Natalia Sierra
Juan Felipe Latorre
Diana Marcela Toro

Entrega 1 - Sistema de recomendación de noticias sobre clientes corporativos

• Problema que abordarán y su contexto.

El problema abordado y su contexto es la necesidad del área comercial del grupo Bancolombia de conocer mejor a sus clientes corporativos a través de la información generada por medios de comunicación locales e internacionales. Actualmente, los comerciales del banco cuentan con miles de noticias relacionadas con cada uno de sus clientes y muy poco tiempo para su lectura y análisis. Por lo tanto, se busca determinar las noticias relevantes para un sector particular de clientes, permitiendo a la fuerza comercial informarse para atender determinado sector y ser más efectivos en su labor. Para esto sugieren segmentar las noticias relacionadas con base a los sectores y determinar las características comunes entre estas, para recomendar otras noticias relevantes, actualizadas y confiables para el área comercial.

• Pregunta de negocio y alcance del proyecto.

¿Cómo puede el área comercial del grupo Bancolombia conocer mejor a sus clientes corporativos a mediante la selección información relevante, actualizada y confiable generada por medios de comunicaciones locales e internacionales?

El objetivo es desarrollar un modelo que pueda analizar y clasificar las noticias de acuerdo a su relevancia para los clientes corporativos del grupo Bancolombia. Esto implica el uso de técnicas de procesamiento de lenguaje natural y aprendizaje automático para entender el contenido de las noticias y determinar su relevancia.

El alcance del proyecto incluirá un preprocesamiento de texto, que incluye tokenización, lematización y eliminación de stopwords. Para el entrenamiento del modelo se evaluará tres opciones: La primera CountVectorizer y similitud del coseno, la segunda TF-IDFVectorizer y producto escalar con Linear Kernel y la tercera utilizando LDA (Latent Dirichlet Allocation) que es el modelado basado en temas y la evaluación de su rendimiento. También se considerará la implementación del modelo en un sistema en producción para que el equipo comercial pueda utilizarlo en su trabajo diario.

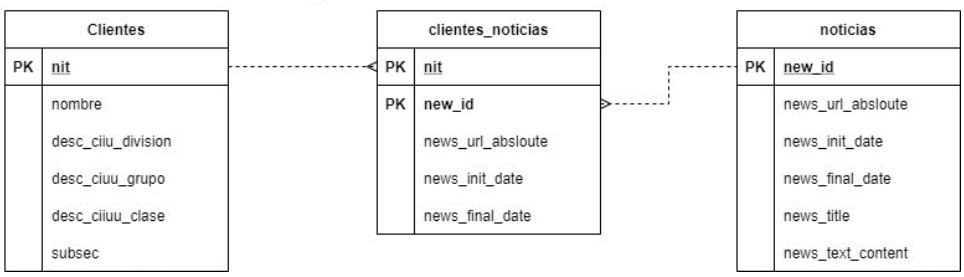
• Descripción de conjuntos de datos a emplear.

Fuente de datos: https://www.kaggle.com/datasets/juancamilodiazzapata/dataton-2022

Los datos utilizados en este estudio provienen del Dataton 2022, una competencia organizada por el Centro de Excelencia en Analítica, Inteligencia Artificial y Gobierno de Información del Grupo Bancolombia[1]. Se dividen en tres bases principales: clientes, noticias y cliente_noticias que es la relación entre ambos.

El diagrama entidad relación se muestra a continuación:

[1] Fuente: https://www.kaggle.com/datasets/juancamilodiazzapata/dataton-2022



Descripción del contenido de las variables

clientes.csv: Archivo con el listado de clientes a consultar, la descripción de su actividad económica y el subsector

- nit: Identificador único del cliente
- nombre: Nombre corporativo del cliente
- desc_ciiu_división: Descripción general de la clasificación Industrial uniforme d todas las actividades económicas
- desc_ciiu_grupo: Descripción por grupo de la clasificación Industrial uniforme d todas las actividades económicas
- desc_ciiu_clase: : Descripción por clase de la clasificación Industrial uniforme d todas las actividades económicas
- subsector: Clasificación de la actividad industrial

noticias.csv: Contenido de cada una de las noticias consultadas

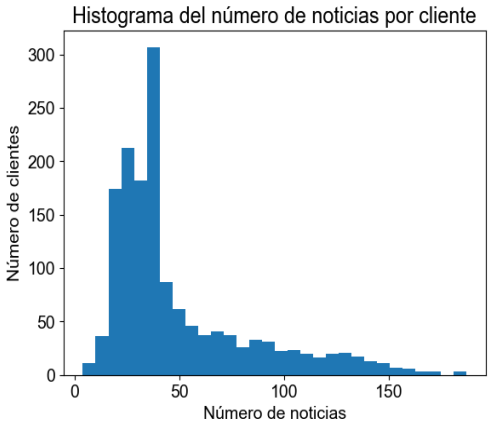
- new_id: Identificador único de noticias
- news_url_absolute: Url de la noticia encontrada
- news_init_date: Fecha mínima del intevalo de tiempo al que pertenece la noticia
- news_final_date: Fecha máxima del del intevalo de tiempo al que pertenece la noticia
- news_title: Título relacionado a la noticia
- news__text__content: Texto contenido de la noticia
-

clientes_noticias.csv: Relación entre cliente y las noticias consultadas mediante el proceso de descarga de información

- new_id: Identificador único del cliente
- news_url_absolute: url de la noticia encontrada
- news_init_date: Fecha mínima del intevalo de tiempo al que pertenece la noticia
- news_final_date: Fecha máxima del del intevalo de tiempo al que pertenece la noticia

• **Exploración de los datos.**

La base de datos clientes.csv alberga información detallada sobre 1507 clientes, incluyendo su actividad económica y el subsector al que pertenecen, todo ello clasificado según el CIIU. Cada cliente está representado por un registro único identificado por su NIT. Esta base de datos está libre de datos nulos y duplicados.De la distribución de noticias por cliente podemos encontrar que cada cliente recibe aproximadamente 49.5 noticias.

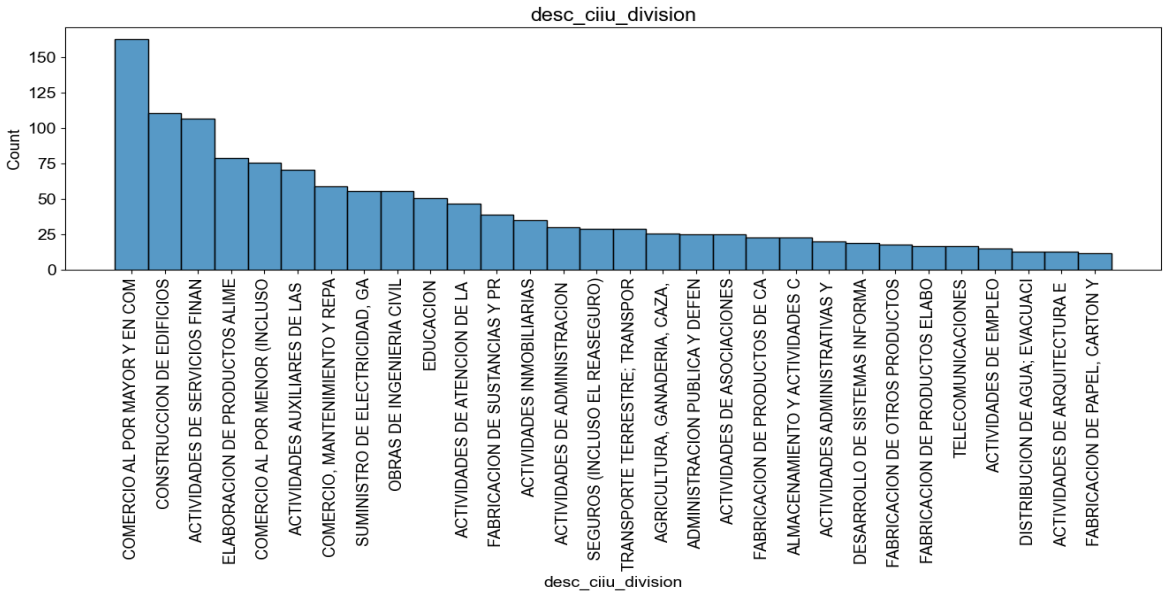


Sin embargo, hay una variabilidad considerable en esta cifra. Por un lado, tenemos clientes que reciben tan solo 4 noticias, que es el mínimo. Por otro lado, el cliente que más noticias recibe llega a las 187.

Si nos centramos en la mediana, que es el valor central de la distribución, vemos que la mitad de los clientes reciben 36 noticias o menos.

Además, si observamos los cuartiles, vemos que el 25% de los clientes reciben 27 noticias o menos y el 75% de los clientes reciben hasta 60 noticias.

Las noticias se encuentran clasificadas en 81 sectores, de acuerdo con la variable desc_ciiu_division, que es la que nos interesa para el desarrollo. La mayor concentración de noticias las encontramos en los sectores comercio al por mayor, construcción de edificios y actividades de servicios financieros. En el siguiente gráfico podemos detallar los sectores restantes.



La base de datos noticias.csv recopila información sobre 23.377 noticias consultadas. Esta información incluye detalles como URLs, fechas, títulos y contenidos. Consta de 11 variables y no presenta datos nulos ni duplicados. Importante resaltar que las noticias se encuentran concentradas en dos fechas 2022-07-15 y 2022-07-29 y la distancia entre ellas es de 15 días.

La tercera base de datos clientes_noticias.csv establece la relación entre los 1507 clientes y las 23.377 noticias consultadas, proporcionando los identificadores correspondientes y las URLs. no presenta datos nulos ni duplicados.

• **Maqueta del prototipo.**

Maqueta de entrenamiento del modelo

Esta maqueta representa el entrenamiento del modelo, en la cual se exhiben las principales actividades económicas que se incorporaron al modelo y que tienen un impacto significativo en los resultados. También se muestra la distribución de noticias por cliente. Esta distribución es crucial ya que, en caso de un reentrenamiento del modelo, me permite determinar si los datos son similares a los de este entrenamiento.

El número diario de noticias por cliente es un factor importante. Si el número de noticias aumenta drásticamente en un día, podría indicar que ha ocurrido un evento muy importante que podría afectar los resultados del modelo y limitar la variedad de las noticias.

Además, se presenta un gráfico de los temas generados por el modelo. Estos temas serán la base para comparar las noticias leídas por los clientes.

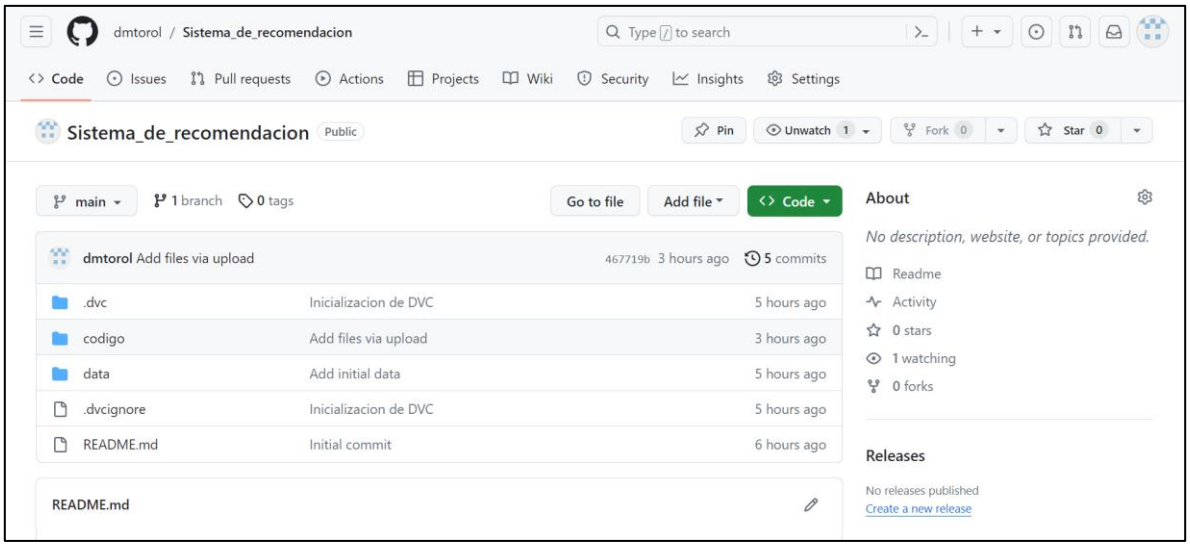
Maqueta de usuario

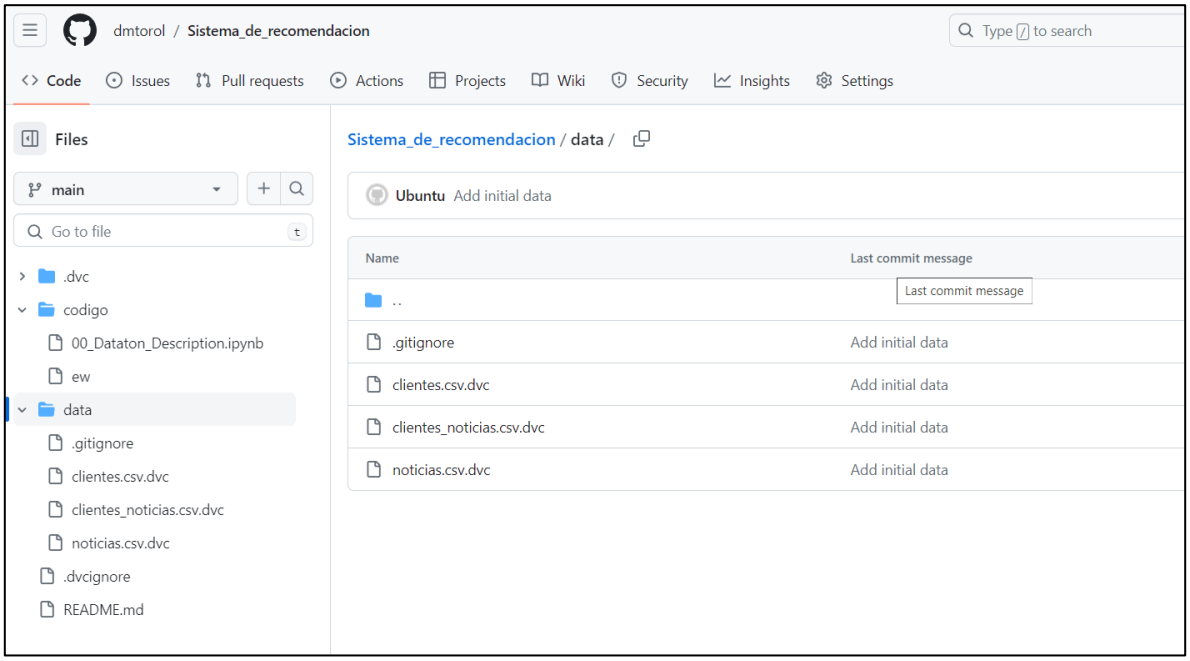
En esta maqueta, el usuario ingresa la cédula del cliente y, en la imagen siguiente, se muestran las noticias, el sector y el porcentaje de similitud de cada una de las noticias que fueron recomendadas por este modelo.

• Repositorios con todo el código (capturas y enlaces que soporten la creación y uso de repositorios).

https://github.com/dmtorol/Sistema_de_recomendacion.git

Repositorio creado en GitHub





Gestión y versiona miento de datos Data Version Control – DVC

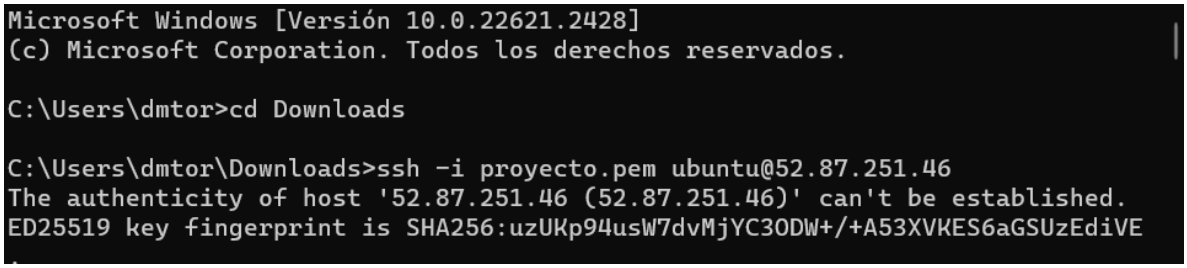
Código

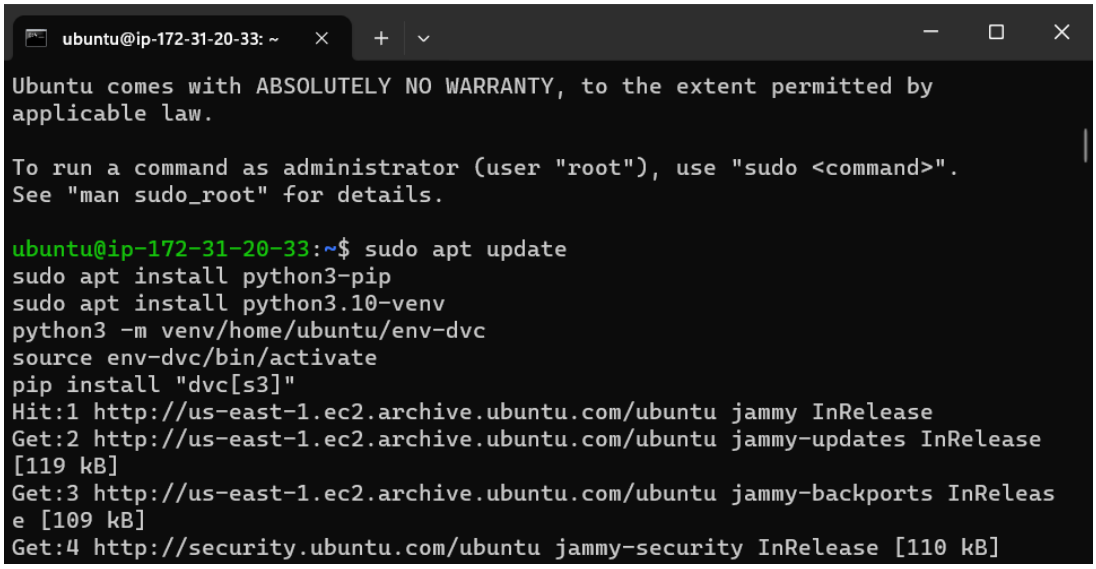
```
cd C:/Users/dmtor/Downloads/Despliegue/SistemaRec/dataton
ssh -i proyecto.pem ubuntu@52.87.251.46

sudo apt update
sudo apt install python3-pip
sudo apt install python3.10-venv
python3 -m venv /home/ubuntu/env-dvc

#####################################################
source env-dvc/bin/activate
#####################################################
pip install "dvc[s3]"

mkdir dvc-proj
cd dvc-proj
mkdir data


#Desde otra terminal
##=================================================================================
=
cd C:/Users/dmtor/Downloads/Despliegue/SistemaRec/dataton
scp -i proyecto.pem
C:/Users/dmtor/Downloads/Despliegue/SistemaRec/dataton/noticias.csv
ubuntu@52.87.251.46:/home/ubuntu/dvc-proj/data/
scp -i proyecto.pem
C:/Users/dmtor/Downloads/Despliegue/SistemaRec/dataton/clientes.csv
ubuntu@52.87.251.46:/home/ubuntu/dvc-proj/data/
scp -i proyecto.pem
C:/Users/dmtor/Downloads/Despliegue/SistemaRec/dataton/clientes_noticias.csv
ubuntu@52.87.251.46:/home/ubuntu/dvc-proj/data/
##=================================================================================
=

cd dvc-proj
git init
```

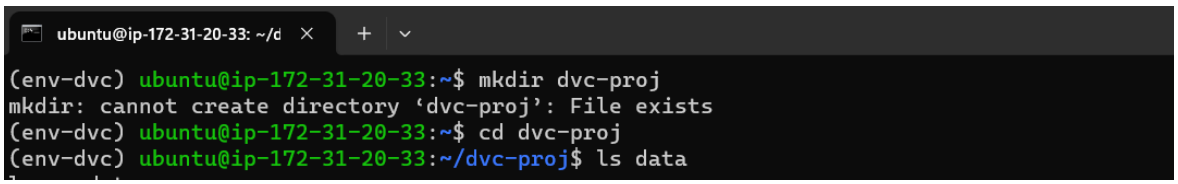
```
git branch -m main
dvc init -f
git status
git commit -m "Inicializacion de DVC"

dvc add data/noticias.csv data/clientes.csv data/clientes_noticias.csv
git add data/noticias.csv.dvc data/clientes.csv.dvc
data/clientes_noticias.csv.dvc data/.gitignore

git commit -m "Add initial data "
git remote add origin https://github.com/dmtorol/Sistema_de_recomendacion.git
git pull origin main --rebase
git push origin main
git push -u origin main
```

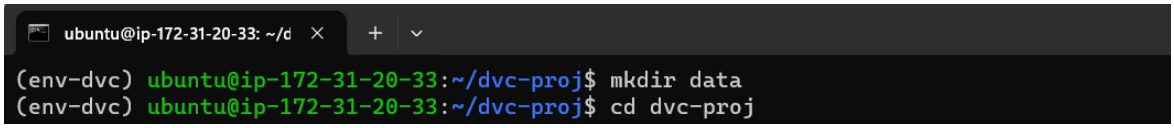
Capturas de pantalla

Conexión con la máquina, instalación de librerías y activación del ambiente y las carpetas dvc-proj y data:

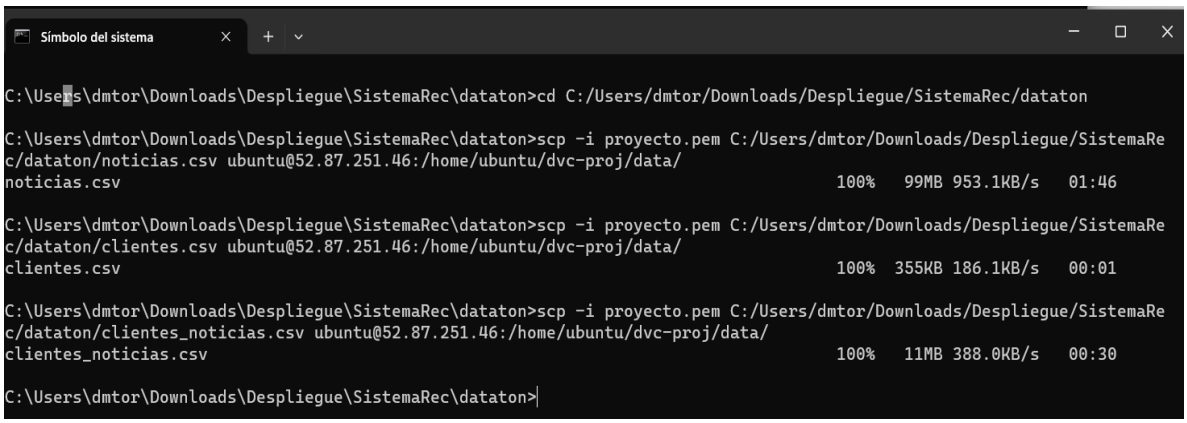








Desde otra terminal se suben los archivos locales a la carpeta dvc-proj/data creada en el ambiente:



Se inicializa git y dvc y el primer commit: