Prueba técnica Departamento de datos no estructurados: Análisis de tweets a través de NLP

Autor: Nicolás Useche Narváez

Una de las principales aplicaciones del procesamiento natural del lenguaje es extraer automáticamente de grandes volúmenes de texto los temas que discuten las personas. Algunos ejemplos de grandes volúmenes de texto pueden ser noticias de redes sociales, opiniones de clientes, comentarios de usuarios, noticias, etc.

Saber de qué hablan las personas y comprender sus problemas y opiniones es muy valioso para empresas, administradores y campañas políticas. Y es realmente difícil leer manualmente volúmenes tan grandes y recopilar los temas.

Por lo tanto, se necesita un algoritmo automatizado que pueda leer los documentos de texto y extraer automáticamente los temas tratados. Tomando como caso de estudio el análisis de tweets extraídos de la red social de Twitter que contienen la palabra **Davivienda** y con el cual se quiere conocer cuál es la interacción que tienen los diferentes usuarios de la red social con el Banco Davivienda.

1. Tecnologías y Librerías Utilizadas

Para el desarrollo del proyecto se empleó el lenguaje de programación Python en su versión 3.9.12 haciendo uso de librerías como “pandas” y “numpy” para el manejo de matrices. “matplotlib”, “seaborn” y “wordcloud” para ilustraciones y gráficos. Para el preprocesamiento y análisis de texto se emplearon librerías como “re” para expresiones regulares, “spacy” para obtener las “STOP_WORDS” para el idioma español y “stanza” para la lematización de los tweets en idioma español. “sklearn” para vectorizar texto y realizar análisis de de N-gramas. Para el modelo LDA y la matriz termino documento utilicé la librería “gensim”.

2. Análisis Exploratorio

El proyecto comenzó con una exploración detallada de una base de datos denominada davivienda_tweets.csv, que contiene tweets relacionados con el Banco Davivienda para el mes de diciembre del año 2021, este dataset cuenta con 1811 registros y 12 variables (“Unnamed: 0”, “UserScreenName” , “UserName”, “Timestamp“ , “Text” , “Embedded_text”, “Comments” ,”Emojis” , “Likes”, “Retweets”, “Tweet URL”, ”Image link”).

2.1 Limpieza del dataset

Los datos crudos a menudo contienen errores, valores faltantes, duplicados, o formatos inconsistentes. La limpieza ayuda a corregir estos problemas, asegurando que el análisis se realice sobre datos precisos y fiables.

Se decide eliminar las columnas “Unamed: 0”, “Tweet URL”, “Image Link” y “Emojis”, principalmente porque no aportan valor semántico para el análisis y la mayoría de los registros son valores NaN.

Posteriormente se realiza la imputación de datos faltantes, comenzando con las variables numéricas (“Likes”, “Retweets”, “Comments”). En este punto se encontró una novedad en la imputación y

conversión a tipo de dato numérico pues por naturaleza twitter abrevia el número de comentarios, likes y retweets por miles y le agrega al final el sufijo "mil". En este punto se diseñó una función especial para la imputación y conversión a dato numérico, eliminando el "mil", multiplicar por 1000 y convertir a numérico.

Por último, se realiza la imputación de los valores faltantes en “UserScreenName” utilizando el valor fijo correspondiente a la columna de UserName. Y se verifica que no haya registros duplicados. Como resultado de estas operaciones se logró obtener un dataset completo y sin valores faltantes, listo para procesar el texto esencial contenido en los tweets.

Análisis de usuarios con mayor número de tweets

Se realiza un conteo de número de tweets realizados por cada usuario con el fin de obtener el top 10, como lo muestra la siguiente figura.

	UserName	TweetCount
0	@Davivienda	245
1	@CNOGUERA20	41
2	@davicorredores	23
3	@Edimejia1979	9
4	@Juanma7725	8
5	@FabioFernandoH1	7
6	@DaviEscucha	7
7	@dataiFX	7
8	@Joacoro	6
9	@EnriqueDelgadoP	6

Figura 1. Top 10 usuarios con mayor número de tweets

Dentro de este top se encuentra en primer lugar el perfil de Davivienda, el cual, de acuerdo con la exploración realizada se observa que la mayoría de los tweets son de soporte para los usuarios. Estos tweets no aportan al análisis, debido a que la mayoría repiten las mismas palabras e invitan al usuario a comunicarse con el banco para resolver sus inquietudes.

Esto puede generar ruido para el entrenamiento del modelo LDA, por lo que se decide eliminar los tweets que provengan de este perfil los cuales en su contenido incluyan la palabra como “En respuesta a”.

Análisis de interacciones por día

Otro tipo de análisis son la cantidad de interacciones (Comentarios, Likes, Retweets) por día, como se muestra en la siguiente ilustración:

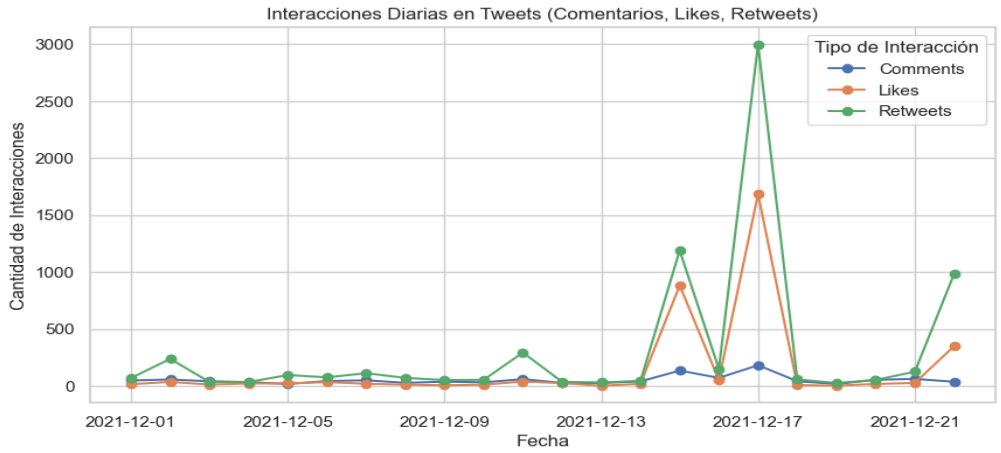

Figura 2. Interacciones diarias en tweets

Los resultados muestran un bajo nivel de interacciones (Comentarios, Likes, Retweets) por día en el periodo de muestra correspondiente al mes de diciembre del 2021, a excepción de dos picos representativos que se presentan el 15 y 17 de diciembre, día en el cual el senador Wilson Neber Arias Castillo realizó una denuncia pública.

Lo anterior se debe a que en la actualidad hay usuarios que dada la gran cantidad de seguidores que tienen en las redes sociales, al momento de realizar una publicación de cualquier temática generan gran impacto en muchos usuarios.

La gráfica de la figura 2, muestra una característica común de los datos de redes sociales, donde pocos contenidos se vuelven virales mientras que la mayoría permanece con bajo nivel de interacción.

Análisis de la Longitud promedio de los Tweets

Otro aspecto importante es el análisis de la longitud promedio de los tweets, el cual puede ser un indicador de la complejidad o la riqueza del contenido.

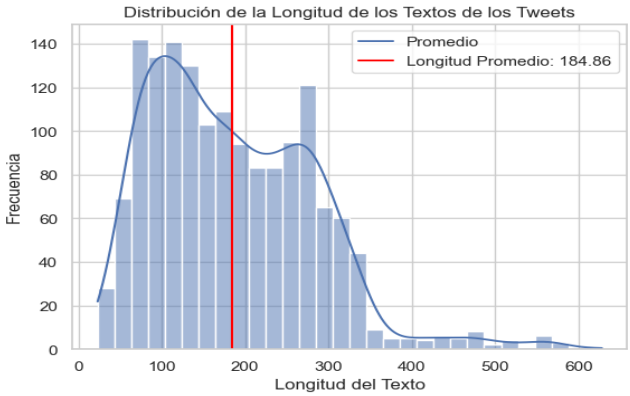

Figura 3. Distribución de la longitud de los tweets

A través de la gráfica de la figura 3, se puede identificar que la mayoría de los tweets tienen una longitud promedio de 184 caracteres, lo que es esperado dada la naturaleza concisa de los tweets. Existe una distribución relativamente normal con una cola hacia la derecha, indicando la presencia de algunos tweets más largos.

La longitud de los tweets puede ayudar en la segmentación de documentos antes de aplicar el modelo LDA, porque permite discernir la complejidad del contenido de un tweet al igual que la interpretación de los tópicos extraídos de este, es decir, los tweets con longitudes extremadamente cortas o largas pueden representar diferentes cosas; por ejemplo, tweets muy cortos pueden ser simplemente enlaces o mensajes sin mucho contenido informativo, mientras que tweets muy largos pueden contener discusiones detalladas.

3. Preprocesamiento de Texto

Para llevar a cabo el entrenamiento del modelo LDA, es importante procesar el texto de los tweets para que sea eficiente y legible para los modelos. Se ha incluido 4 pasos principales en el preprocesamiento:

- **Limpieza**:
    - Poner el texto en minúsculas para que el algoritmo no trate las mismas palabras en distintos casos como diferentes palabras.
    - Eliminación de texto entre corchetes, URL's, HTML,saltos de linea, puntuación , caracteres especiales, palabras de 3 caracteres o menos.
- **Eliminación de palabras vacías (stop words)**: Estas hacen referencia a palabras muy comunes que parecen tener poco valor para ayudar a seleccionar los tópicos para el modelo LDA, por tanto, se eliminan o se omiten.
- **Tokenización**: Se realiza la conversión de cadenas de texto del tweet a una lista de tokens, dividiendo las palabras por espacios en blanco, similar al método. split() de Python.
- **Lematización**: se decide lematizar las palabras en lugar de hacerles un estematización, ya que la lematización considera la estructura y propiedades gramaticales de las palabras y las convierte a una forma base, mientras que la estematización es una técnica que se utiliza para reducir una palabra hasta la raíz de la palabra. Esto puede resultar en raíces que no son realmente palabras reales. Por lo tanto, la lematización es mejor para la modelización de temas ya que conserva el significado semántico de las palabras. Sin embargo, computacionalmente, es más costoso.

4. Análisis de Nube de Palabras

La nube de palabras proporcionada refleja términos y frases que se usaron frecuentemente en tweets relacionados con el banco Davivienda en diciembre de 2021. Estas son útiles para visualizar de forma rápida y directa los términos más mencionados dentro de un gran conjunto de datos de texto.

Algunas observaciones y análisis basados en esta nube se presentan a continuación:

- **Posibles problemas de servicio**: Palabras como "problema", "esperar", "necesitar", "funcionar" y "solución" sugieren que los usuarios podrían estar discutiendo problemas o presentando quejas de los servicios del banco, buscando asistencia o soluciones a esos problemas.

- **Interacciones con el cliente**: La frecuencia de palabras como "llamar", “sfcsupervisor”, “ayudar", "pedir", “favor”, "gracias" sugiere que hay una cantidad significativa de comunicación entre los clientes y el banco. Esto podría incluir solicitudes de ayuda, agradecimientos por el servicio o necesidad de asistencia de un representante del banco.
- **Transacciones y servicios**: Términos específicos de transacciones como “daviplata”, “cajero”, “tarjeta”, “dinero”, "servicio", "pagar", y "crédito" sugieren que los tweets también incluyen discusiones detalladas sobre los servicios financieros proporcionados por el banco.

La naturaleza de las palabras sugiere que podría haber una preocupación general sobre la eficiencia del servicio, la seguridad y la atención al cliente.

Figura 4. Nube de palabras

5. Matriz termino-documento

Tras la fase inicial de preprocesamiento, se procede a preparar los datos para el entrenamiento del modelo LDA, esto significa que se debe transformar el texto en un vector (o matriz) de números con sentido. Esto puede hacerse mediante la técnica de Matriz termino-documento, la cual es una representación del texto que describe la aparición de palabras en un documento.

De esta manera, un término que aparece con frecuencia en el corpus (cuerpo) del texto tendrá una puntuación mucho más alta, y un término que aparece raramente en el corpus de texto tendrá una puntuación mucho más baja. Se descarta cualquier información sobre el orden o la estructura de las palabras en el documento y el modelo sólo se preocupa de si las palabras conocidas aparecen en el documento, no de dónde aparecen en el documento.

	afectar	asociado	aumento	cambio	caída	confianza	deteriorar	especialmente	indicador	leve	...	enormemente	capacitado	ginaa1931	gleniiaaa
0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	...	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0

5 rows × 4416 columns

Figura 5. Matriz termino-documento

6. Histogramas

Análisis de N-gramas

Para analizar el contenido de cada tweet se opta por extraer las características de N-gramas. Los cuales, capturan la estructura del lenguaje desde el punto de vista estadístico, como qué letra o palabra es probable que siga a la dada. Si los n-gramas son demasiado cortos, es posible que no capten diferencias importantes, es decir, se limitan a casos particulares y no se tiene un "conocimiento general" del tema. Por otro lado, cuanto más largo sea el n-grama (cuanto mayor sea n), más contexto tendrá para trabajar, como se observa a continuación:

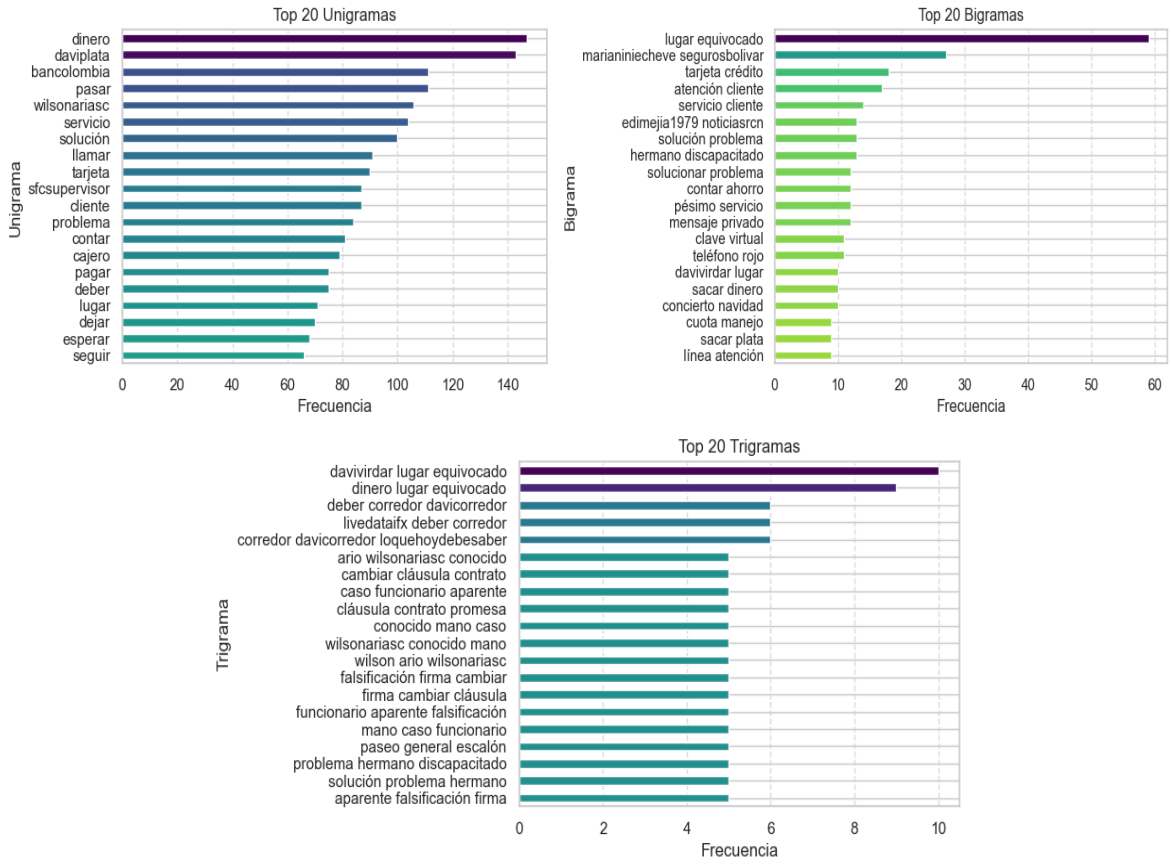

Figura 6. Top 20 Unigramas, Bigramas, Trigramas

Del análisis se puede inferir que las preocupaciones o temas comunes incluyen transacciones monetarias, problemas con servicios digitales o transacciones entre plataformas, atención al cliente, términos de contratos y posiblemente errores o fraudes.

Dada la presencia de palabras como “lugar equivocado” son indicadores del gran éxito que presentó esta campaña publicitaria por parte del banco Davivienda. Donde las personas utilizan estos términos para referirse a problemas con sus productos como tarjetas de crédito, cuentas y cajeros.

7. Latent Dirichlet Allocation- Modelo no supervisado de Tópicos.

Finalmente se procede a entrenar el modelo LDA el cual es capaz de detectar y extraer de manera automática relaciones semánticas latentes de grandes volúmenes de datos. Estas relaciones son los llamados tópicos, que son un conjunto de palabras que suelen aparecer juntas en los mismos contextos y permiten observar relaciones que son indetectables a primera vista.

En los parámetros de entrenamiento del modelo LDA se necesitan definir el número de temas, para esto se hace uso del puntaje de coherencia, es decir, calculando si la similitud semántica entre las palabras principales (por tema) es alta o baja. En este caso, se realizaron pruebas construyendo el modelo con un mínimo de 5 temas hasta 25, calculando el puntaje de coherencia, como se muestra en la figura 7.

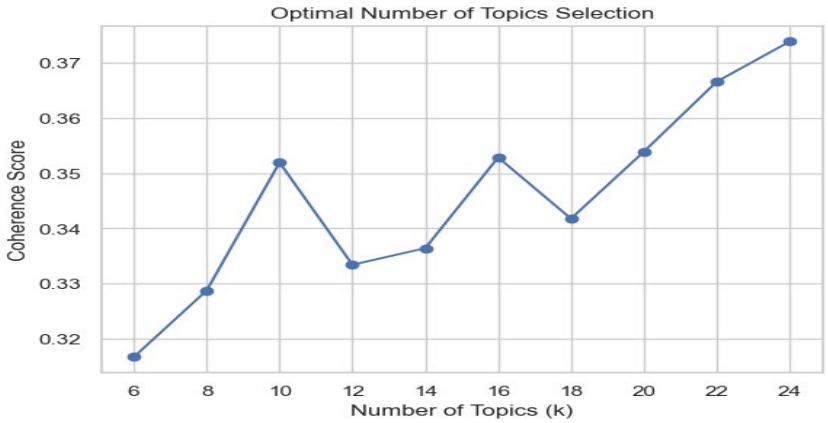

Figura 7. Número óptimo de temas

A pesar de que, según el puntaje de coherencia, el mejor número de temas es entre 22 y 24, se decide trabajar con 10 temas como se muestra en la figura 8, pues de esta forma se puede identificar de forma sencilla el significado de cada tema, de lo contrario se presentarían solapamientos entre temas y estarían muy cerca uno de otro.

8. Interpretación de Resultados y conclusiones generales.

Ahora que el modelo LDA está listo, el siguiente paso es examinar los temas producidos y las palabras clave asociadas. La herramienta que se escoge es el gráfico interactivo del paquete pyLDAvis. Donde cada burbuja del gráfico de la izquierda representa un tema. Cuanto más grande es la burbuja, mayor es la prevalencia de ese tema. El gráfico de barras horizontales de la derecha representa las palabras más relevantes de cada tema.

Como se mencionaba anteriormente, el gráfico es interactivo, lo que permite seleccionar temas específicos y ver las palabras relacionadas con cada tema, con la esperanza de inferir el significado de cada tema.

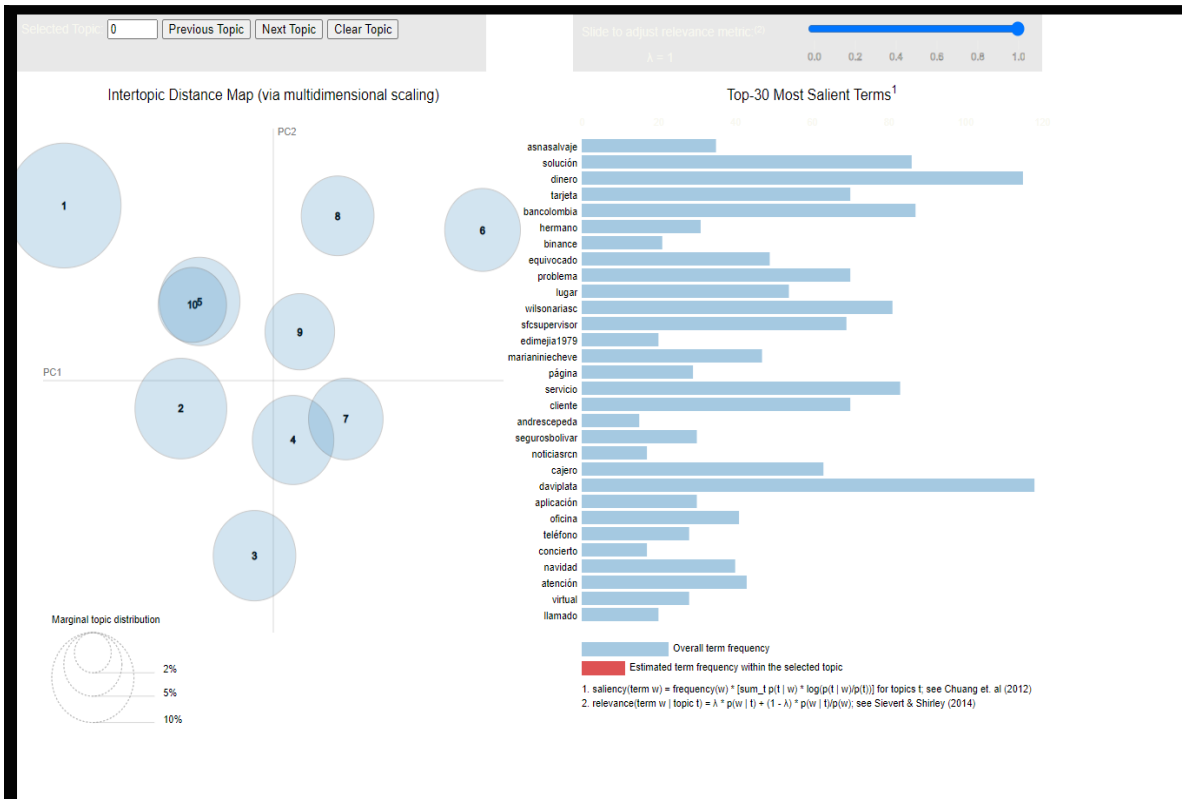

Figura 8. Visualización de tópicos.

Del modelo de 10 tópicos se puede interpretar que:

- Tópico 1: "Problemas con la app de Daviviplata"
- Tópico 2: "Problemas con las transferencias de dinero cuentas Davivienda a otros bancos"
- Tópico 3 y 4: "Publicaciones a modo de queja de usuarios con un número considerable de seguidores"
- Tópico 5 y 10: "Problemas con el servicio al cliente, tarjetas y cajeros"
- Tópico 6: "Piloto con la plataforma de Binance y Davivienda"
- Tópico 7: "Referencias a Davivienda utilizando la famosa frase de “En estos momentos su dinero puede estar en el lugar equivocado, tráigalo a Davivienda”
- Tópico 8: "Problemas con el dinero de los clientes y daviplata"
- Tópico 9: "Publicaciones para reportar fraude o estafa por parte de usuario con un número considerable de seguidores"

En general, se evidencia que la mayoría de los usuarios utilizan la plataforma de Twitter para reportar problemas con la app de Daviplata, problemas con las transferencias de dinero a otros bancos, problemas con el servicio al cliente, tarjetas y cajeros, y publicaciones a modo de queja de usuarios con un número considerable de seguidores. Adicionalmente, se evidencia que algunos usuarios utilizan la plataforma para reportar fraude o estafa.

Se probaron distintos métodos de agrupación y búsqueda de temas subyacentes a partir de los tweets del conjunto de datos, como el modelado de temas LDA. Además, se exploraron formas de visualizar los resultados y de compararlos con los temas originales de los tweets.

Sin embargo, probablemente debido a la propia naturaleza del texto (por ejemplo, los tweets son demasiado similares en general, o tienen muy pocas palabras para permitirnos segmentar el tema subyacente), acabamos con algunas distinciones borrosas entre cada tema, como se ve en las palabras principales por cada tópico tras el modelado LDA.

9. Referencias

Natural Language Processing Master ♔ | Kaggle

EDA and Preprocessing for BERT | Kaggle

Getting started with NLP - A general Intro | Kaggle

Topic modelling headlines: KMeans & LDA | Kaggle

Topic Modelling (LDA) on Elon Tweets | Kaggle

detecting_hatespeech | Kaggle

Gensim Topic Modeling - A Guide to Building Best LDA models (machinelearningplus.com)

Quick Text Pre-Processing. Making sense of messy tweets | by Rob Zifchak | The Startup | Medium

Stemming and Lemmatization in Python | DataCamp

Document-Term Matrix in NLP: Count and TF-IDF Scores Explained | HackerNoon

LDA Topic modeling with Tweets. Making sense of unstructured text | by Rob Zifchak | Towards Data Science