Natalia Sierra
Juan Felipe Latorre
Diana Marcela Toro

Entrega Final - Sistema de recomendación de noticias sobre clientescorporativos

• 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.

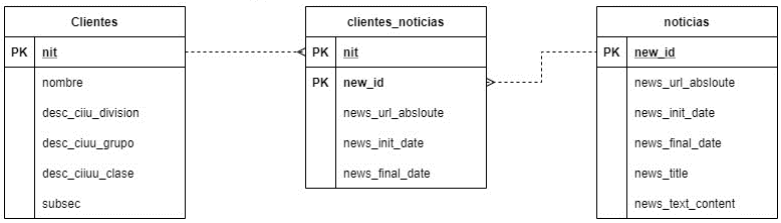
El alcance del proyecto incluirá un preprocesamiento de texto, que incluye tokenización, lematización y eliminación de stopwords. Para el entrenamiento del modelo se evaluará tres opciones: La primera CountVectorizer y similitud del coseno, la segunda TF-IDFVectorizer y producto escalar con Linear Kernel y la tercera utilizando LDA (Latent Dirichlet Allocation) que es el modelado basado en temasy 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

Modelos desarrollados y su evaluación

1. **CountVectorizer:** en este caso, usando el título de las noticias calclulamos la similitud coseno de la noticia de entrada y ordenamos las similitudes en orden descendente para obtener las noticias más similares primero. Este algoritmo nos recomienda 5 noticias, y les asigna un puntaje de "similitud". Tomamos como ejemplo la noticia con el titular "Por no contar la verdad, Hugo Aguilar quedó sin cupo en la JEP" y obtuvimos los siguientes resultados:

	news_title	subsec	sim_scores
0	9 de cada 10 adultos en Colombia contaba con a...	SEGUROS	0.289540
1	Sube de 43% a 57% porcentaje de peruanos con c...	SEGUROS	0.286897
2	Las ventajas de emplear un software de contabi...	TRANSPORTE TERRESTRE	0.256613
3	El 49% de peruanos ya recibe o realiza pagos d...	SEGUROS	0.247232
4	La penetracion de los seguros se mantuvo en 3%...	SEGUROS	0.243572

Vemos acá los subsectores de cada recomendación y los respectivos puntajes, que van desde un 24.35% hasta un 28.95%.

2. **TF-IDFVectorizer:** en esta aproximación, utilizamos este algoritmo para hacer las recomendaciones; convertimos el contenido de las noticias en recomendaciones numércias. Este algoritmo le da mayor peso a las palabras que son más relevantes para un documento en particular, en relación con una colección de otros documentos. Luego hacemos un cálculo del kernel lineal y se lo aplicamos a la matriz TF-IDF. Este kernel lineal es una función que calcula el producto escalar (similaridad) entre 2 vectores. En este caso, los vectores son las representaciones TF-IDF de las noticias. Para evaluar este algoritmo, seguimos usando el mismo ejemplo anterior, y obtuvimos los siguientes resultados

	news_title	subsec	sim_scores
0	El 60% de los valencianos consideran que exist...	SEGUROS	0.206347
1	9 de cada 10 adultos en Colombia contaba con a...	SEGUROS	0.196525
2	Por que las empresas y la academia cooperan pa...	SEGUROS	0.179647
3	Sube de 43% a 57% porcentaje de peruanos con c...	SEGUROS	0.174100
4	La educacion gratuita | Columna de Jesus Ferro...	TRANSPORTE TERRESTRE	0.162873

En este caso, los score de similitud son más bajos de lo que vimos en la aproximación anterior, ya que oscilan entre 16.28% y 20.63%.

3. **LDA:** en este caso, la recomendación es basada en temas, donde posterior a obtener el modelo, implementamos la función de recomendación en la cual obtenemos el tema al cual pertenece la noticia recibida como parámetro y el subsector de la misma. Con base en este resultado, buscamos todas las noticias asociadas a este mismo tópico; tenemos 4 grandes temas en esta aproximación.

   En este caso, la relevancia de una noticia se mide en relación con el tema principal asignado por el modelo LDA y la probabilidad asociada a este tema. Noticias con una probabilidad más alta para el tema principal se consideran más relevantes en ese tema específico.

   Para evaluar esta aproximación usamos la noticia de prueba "Javier Rodriguez Soler, nuevo presidente del Consejo Asesor del Centro de Educacion y Capacidades Financieras de BBVA" y obtuvimos los siguientes resultados:

	news_title	subsec	topic_proba
1	Ana Paula Marques (EDP): "Si tenemos una carga...	SEGUROS	0.999086
2	Economia peruana crecio 2.28% en mayo, se desa...	SEGUROS	0.998983
3	Mafia de combustibles y alianza Marti-Total	SEGUROS	0.998971
4	Volaris reactiva ruta Guanajuato-Merida; Zoho ...	SEGUROS	0.998963
5	MAPFRE gana 338 millones de euros en los seis ...	SEGUROS	0.998932

A diferencia de las aproximaciones anteriores, vemos que el score de similitud es considerablemente más alto que en nuestros modelos anteriores, este score está entre 99.89% y 99.90%.

Observaciones y conclusiones sobre los modelos

La principal limitación del Modelo #1 (CountVectorizer) es la sensibilidad al título, ya que si la información está en el cuerpo del correo y no en el título, la recomendación no va a ser tan efectiva.

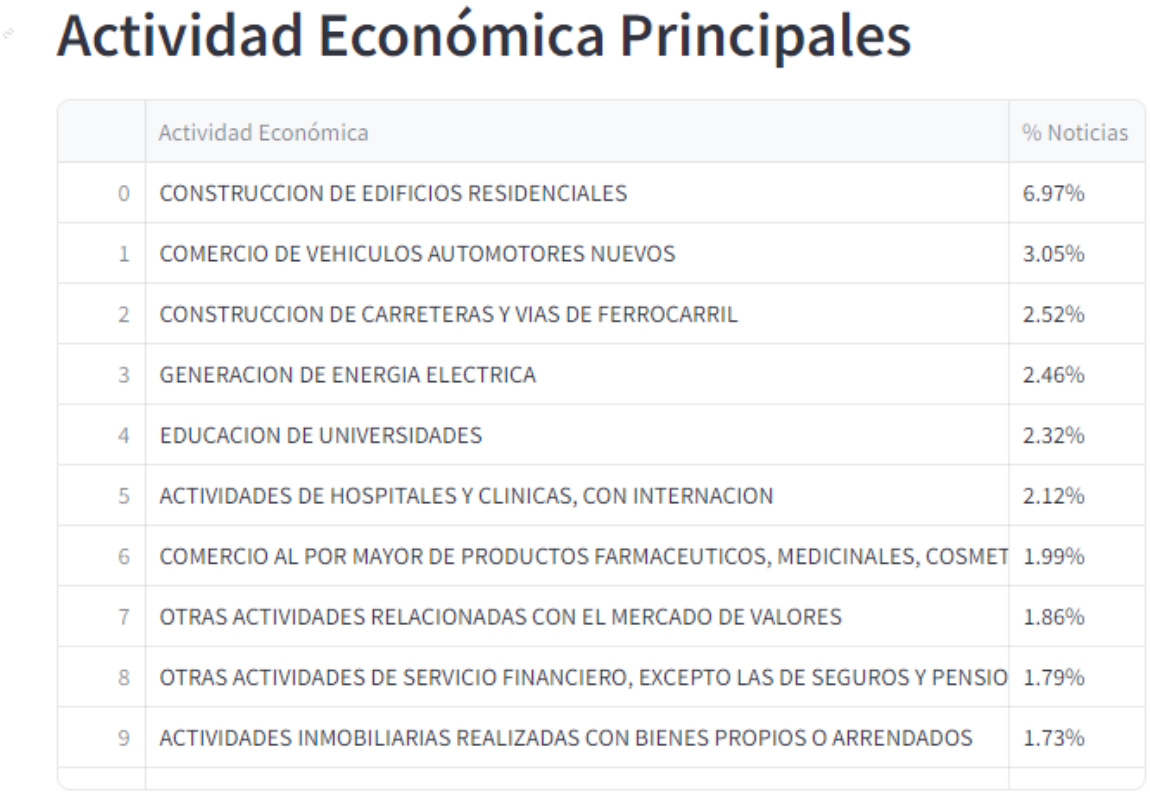
Dentro de las limitaciones que tiene del Modelo #2. (TF-IDFVectorizer), está que la calidad de las recomendaciones puede estar influenciada por el tamaño y la representatividad de los datos utilizados.

Finalmente, una limitación del Modelo #3. (LDA) es la sensibilidad a los hiperparámetros, como es el número de los temas (en nuestro caso 4).

A pesar de estas limitaciones, y dado el buen desempeño del modelo que usa LDA, tanto a nivel absoluto como relativo a los otros dos, consideramos que es el algoritmo ideal para darle solución al problema de negocio.

Descripción del tablero desarrollado y la funcionalidad que éste ofrece

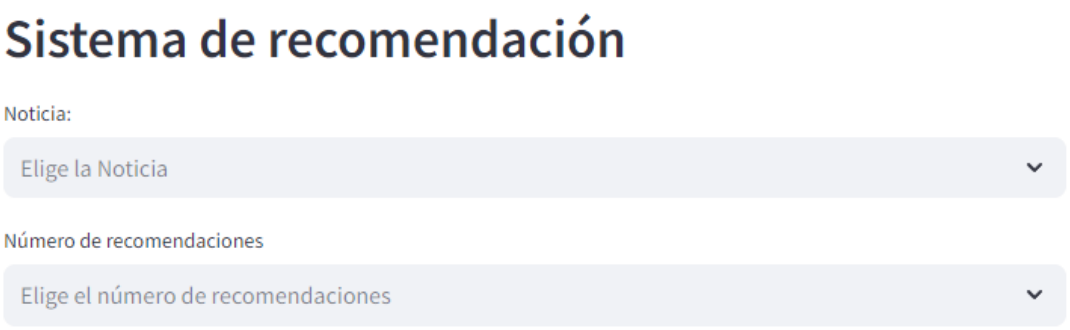
En la URL http://44.202.237.158:8502/ se desarrolla el tablero “Sistema de Recomendación de Noticias Sobre Clientes Corporativos” en donde en una primera parte se pueden ver las actividades económicas a las que pertenecen las noticias, y el porcentaje de noticias en referencia a ese tema.



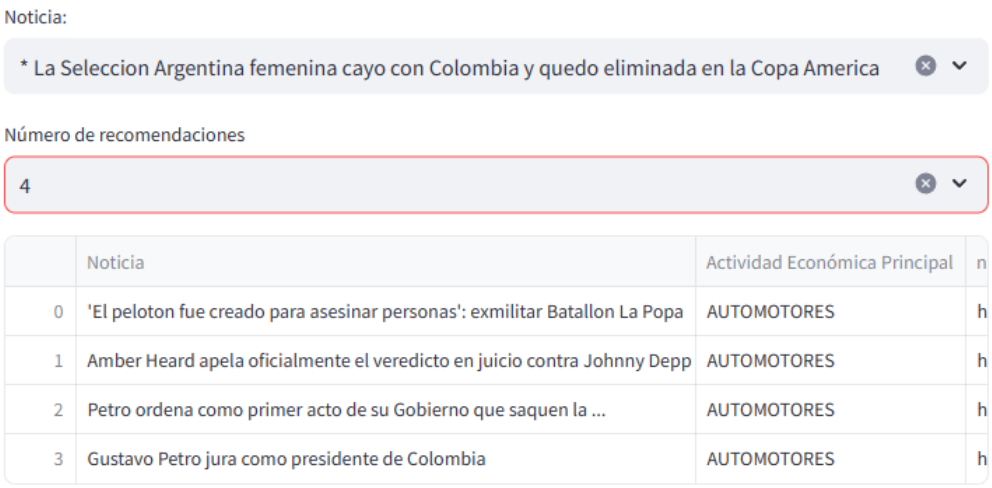
Actividad Económica Principales

	Actividad Económica	% Noticias
0	CONSTRUCCION DE EDIFICIOS RESIDENCIALES	6.97%
1	COMERCIO DE VEHICULOS AUTOMOTORES NUEVOS	3.05%
2	CONSTRUCCION DE CARRETERAS Y VIAS DE FERROCARRIL	2.52%
3	GENERACION DE ENERGIA ELECTRICA	2.46%
4	EDUCACION DE UNIVERSIDADES	2.32%
5	ACTIVIDADES DE HOSPITALES Y CLINICAS, CON INTERNACION	2.12%
6	COMERCIO AL POR MAYOR DE PRODUCTOS FARMACEUTICOS, MEDICINALES, COSMET	1.99%
7	OTRAS ACTIVIDADES RELACIONADAS CON EL MERCADO DE VALORES	1.86%
8	OTRAS ACTIVIDADES DE SERVICIO FINANCIERO, EXCEPTO LAS DE SEGUROS Y PENSIO	1.79%
9	ACTIVIDADES INMOBILIARIAS REALIZADAS CON BIENES PROPIOS O ARRENDADOS	1.73%

Luego el usuario elige la noticia de su interés, y el número de recomendaciones que desea



Con base en el algoritmo de LDA, al seleccionar una noticia, el tablero brinda el número de recomendaciones que el usuario elija. Adicionalmente, el usuario puede encontrar el link que lo direcciona a la noticia recomendada.



	Noticia	Actividad Económica Principal	n
0	'El peloton fue creado para asesinar personas': exmilitar Batallon La Popa	AUTOMOTORES	h
1	Amber Heard apela oficialmente el veredicto en juicio contra Johnny Depp	AUTOMOTORES	h
2	Petro ordena como primer acto de su Gobierno que saquen la ...	AUTOMOTORES	h
3	Gustavo Petro jura como presidente de Colombia	AUTOMOTORES	h

Como principales conclusiones, vemos que el tablero ofrece solución justamente lo que se propone como problema de negocio, haciendo recomendaciones de noticias relevantes para la fuerza comercial de Bancolombia. Dichas recomendaciones tienen una base cuantitativa y estadística que dan confiabilidad a las mismas.