Marcela Toro
Natalia Sierra
Juan Rave
Juan Latorre

Sistema de recomendación de noticias sobre clientes corporativos

Resumen

El grupo comercial de Bancolombia desea tener un mayor conocimiento de sus clientes corporativos a través de la lectura de noticias de diferentes medios de prensa locales e internacionales, en pro de poder mejorar la oferta de servicios. 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.

Los datos están conformados de la siguiente manera: *clientes:* listado de clientes, descripción de la actividad económica y el subsector, *clientes_noticias:* relación entre cliente y noticias consultadas mediante el proceso de descarga de información y *noticias:* contiene cada una de las noticias consultadas.

Para la recomendación de noticias, encontramos un buen acercamiento en un algoritmo basado en el modelado de tópicos, el cual hemos implementado encontrando las probabilidades más altas de que una noticia pertenezca a un tópico y sector. El resultado constituye un muy buen acercamiento al recomendar noticias con probabilidades superiores al 99,8% de tener información similar a la noticia usada como base, lo que facilita la recomendación de noticias relevantes para sectores específicos.

Introducción

El área comercial del grupo Bancolombia requiere conocer mejor a sus clientes corporativos a través de información relevante, actualizada y confiable generada por medios de comunicaciones locales e internacionales. Actualmente los comerciales del banco cuentan con miles de noticias relacionadas a cada uno de sus clientes y muy poco tiempo para su lectura y análisis; incluso dentro de la gran cantidad de noticias que deben leer, un gran porcentaje corresponden a noticias falsas o engañosas y no vigentes. Es de interés poder determinar las noticias relevantes para un sector particular de clientes, que permitan a la fuerza comercial informarse para atender determinado sector y ser más efectivos en su labor.

Debido a lo anterior, el banco requiere una herramienta que le permita a la fuerza comercial tener un recomendador de noticias con base en otras noticias que han sido definidas como relevantes para cada uno de los clientes y sectores de la economía. Desde nuestro equipo planteamos solucionar mediante la aplicación de técnicas de NLP y Aprendizaje No Supervisado, que permitan segmentar el contenido de las noticias relacionadas con los clientes y determinar qué características tienen en común para ser recomendadas.

Para afrontar este problema, se plantea extraer información característica de los textos de las noticias que permitan su representación matemática, con técnicas de NLP, esto corresponde a convertir cada texto de las noticias en una representación matricial con cierta dimensionalidad que depende de la técnica usada. Cuando se tiene esta representación, se

calcula la similitud entre estas o se agrupan las noticias que tengan características similares y formen grupos representativos de sus características principales, permitiendo determinar cuán relacionada se encuentra una noticia con otra y esta sea de interés para la fuerza comercial.

Por un lado, el aprendizaje no supervisado, tiene diferentes campos de aplicación en que se destaca, la agrupación o Clustering, donde el UL ha sido ampliamente utilizado en el campo de la agrupación de datos, donde resaltan algoritmos como k-means, DBSCAN y clustering jerárquico se han aplicado con éxito en diversas aplicaciones, como segmentación de usuarios, análisis de redes sociales y procesamiento de imágenes y texto.

Por su parte, el Procesamiento del Lenguaje Natural (NLP), también ha resaltado últimamente por los grandes problemas que es posible solucionar con estas técnicas. Uno de los principales aportes son los modelos de lenguaje, la aparición de modelos de lenguaje pre-entrenados como BERT, GPT-3 y sus sucesores ha revolucionado el campo del NLP. Estos modelos han mejorado significativamente el rendimiento en tareas de comprensión de texto, traducción automática y generación de texto[2].

Aunque se presentaron las diferentes aplicaciones actuales del aprendizaje no supervisado y el procesamiento de lenguaje natural, es importante resaltar que existen múltiples problemas los cuales se pueden abordar aplicando mezclas de estas técnicas. Una de estas técnicas es el clustering de texto, esta técnica se ha utilizado para agrupar documentos similares, identificar temas en grandes conjuntos de texto y mejorar la recuperación de información, que es particularmente útil para la solución del problema que se presenta en este trabajo [3].

En conclusión, el Aprendizaje No Supervisado y el Procesamiento del Lenguaje Natural son áreas de investigación y desarrollo en constante evolución con numerosas aplicaciones prácticas en una variedad de campos, dando lugar a avances significativos en el análisis de texto y la comprensión del lenguaje humano.

Materiales y Métodos

Los datos fueron extraídos del Dataton 2022, esta es una competencia que realiza el centro de excelencia en analítica, inteligencia artificial y Gobierno de Información del Grupo Bancolombia, competencia que fue recomendada por el profesor. Los datos se encuentran conformados por tres bases de datos, contienen información sobre clientes, noticias y la relación entre ellos.

La base de clientes contiene información sobre 1507 clientes, su actividad económica y el subsector al que pertenece, según ciiu agrupado por división, grupo, clase y subsector. Cada registro en esta base representa a un cliente único, identificado por su nit. No hay datos nulos ni duplicados. La mayoría de las variables son de tipo objeto, excepto el nit que es de tipo entero. Se observa una gran variedad de subsectores, divisiones, grupos y clases en la clasificación industrial. En subsectores encontramos 83 categorías, 81 categorías en la descripción del ciiu por división, 156 por grupo y 244 por clase.

La base de datos de noticias contiene información sobre 23.377 noticias consultadas, en 23.116 urls, 2 fechas, 22.772 títulos y 21.621 contenidos. La dimensión de la base de datos es de 23.377 por 11 variables. No hay datos nulos ni duplicados. La clave principal es `new_id`.

La base de datos de clientes_noticias contiene información sobre la relación entre 1507 clientes y 23.377 las noticias consultadas, con los identificadores y las urls correspondientes. La dimensión de la tabla es de 74.709 registros por 5 columnas. No hay datos nulos ni duplicados. Todas las variables son de tipo objeto.

A partir de la distribución de noticias por cliente se puede observar que el número promedio de noticias por cliente es de aproximadamente 49.54, con una desviación estándar de 34.21. Esto indica que hay una variabilidad considerable en el número de noticias por cliente.

El número mínimo de noticias por cliente es 4, mientras que el número máximo es 187. El primer cuartil (25%) es 27, lo que significa que el 25% de los clientes tienen 27 noticias o menos. La mediana (50%) es 36, lo que indica que el 50% de los clientes tienen 36 noticias o menos. El tercer cuartil (75%) es 60, lo que significa que el 75% de los clientes tienen 60 noticias o menos.

El algoritmo para el preprocesamiento de las noticias consistió de los siguientes pasos: primero, se eliminan caracteres especiales y acentos, convirtiendo caracteres acentuados en sus equivalentes no acentuados, se eliminan todos los caracteres que no son letras o espacios en blanco, incluyendo números y caracteres especiales, se reemplazan múltiples espacios en blanco seguidos, tabulaciones o saltos de línea por un solo espacio, se eliminan palabras con representaciones menores a dos letras y finalmente se convierte todo el texto a minúsculas.

Se realiza la tokenización y lematización (obtiene las formas base de las palabras) del texto utilizando el modelo de procesamiento de lenguaje natural de spaCy., se eliminan los stopwords, las cuales no aportan relevancia al análisis del texto.

Resultados y Discusión

La implementación del algoritmo consistio en unir cada uno de las bases de datos, agrupando los datos presentes en clientes_noticias por el código interno de la noticia el cual es único y la clasificación de la actividad industrial con mayor número de incidencias a través de la moda. De esta manera obtenemos la noticia que es más relevante por subsector. Posteriormente a la unión de las bases de datos realizamos el proceso de homogenización del texto, algoritmo descrito en el literal anterior.

El primer acercamiento se realizó con el algoritmo **CountVectorizer**, el cual convierte cada noticia en una matriz de recuento de palabras, paso seguido obtenemos la matriz de similitud del coseno y construimos nuestra función de recomendación, la cual se encuentra presente en github.https://github.com/juanrave/Sistema-de-recomendacion-clientes-con-potencial-comercial/blob/main/code/ProyectoANS.ipynb

A continuación se describe el algoritmo, paso 2 y 3: obtenemos todos los títulos de noticias y obtenemos el índice de la noticia que usamos como base para la recomendación, paso 4: con la matriz del coseno de similitud, obtenemos las medidas de similitud de cada noticia a la noticia recibida como parámetro y las almacenamos en sim scores, en el paso 5: ordenamos los valores en mayor a menor similitud y en el paso 6 obtenemos la cantidad de noticias a recomendar descartando la primera. En el paso 7 obtenemos el título de la noticia y el subsector al que pertenece, agregamos la columna título original de la noticia, sector original y la medida de similitud, finalmente definimos una columna llamada match, la cual corresponde a si la noticia que ha sido recomendada por nuestro algoritmo tiene como subsector el mismo de la noticia base.

Para nuestra prueba inicial, tomamos la noticia asociada al cliente corporativo con nit 804001062 RUITOQUE SA ESP, “Por no contar la verdad, Hugo Aguilar quedo sin cupo en la JEP”, perteneciente al subsector ELECTRICIDAD, donde le solicitamos al sistema que nos recomiende 5 noticias.

```
recomendador(df_news['news_title'][0], cosine_sim2, df_news, num_reco)[['news_title', 'subsec', 'sim_scores']]
```

	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

En la imagen anterior notamos el porcentaje de similitud de cada una de las noticias que fueron recomendadas por nuestro sistema.

Realizamos un segundo acercamiento vectorizando las noticias a través de **TF-IDFVectorizer** el cual tiene en cuenta la importancia relativa de las palabras en el conjunto completo de documentos. Posteriormente calculamos el producto escalar a través de Linear Kernel y usamos de nuevo nuestra función de recomendación.

```
recomendador(df_news['news_title'][0], cosine_sim, df_news, num_reco)[['news_title', 'subsec', 'sim_scores']]
```

	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

Para este caso el puntaje de recomendación que se obtiene se encuentra en el rango de 0.16 a 0.2, tal como se aprecia en la imagen anterior. Estas recomendaciones se basan en la importancia relativa de cada palabra y no en la frecuencia absoluta como en el caso anterior.

Por último, realizamos una nueva implementación de recomendación de noticias basado en temas, para lo cual usamos el modelado de temas con LDA, en donde posterior a obtener el modelo, implementamos la función de recomendación en la cual obtenemos el tópico o el tema al que pertenece la noticia recibida como parámetro y el subsector y con base a esta información buscamos todas las noticias asociadas al mismo tópico ordenadas de manera descendente por la probabilidad de pertenecer al tema, retornando las noticias con mayor probabilidad de pertenecer al tema.

Los temas generados fueron 4 y agrupan los siguientes subsectores:

Tópico 0: subsectores: ELECTRICIDAD 259, MEDIOS 207, SEGUROS 425, TRANSPORTE TERRESTRE 207

Tópico 1: ELECTRICIDAD 60, MEDIOS 129, SEGUROS 55, TRANSPORTE TERRESTRE 29

Tópico 2: ELECTRICIDAD 47, MEDIOS 67, SEGUROS 43, TRANSPORTE TERRESTRE 82,

Topico 3: ELECTRICIDAD 42, MEDIOS 48, SEGUROS 21, TRANSPORTE TERRESTRE 66

Para nuestro caso de prueba usamos la noticia ‘Javier Rodriguez Soler, nuevo presidente del Consejo Asesor del Centro de Educación y Capacidades Financieras de BBVA’. para la cual nos genera las siguientes recomendaciones:

	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

El algoritmo que mejor genera recomendaciones es el basado en temas, para el cual se obtienen probabilidades más altas de pertenecer una noticia a un tópico o tema en particular.

Conclusión

En el campo de la recomendación de noticias, un estudio destacado es el realizado por Gisela Yunanda, Dade Nurjanah y Selly Meliana, publicado en la revista “Building of Informatics, Technology and Science (BITS)” en 2022. Este estudio se centra en la recomendación de noticias utilizando el método TF-IDF, con un enfoque particular en las noticias que los lectores han visitado previamente en el portal de Microsoft News. Lograron una impresionante tasa de aciertos del 80.77%.

Nuestro trabajo, presentado en este documento, adopta un enfoque diferente. Nos enfocamos en recomendar noticias que son relevantes para un sector económico específico. La eficacia de nuestras recomendaciones se evalúa mediante la identificación de si la noticia recomendada ha sido leída por un cliente que pertenece al mismo sector económico. De esta manera, esperamos proporcionar un servicio de recomendación de noticias altamente personalizado y relevante para nuestros usuarios.

El proceso de recomendación de noticias consistió en la construcción de 3 diferentes estrategias, para lo cual la primera implementación se realizó al vectorizar los textos con **CountVectorizer** y encontrar la similitud del coseno, este acercamiento nos entregó niveles de similitud de hasta 0.289540, con lo cual no logramos tener una alta probabilidad de que la noticia recomendada realmente sea relevante para el lector. Posteriormente, realizamos la vectorización de textos con **TF-IDFVectorizer** y calculamos el producto escalar a través de Linear kernel y usamos de nuevo nuestra función de recomendación, encontrando porcentajes de hasta el 0.206347, lo cual es más bajo que los resultados de la primera implementación. Por último, realizamos la implementación del modelado basado en temas, para lo cual tenemos probabilidades superiores al 0.998932 de que una noticia sea similar a otra.

De acuerdo a lo anterior la última implementación basada en temas, arroja los mejores resultados, al permitir recomendar una noticia con base a otras con probabilidades muy altas para un tópico y sector en particular.

Bibliografía

[1] Khurana, D., Koli, A., Khatter, K. et al. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 82, 3713–3744 (2023). https://doi.org/10.1007/s11042-022-13428-4.

[2] Wang, Jue & Tao, Qing. (2008). Machine Learning: The State of the Art. IEEE Intelligent Systems. 23. 49-55. 10.1109/MIS.2008.107.

[3] Dogra, V. Verma, S. et al. A Complete Process of Text Classification System Using State-of-the-Art NLP Models (2022). https://doi.org/10.1155/2022/1883698.

[4] Dataton (2022). Datatón es una competencia que realiza el CdE en analítica IA, GI de Bancolombia https://www.kaggle.com/datasets/juancamilodiazzapata/dataton-2022.