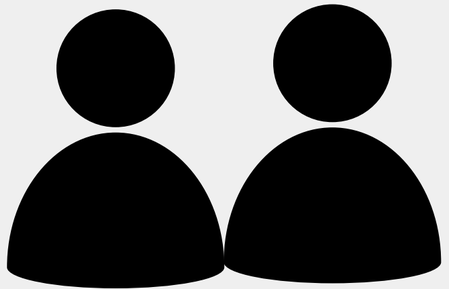
ESTRUCTURA FINAL

SISTEMA DE RECOMENDACION

RESUMEN INFORMATIVO

API

STREAMLIT



GRACIAS POR SU ATENCIÓN