Trabajo Práctico N° 1

Procesamiento del lenguaje natural

Tecnicatura Universitaria en Inteligencia Artificial

FCEIA - UNR

2do año

Integrantes

Fernández, Florencia

Palermo, Leonel

Salvañá, Leandro

1.Introducción

Este informe se enfoca en aplicar técnicas de análisis de datos y procesamiento del lenguaje natural a un conjunto de datos que corresponden a noticias de distintas categorías. Las actividades incluyen análisis de datos, estandarización, reducción de dimensionalidad con PCA, clasificación de texto y armado de resúmenes. Estas técnicas se aplican con el objetivo de analizar las diversas herramientas que existen en esta rama de la inteligencia artificial y los rendimientos encontrados.

Nota: los resultados del código aplicado se analizan en este informe y también se presentan las tablas y gráficos pertinentes correspondientes a cada sección del trabajo. Se agregan aquí para que las explicaciones dadas sobre los resultados obtenidos estén acompañadas visualmente y sea más sencillo de entender. Además, correr el código insume mucho tiempo ya que la cantidad de gráficos es extensa.
El código utilizado se anexa en un archivo .ipynb y las secciones se encuentran claramente distinguidas en forma de títulos de distintos niveles y texto acorde. Dichas secciones coinciden con las del informe en cada punto del trabajo.

Si desea correr por su cuenta el código, deberá instalar previamente las librerías necesarias en su entorno de trabajo. Las librerías utilizadas están detalladas al inicio del archivo .ipynb.

2. Ejercicio 1

Construir un dataset haciendo web scraping de páginas web de su elección.

- Definir 4 categorías de noticias/artículos.
- Para cada categoría, extraer los siguientes datos de 10 noticias diferentes:
  - url (sitio web donde se publicó el artículo)
  - título (título del artículo)
  - texto (contenido del artículo)

Recomendaciones: elegir blogs para evitar los límites de lectura para los medios que exigen suscripción. Investigue sobre el archivo robots.txt y téngalo en cuenta. Considere también espaciar las consultas para evitar saturar el sitio.

Utilizando los datos obtenidos construya el dataset en formato csv.

Sitios web sobre los que se realiza el web scraping:

- 'https://eurofitness.com'
- 'https://www.mijuegobonito.com'
- 'https://universidadeuropea.com'
- 'https://recursosparapymes.com'

2.1. Diccionario de permisos de los sitios web

```
pp.pprint(permisos_web)
```

```
{   'https://eurofitness.com': {   'aviso_legal': 'UBAEFITNESS, S.L. autoriza '
                                                  'a los usuarios a acceder y '
                                                  'navegar en el Sitio Web, '
                                                  'utilizando los servicios y '
                                                  'visualizando los contenidos '
                                                  'que allí se incorporen.\n'
                                                  'El acceso, visualización y, '
                                                  'si acaso, descarga de los '
                                                  'Contenidos y/o Servicios se '
                                                  'realizará siempre y en todo '
                                                  'caso con finalidades '
                                                  'estrictamente personales y '
                                                  'no comerciales.\n'
                                                  '\n'
                                                  'ENLACES',
                                'robots.txt': 'User-agent: *\n'
                                              'Disallow: /wp-admin/\n'
                                              'Allow: '
                                              '/wp-admin/admin-ajax.php'},
    'https://recursosparapymes.com': {   'aviso_legal': 'El Usuario se '
                                                        'compromete a respetar '
                                                        'los derechos de '
                                                        'propiedad intelectual '
                                                        'e industrial de '
                                                        'RecursosParaPymes.com. '
...
                                                       'Crawl-Delay: 10\n'
                                                       '\n'
                                                       'User-agent: Pinterest\n'
                                                       'Crawl-delay: 1'}}
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Se ha encontrado que los sitios web a consultar para obtener noticias permiten el uso de la información que allí se encuentra si es para uso personal y con fines no comerciales.

2.2. Creación del dataframe de noticias que será guardado como archivo .csv

```
# Se observa el dataframe obtenido
df_noticias.head()
```

	urls	titulo	contenido	categoria
0	https://eurofitness.com/blog-deportes/prensa-s...	¿Qué es mejor, prensa o sentadilla?	\nEl uso de ejercicios de prensa o sentadilla ...	fitness y deporte
1	https://eurofitness.com/blog-deportes/motivaci...	La clave del éxito para entrenar: motivación...	\nTodos buscamos alcanzar el éxito en diferent...	fitness y deporte
2	https://eurofitness.com/blog-deportes/efectos-...	¿Qué le pasa a tu cuerpo cuando dejas de hac...	\nEl ejercicio es una parte importante de un e...	fitness y deporte
3	https://eurofitness.com/blog-deportes/ejercici...	Los mejores ejercicios de fuerza para corred...	\n \n \nLos ejercicios de fuerza son esenciale...	fitness y deporte
4	https://eurofitness.com/blog-deportes/entrenar...	Cómo entrenar con pesas sin ganar volumen	\nNo todo el mundo tiene los mismos objetivos ...	fitness y deporte

Se han obtenido 40 noticias en total, 10 por cada categoría.

2.3. Observaciones y conclusiones

Al realizar el ejercicio solicitado y utilizar diferentes librerías para llevarlo a cabo se encuentran algunas reflexiones:

- El web scraping permite extraer información específica y estructurada de páginas web y esto facilita el análisis y la posterior utilización de los datos. Además, puede automatizar la extracción de datos de múltiples páginas web, lo que ahorra tiempo y esfuerzo en comparación con la recopilación manual. Otra ventaja es que no está limitado por las restricciones de una API. Puede extraer cualquier dato visible en la página web. Todo esto los hace una herramienta valiosa para realizar análisis de contenido de interés para las personas y la comprensión de los comportamientos de las mismas.

- Nos hemos enfrentado a la realidad de que muchos sitios son muy restrictivos con las políticas de acceso a los datos por lo que las fuentes se ven limitadas. Consideramos que es muy importante asegurarse de que se están respetando los permisos de cada sitio y es por ello que se ha generado el diccionario de permisos donde se hace visible qué partes del sitio pueden ser accedidas.

- En cuanto al uso de las librerías en sí mismas, se requiere conocimiento y análisis previo del sitio para obtener la información deseada. Existe una fragilidad en este sentido ya que si el sitio web tiene una estructura variable, puede ser más complicado definir un código que funcione cada vez que se utiliza.

- Comparando Selenium con BeautifulSoup, en esta experiencia resultó más engorroso el uso de la primera librería ya que requiere contar con drivers que se conecten con el navegador y realizar más configuraciones. Además, ha ocurrido en repetidas ocasiones que el driver no ha podido establecer la conexión mientras que BeautifulSoup ha podido acceder al contenido. Por ello, para esta aplicación particular, se prefiere el uso de BeautifulSoup.

3. Ejercicio 2

Utilizando los datos de título y categoría del dataset del ejercicio anterior, entrenar un modelo de clasificación de noticias en categorías específicas.

Escribir un análisis general del resultado obtenido.

Para entrenar un modelo de clasificación se debe primeramente acondicionar los datos que se utilizarán con tal fin para asegurar la calidad de los mismos y optimizar así los resultados obtenidos.

3.1. Procesamiento de los datos (títulos)

- Conversión a minúsculas:
  La conversión a minúsculas ayuda a homogeneizar el texto, evitando que las palabras escritas en mayúsculas se interpreten como diferentes de las escritas en minúsculas.
- Eliminación de acentos:

La eliminación de acentos simplifica el texto y evita posibles problemas de codificación que podrían surgir con caracteres acentuados.

- Eliminación de puntuación:
  La eliminación de puntuación permite centrarse en las palabras y su significado, sin verse distraído por signos de puntuación que no aportan información semántica.

- Remover símbolos innecesarios:
  La eliminación de símbolos innecesarios es similar a la eliminación de puntuación, se enfoca en quitar caracteres que no aportan información relevante para el análisis de texto.
- Remover números en títulos:
  Se remueven los números porque no son caracteres específicos de una categoría que puedan ayudar a indentificarla de manera más precisa. Las cuatro categorías utilizadas pueden presentar números en sus títulos para representar diferentes significados.
- Remover stopwords:
  Las stopwords son palabras comunes como "el", "la", "y", etc., que no aportan mucha información en términos de contexto y significado. Se eliminan porque ayuda a reducir la dimensionalidad del texto y a concentrarse en las palabras clave.
- Tratamiento de abreviaturas:
  El tratamiento de abreviaturas desambigua el texto al expandir las abreviaturas a sus formas completas, permitiendo una interpretación más precisa. Además puede que haya palabras en su forma abreviada y expandida y tratando las abreviaturas se consigue homogeneizar las palabras y que no dé lugar a que se interpreten como palabras distintas.

- Corrección de ortografía:
  Se aplica corrección de ortografía ya que mejora la calidad del texto al corregir posibles errores tipográficos o de escritura, lo que facilita su interpretación y análisis.

- Lematización:
  La lematización reduce las palabras a sus formas base o lemas, lo que ayuda a agrupar las diferentes formas de una palabra en una única representación. Esto simplifica el análisis y la extracción de información.

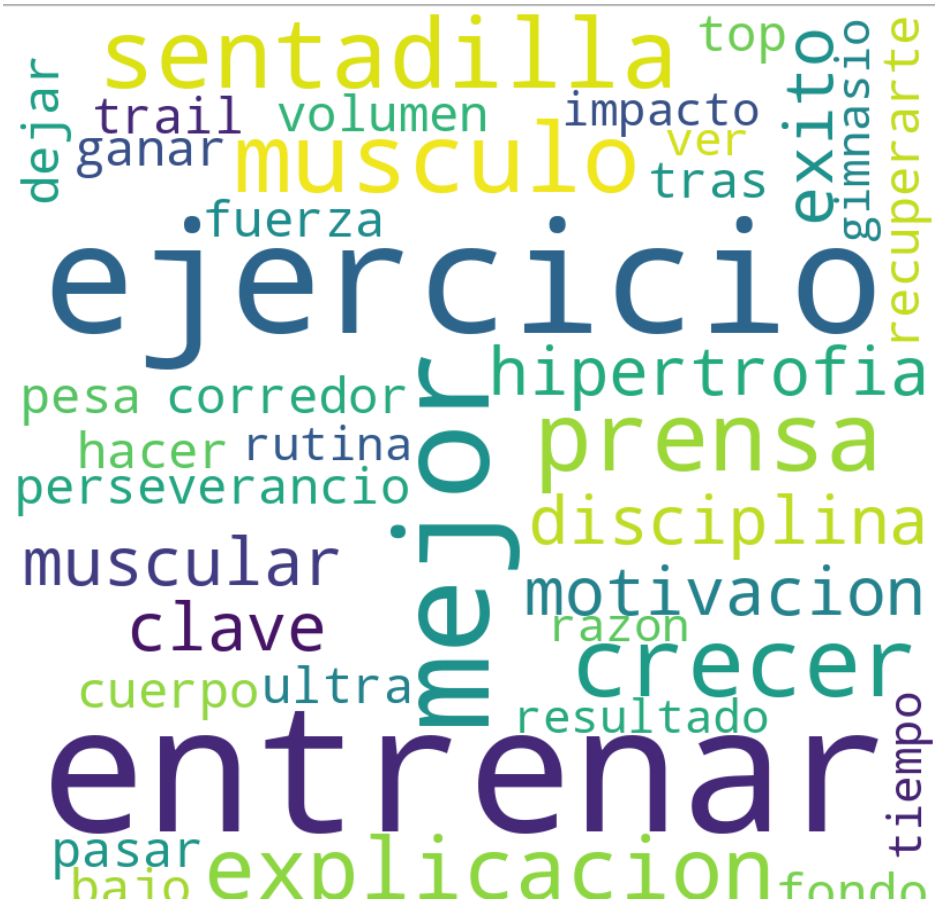
- Tratamiento de emojis:
  Se buscará la presencia de emojis y en caso de hallarlos se elimnarán para centrarse en las palabras. En este caso no se pretende hacer un análisis de sentimientos, sino una clasificación por lo que se prescinde de los emojis.

3.2. Frecuencia de palabras

Categoría economía y finanzas:

Categoría entretenimiento:

Categoría fitness y deporte:



Categoría medicina y salud:

Se encuentra que las palabras más representativas por categoría son:

- Economía y finanzas:
  - empresa
  - autónomo
  - ingreso
  - ayudar
  - digital
  - poder

- Entretenimiento:
  - juego
  - mesa
  - aula
  - primario
  - atención

- Fitness y deporte:
  - entrenar
  - ejercicio
  - sentadilla
  - mejor
  - sentadilla
  - explicación
  - crecer

- Medicina y salud:
  - reanimación
  - cardiopulmonar
  - causa
  - resonancia
  - aprendizaje

En general, se observa que las palabras más destacadas para cada categoría son representativas en su contenido y significado. También se encuentran palabras que son más generalizables en cuanto a pertenencia a algún grupo particular, poe ejemplo 'crecer' en fitness y deporte tendría sentido también en medicina y salud.

3.3. Vectorización TF-IDF

Como se requiere que los datos estén vectorizados para que el modelo los pueda entender, la variable con los datos de entrenamiento, así como la de test serán vectorizadas.

En este caso se utilizará TF-IDF para ello. Se ha elegido esta técnica porque tiene en cuenta la ocurrencia de palabras en los documentos para ponderar su importancia. Esto resulta muy importante porque equilibra la influencia de las mismas en los textos. Es decir, no pasan a ser significativas por el simple hecho de repetirse muchas veces.

3.4. Entrenamiento del modelo y análisis de métricas

```
labels = {
    0: 'economia y finanzas',
    1: 'entretenimiento',
    2: 'fitness y deporte',
    3: 'medicina y salud'
}
```

```
Precisión Regresión Logística: 0.625
Reporte de clasificación Regresión Logística:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       1.00      0.50      0.67         4
           2       1.00      0.50      0.67         2
           3       0.25      1.00      0.40         1

    accuracy                           0.62         8
   macro avg       0.81      0.75      0.68         8
weighted avg       0.91      0.62      0.68         8
```

Se realizan las siguientes observaciones:

Para la clase 0 ('economia y finanzas'), se obtiene un f1-score perfecto de 1.00, lo que indica un rendimiento excelente tanto en precisión como en recall.

Sin embargo, para las otras tres clases, el modelo tiene un rendimiento mixto. La clase 1 ('entretenimiento') tiene un recall bajo (0.50), lo que indica que el modelo tiene dificultades para identificar correctamente esta clase. Lo mismo sucede con la clase 2 ('fitness y deporte').

La clase 3 ('medicina y salud') tiene un recall perfecto, lo que indica que el modelo es muy bueno para identificar esta clase, pero el f1-score es bajo debido a un bajo valor de precisión.

3.5. Clasificación de nuevos textos

La frase 'Los 10 juegos de atención, estrategia y cálculo para tener diversión en familia.' pertenece a la categoría: entretenimiento

La frase 'Es recomendable realizar ejercicio por lo menos 1 hora dos veces a la semana.' pertenece a la categoría: fitness y deporte

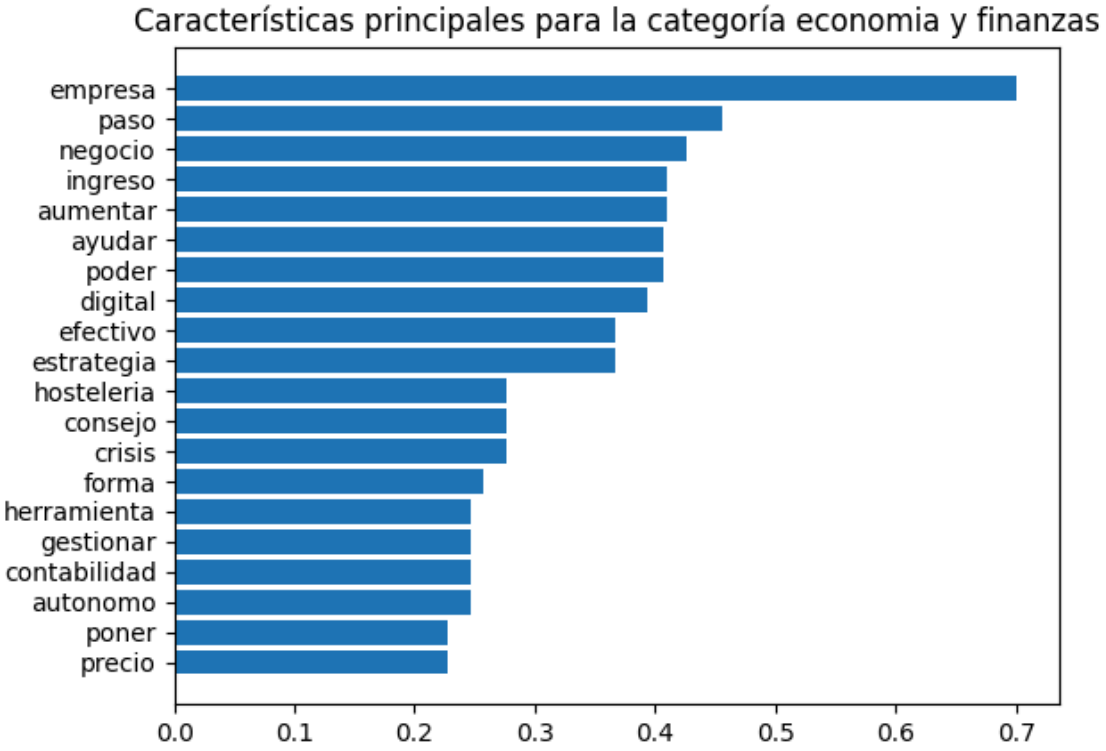
La frase 'La Inteligencia Artifial es genial!' pertenece a la categoría: medicina y salud

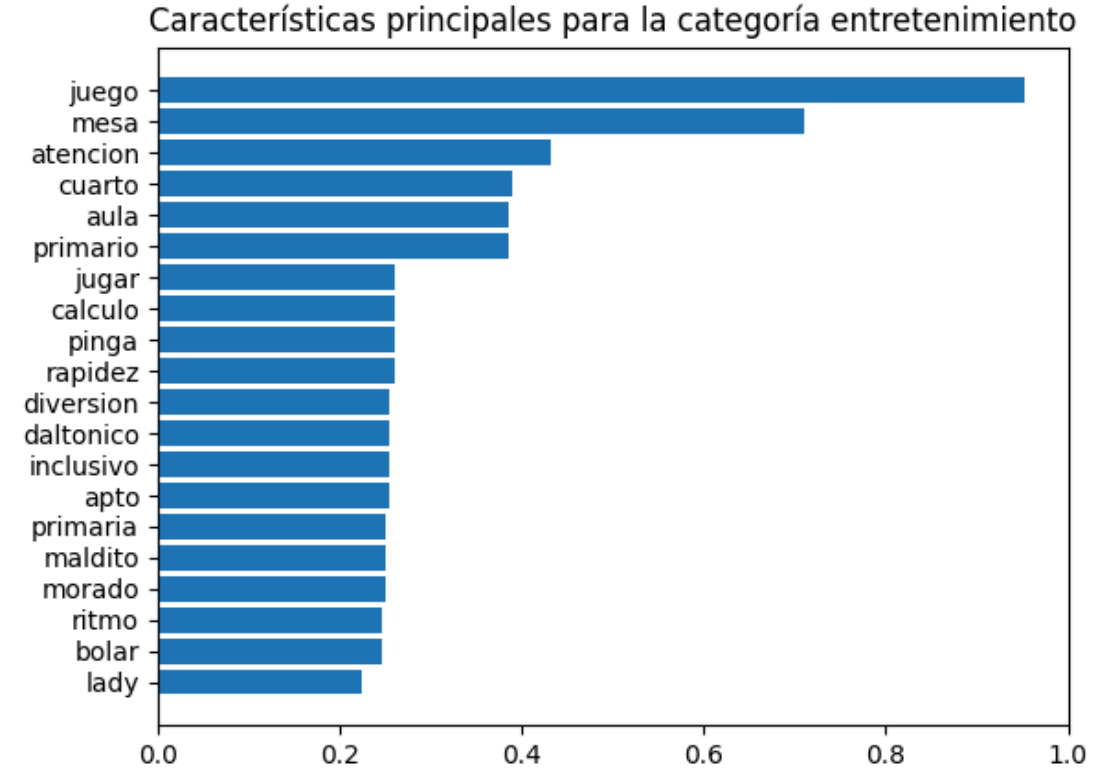
La frase '5 cosas importantes de saber a la hora de invertir tu dinero.' pertenece a la categoría: medicina y salud

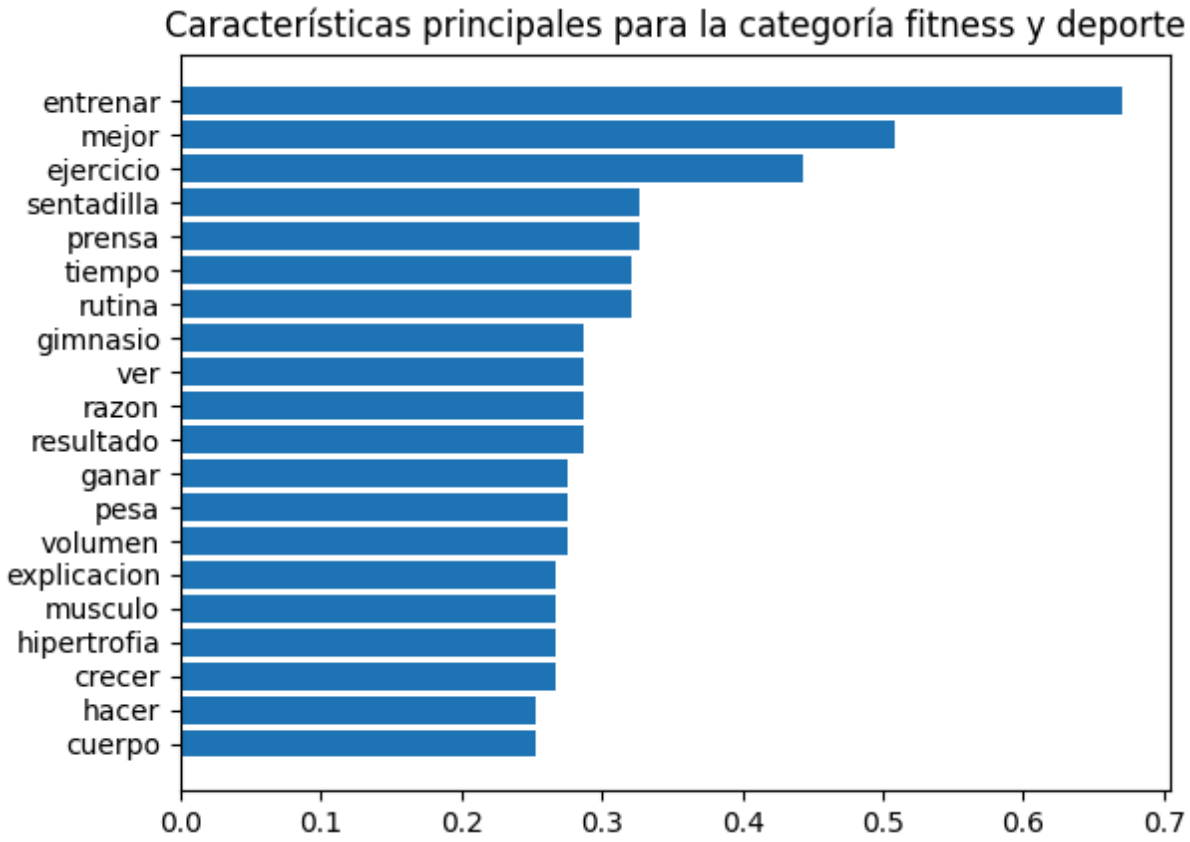
La frase 'Conocer sobre primeros auxilos, es importante para saber cómo cuidar nuestra salud.' pertenece a la categoría: medicina y salud

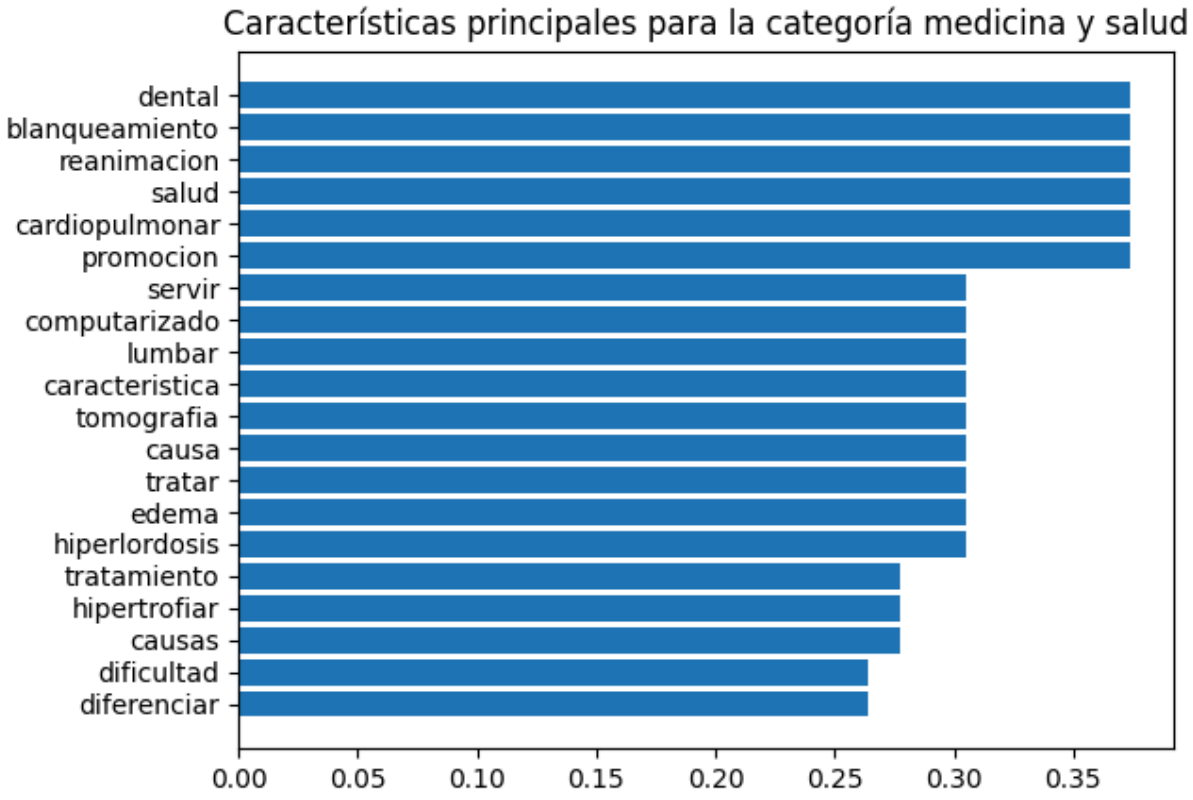
Se observa que para la frase '5 cosas importantes de saber a la hora de invertir tu dinero.' predice una categoría incorrecta, ya que se esperaba que fuera 'economia y finanzas'. Esto puede deberse a la calidad de los datos que se han usado para entrenar el modelo correspondientes a dicha categoría. También puede deberse a que la cantidad de registros de entrenamiento para ésa categoría no sea suficiente para detectar que la nueva frase pertenece a 'economia y finanzas'.

Para comprender mejor lo resaltado anteriormente, se analizará cuáles son las palabras que más influyen en el clasificador para definir una categoría determinada.









Se encuentra que las palabras invertir y dinero, que se hubiesen esperado sean las más relevantes para ubicar a la frase '5 cosas importantes de saber a la hora de invertir tu dinero.' en la categoría 'economia y finanzas' no se encuentran presentes en el gráfico de características más influyentes para dicha categoría. El modelo halló similitudes para dicha frase con la categoría 'medicina y salud'.

Lo mismo sucedió con la frase 'La Inteligencia Artifial es genial!'. En este caso, la frase no pertenecería a ninguna categoría. No obstante, el modelo nuevamente encuentra similitud con la categoría 'medicina y salud'.

Estos resultados dan lugar a la reflexión de que, con las librerías y herramientas utilizadas en este caso, no solo el conjunto de entrenamiento para la categoría 'economia y finanzas' carece de ciertas palabras relevantes para la clasificación sino que la posibilidad de que la categoría 'medicina y salud' cuente con características que sean más generalizables a otras frases y categorías acompaña a la obtención de resultados erróneos en algunos casos.

3.6. Vectorización BERT (embeddings)

Se procederá a continuación a realizar otro modelo que utilice embeddings.

Como de esta forma se captura la similitud semántica de los textos, se espera que mejore el desempeño al momento de clasificar nuevos títulos.

3.7. Entrenamiento del modelo y análisis de métricas

```
Precisión Regresión Logística: 0.625
Reporte de clasificación Regresión Logística:
              precision    recall  f1-score   support

           0       0.50      1.00      0.67         1
           1       1.00      0.50      0.67         4
           2       0.33      0.50      0.40         2
           3       1.00      1.00      1.00         1

    accuracy                           0.62         8
   macro avg       0.71      0.75      0.68         8
weighted avg       0.77      0.62      0.64         8
```

Se realizan las siguientes observaciones:

La clase 0 ('economia y finanzas') tiene un recall del 1.00, lo que indica un rendimiento excelente en la identificación de esta clase. Sin embargo, la precisión es baja (0.50), lo que sugiere que el modelo puede estar clasificando incorrectamente algunas instancias de otras clases como esta.

La clase 1 ('entretenimiento') tiene un f1-score bajo debido a un bajo valor de recall, lo que indica que el modelo tiene dificultades para identificar correctamente esta clase. Lo mismo sucede con la clase 2 ('fitness y deporte').

La clase 3 ('medicina y salud') tiene un rendimiento perfecto en todas las métricas, lo que indica que el modelo es muy bueno para identificar esta clase.

3.8. Clasificación de nuevos textos

La frase 'Los 10 juegos de atención, estrategia y cálculo para tener diversión en familia.' pertenece a la categoría: entretenimiento

La frase 'Es recomendable realizar ejercicio por lo menos 1 hora dos veces a la semana.' pertenece a la categoría: medicina y salud
La frase 'La Inteligencia Artifial es genial!' pertenece a la categoría: medicina y salud
La frase '5 cosas importantes de saber a la hora de invertir tu dinero.' pertenece a la categoría: economia y finanzas
La frase 'Conocer sobre primeros auxilos, es importante para saber cómo cuidar nuestra salud.' pertenece a la categoría: fitness y deporte

Se encuentra que la nueva frase que anteriormente era clasificada incorrectamente por el modelo vectorizado con TF-IDF ahora se halla bien clasificada. Sin embargo, 2 títulos son etiquetados con la categoría errónea. Es decir, obtuvimos más cantidad de títulos categorizados de manera equívoca.

La frase "Es recomendable realizar ejercicio por lo menos 1 hora dos veces a la semana.", en los embeddings obtenidos con BERT se posiciona más cercana a títulos de entrenamiento para la categoría 'medicina y salud'. Si bien no es lo que se esperaba, no es completamente erróneo ya que realizar ejercicio favorece la salud física. Se entiende entonces que el modelo encuentra cierta asociación semántica entre realizar ejercicio con la salud. Se destaca este detalle ya que el modelo podría haberlo categorizado como 'economía y finanzas' o 'entretenimiento', pero halló una categoría más acorde para el título.

A su vez, la frase "Conocer sobre primeros auxilos, es importante para saber cómo cuidar nuestra salud." fue etiquetada como 'fitness y deporte'. Con esta observación y la anterior, surge la idea de que posiblemente el modelo encuentre a las categorías 'medicina y salud' y 'fitness y deporte' mayor asociadas semánticamente que las demás.

"La Inteligencia Artifial es genial!" es nuevamente clasificada como 'medicina y salud'. Se sostiene la posibilidad de que los embeddings generados para dicha categoría sean más generalizables a nuevos datos.

3.9. Comparación de métricas y conclusiones

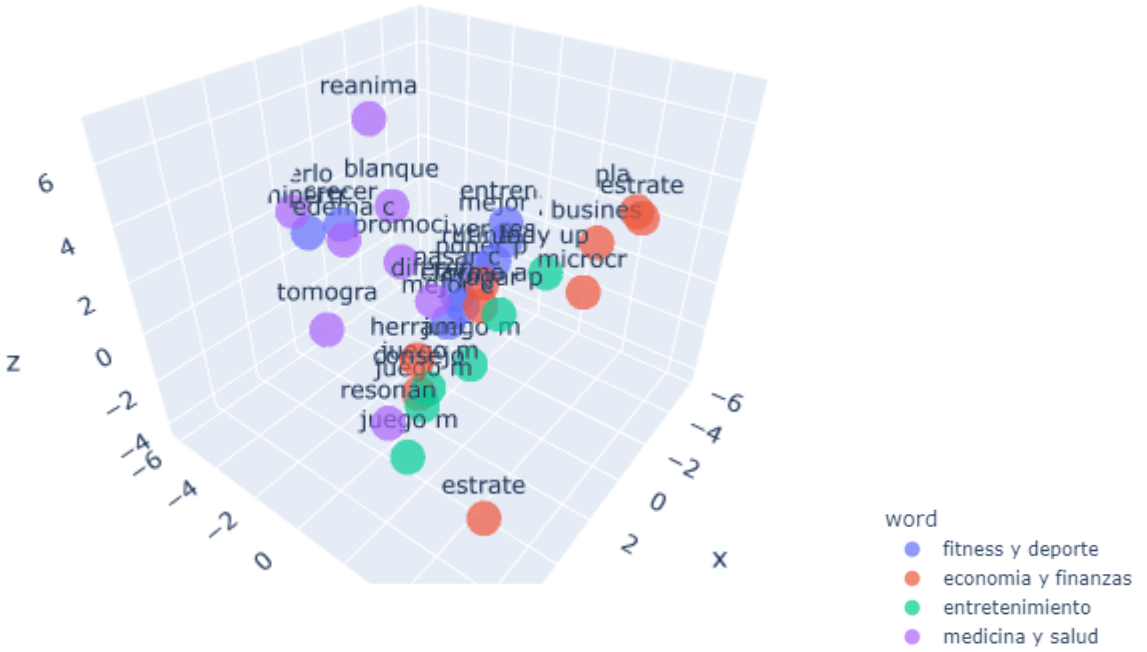
Ambos modelos tienen un rendimiento similar en términos de precisión y f1-score. Sin embargo, el modelo con BERT tiene un mejor rendimiento en la identificación de la clase 3 ('medicina y salud') en términos de l, mientras que el modelo con TF-IDF tiene un mejor rendimiento en la identificación de la clase 0 ('economia y finanzas') en términos de f1-score.

En el modelo con TF-IDF, la precisión es perfecta para tres de las clases, pero es baja (0.25) para la clase 3 ('medicina y salud'). Esto indica que el modelo tiene dificultades para predecir correctamente esta clase.
En el modelo con BERT, la precisión es alta para la mayoría de las clases, pero es baja (0.33) para la clase 2 ('fitness y deporte') y 0.50 para la clase 0 ('economia y finanzas). Esto indica que el modelo tiene dificultades para predecir correctamente estas clases.

En líneas generales el modelo para el que se usó TF-IDF sería mejor clasificando la clase 'economía y finanzas' y relativamente malo para la categorización de 'medicina y salud'. Se había observado que para nuevos títulos, el modelo clasificó uno correspondiente a economía y finanzas como medicina y salud, evidenciando la baja precisión para esta última categoría, donde es muy factible encontrar falsos positivos.

Y, con el modelo en el que se ha utilizado BERT se encuentra que para nuevos títulos ha clasificado erróneamente como fitness y deporte cuando correspondía medicina y salud, evidenciando la baja precisión para la primera categoría, donde es muy factible encontrar falsos positivos.

Gráfica tridimensional de los datos de training usando PCA:



Mediante el gráfico se puede observar que los títulos de entrenamiento para las categorías fitness y deporte y medicina y salud se encuentran más próximos entre ellos en el espacio vectorial que con el resto de las categorías. Esta visualización sugiere un apoyo a la hipótesis de asociación semántica del modelo para dichas categorías.

4. Ejercicio 3

4.1. Comparación de métricas y conclusiones

Para cada categoría, realizar las siguientes tareas:
- Procesar el texto mediante recursos de normalización y limpieza.
- Con el resultado anterior, realizar conteo de palabras y mostrar la importancia de las mismas mediante una nube de palabras.

Escribir un análisis general del resultado obtenido.

4.2. Procesamiento de los datos

- Conversión a minúsculas:

La conversión a minúsculas ayuda a homogeneizar el texto, evitando que las palabras escritas en mayúsculas se interpreten como diferentes de las escritas en minúsculas.

- Eliminación de acentos:
  La eliminación de acentos simplifica el texto y evita posibles problemas de codificación que podrían surgir con caracteres acentuados.

- Eliminación de puntuación:
  La eliminación de puntuación permite centrarse en las palabras y su significado, sin verse distraído por signos de puntuación que no aportan información semántica.

- Remover símbolos innecesarios:
  La eliminación de símbolos innecesarios es similar a la eliminación de puntuación, se enfoca en quitar caracteres que no aportan información relevante para el análisis de texto.
- Remover números en títulos:
  Se remueven los números porque no son caracteres específicos de una categoría que puedan ayudar a indentificarla de manera más precisa. Las cuatro categorías utilizadas pueden presentar números en sus títulos para representar diferentes significados.
- Remover stopwords:
  Las stopwords son palabras comunes como "el", "la", "y", etc., que no aportan mucha información en términos de contexto y significado. Se eliminan porque ayuda a reducir la dimensionalidad del texto y a concentrarse en las palabras clave.
- Tratamiento de abreviaturas:
  El tratamiento de abreviaturas desambigua el texto al expandir las abreviaturas a sus formas completas, permitiendo una interpretación más precisa. Además puede que haya palabras en su forma abreviada y expandida y tratando las abreviaturas se consigue homogeneizar las palabras y que no dé lugar a que se interpreten como palabras distintas.

- Corrección de ortografía:
  Se aplica corrección de ortografía ya que mejora la calidad del texto al corregir posibles errores tipográficos o de escritura, lo que facilita su interpretación y análisis.

- Lematización:
  La lematización reduce las palabras a sus formas base o lemas, lo que ayuda a agrupar las diferentes formas de una palabra en una única representación. Esto simplifica el análisis y la extracción de información.

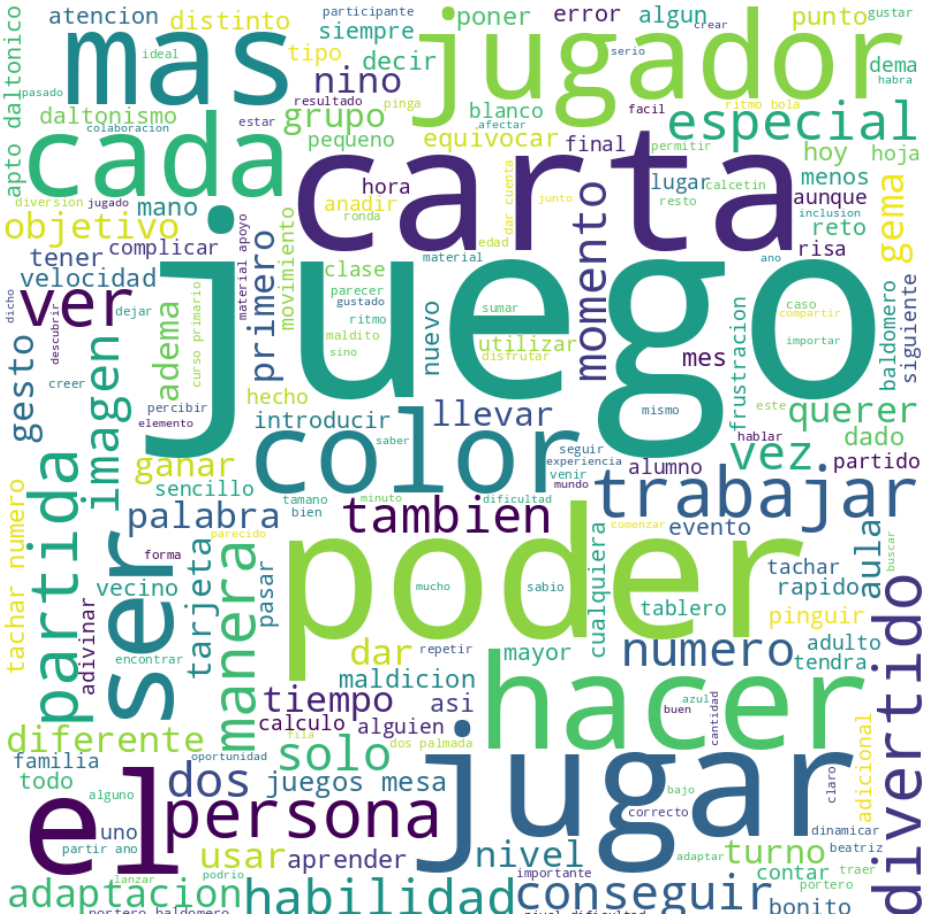
- Tratamiento de emojis:
  Se buscará la presencia de emojis y en caso de hallarlos se elimnarán para centrarse en las palabras. En este caso no se pretende hacer un análisis de sentimientos, sino una clasificación por lo que se prescinde de los emojis.

4.3. Frecuencia de palabras

```
Diccionario de frecuencias para las categorías de las noticias
{   'economia y finanzas': {   'abanico': 1,
                               'abarca': 3,
                               'abarcar': 1,
                               'abarco': 1,
                               'abordarer': 1,
                               'abracer': 1,
                               'abrazar': 2,
                               'abrazar el': 1,
                               'abrir': 2,
                               'absolutamente': 3,
                               'acabar': 4,
                               'acabe': 1,
                               'academico': 1,
                               'acceder': 4,
                               'accesibl': 1,
                               'accesible': 1,
                               'acceso': 7,
                               'accesorio': 1,
                               'accion': 7,
                               'aceptable': 1,
                               'aceptar': 1,
                               'acercamiento': 1,
                               'acercar': 1,
                               'acertar': 1,
...
                           'zona': 9,
                           'zonaasimismo': 1,
                           'zonar': 1,
                           'zoologico': 2}}
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Categoría economia y finanzas:

Categoría entretenimiento:



Categoría fitness y deporte:

Categoría medicina y salud:

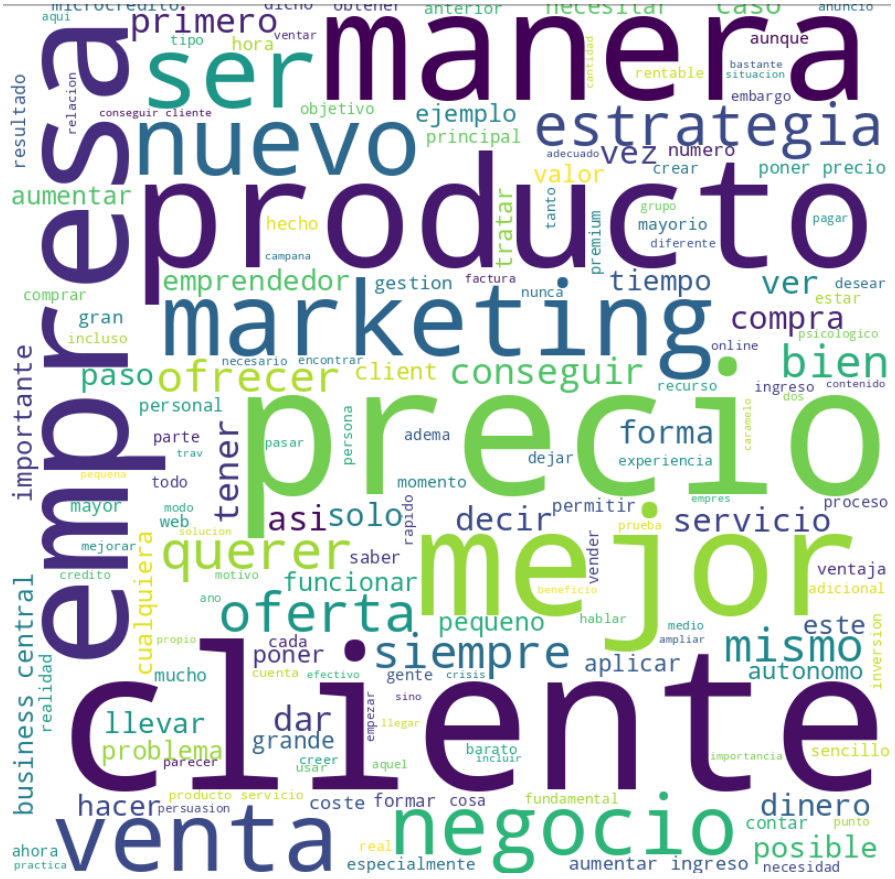
Se observa que existen palabras que permiten distinguir claramente la pertenencia a una categoría específica y son palabras con considerablemente destacadas.

No obstante, hay ciertas palabras que se advierte que presentan una frecuencia elevada en los textos: 'poder' y 'mas'. Las palabras "poder" y "más" ocupan mucho espacio en la nube de palabras en todas las categorías, lo que sugiere que son términos muy comunes en tu conjunto de noticias.

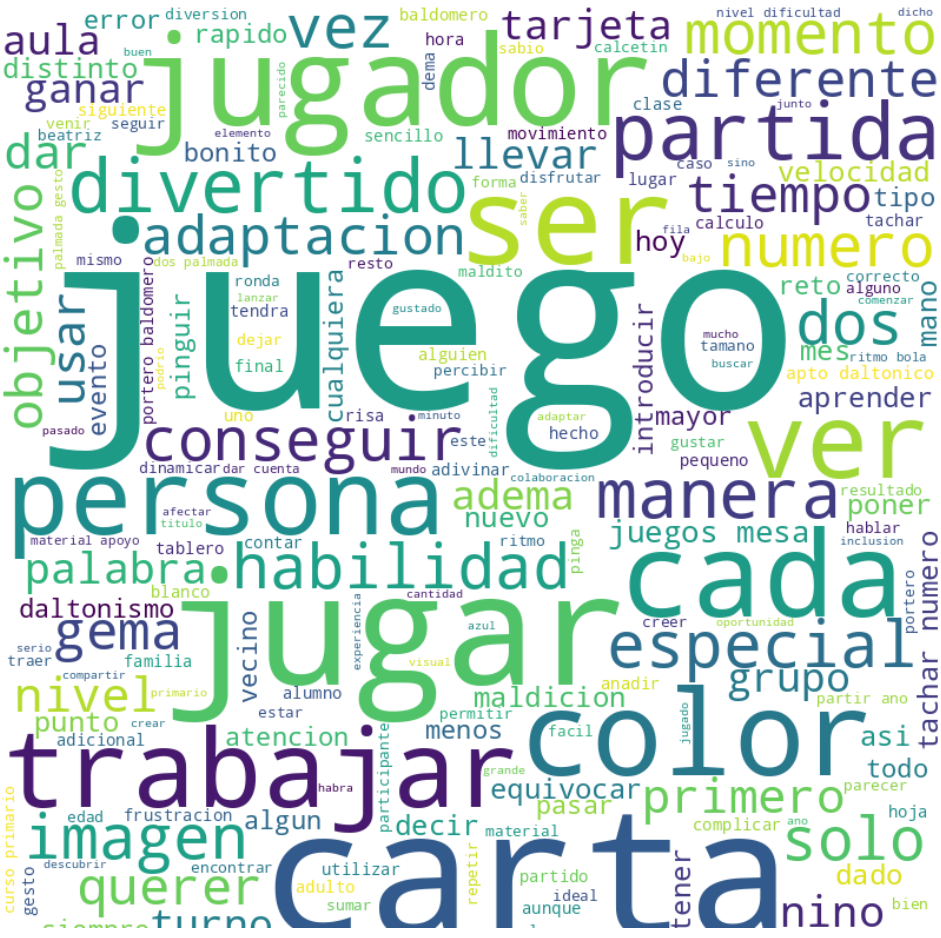
Dado que estas palabras no proporcionan una discriminación clara entre las categorías y no aportan información relevante para distinguir el contenido de las noticias, se considera beneficioso eliminarlas antes de realizar cualquier análisis de texto o clasificación. Es decir, dichas palabras serán consideradas stop words, ya que no aportan a la diferenciación nítida de los contenidos por categoría, lo que permitirá enfocarse en términos más informativos y relevantes para la clasificación.

En resumen, se considera que eliminar las palabras "poder" y "más" de las noticias podría ayudar a mejorar la calidad y relevancia de las características utilizadas para una eventual clasificación, lo que posiblemente conduzca a un mejor rendimiento en los modelos de clasificación.

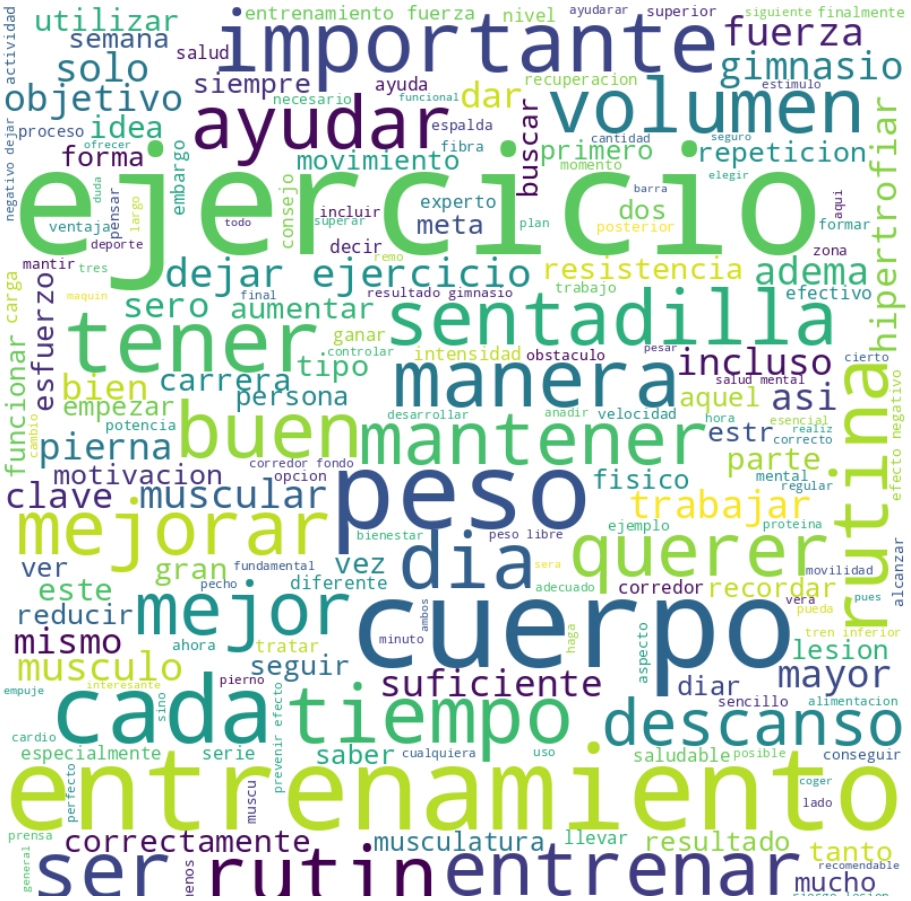
Categoría economia y finanzas luego de reacondicionamiento:



Categoría entretenimiento luego de reacondicionamiento:



Categoría fitness y deporte luego de reacondicionamiento:



Categoría medicina y salud luego de reacondicionamiento:

Se ha conseguido que las palabras que aportan información relevante y discriminativa para cada categoría sean de las más destacadas en los textos, viéndose reflejado en los gráficos.

4.4. Observaciones y conclusiones

Es oportuno destacar, luego del extenso proceso de acondicionamiento de los datos, que al observar en detalle las nuevas nubes de palabras, se pueden detectar nuevamente elementos que podrían considerarse stopwords. Es decir, el proceso de acondicionamiento de los datos es inexorablemente iterativo y cuan exhaustivo sea dependerá de los objetivos a alcanzar.

5. Ejercicio 4

Use los modelos de embedding propuestos sobre el final de la Unidad 2 para evaluar la similitud entre los títulos de las noticias de una de las categorías.

Reflexione sobre las limitaciones del modelo en base a los resultados obtenidos, en contraposición a los resultados que hubiera esperado obtener

5.1. Similaridad Doc2Vec

Para evaluar la similaridad entre títulos se hará uso de la librería Doc2Vec que es una extensión de Word2Vec pero para utilizar sobre oraciones o documentos completos.

Como los títulos son oraciones y queremos que se establezcan similaridades entre los títulos completos y no entre palabras aisladas, Doc2Vec parece adecuado para tal fin.

El título con mayor similitud para "" ¿Qué es mejor, prensa o sentadilla?"" es: "" La clave del éxito para entrenar: motivación, disciplina y perseverancia"" con una similitud de ""0.6796526908874512"".
El título con menor similitud para "" ¿Qué es mejor, prensa o sentadilla?"" es: "" ¿Qué le pasa a tu cuerpo cuando dejas de hacer ejercicio?"" con una similitud de ""0.470191091299057"".

El título con mayor similitud para "" La clave del éxito para entrenar: motivación, disciplina y perseverancia"" es: "" ¿Qué es mejor, prensa o sentadilla?"" con una similitud de ""0.6796526908874512"".
El título con menor similitud para "" La clave del éxito para entrenar: motivación, disciplina y perseverancia"" es: "" ¿Por qué no veo resultados en el gym? Estas son las razones"" con una similitud de ""0.31812843680381775"".

El título con mayor similitud para "" ¿Qué le pasa a tu cuerpo cuando dejas de hacer ejercicio?"" es: "" Cómo recuperarte tras un ultra trail"" con una similitud de ""0.7073678970336914"".

El título con menor similitud para "" ¿Qué le pasa a tu cuerpo cuando dejas de hacer ejercicio?"" es: "" Top 5 ejercicios de bajo impacto para entrenar"" con una similitud de ""0.3773544430732727"".

El título con mayor similitud para "" Los mejores ejercicios de fuerza para corredores de fondo"" es: "" Cómo entrenar con pesas sin ganar volumen"" con una similitud de ""0.72984379529953"".
El título con menor similitud para "" Los mejores ejercicios de fuerza para corredores de fondo"" es: "" ¿Qué le pasa a tu cuerpo cuando dejas de hacer ejercicio?"" con una similitud de ""0.45306795835494995"".

El título con mayor similitud para "" Cómo entrenar con pesas sin ganar volumen"" es: "" Los mejores ejercicios de fuerza para corredores de fondo"" con una similitud de ""0.7298436760902405"".
El título con menor similitud para "" Cómo entrenar con pesas sin ganar volumen"" es: "" ¿Qué le pasa a tu cuerpo cuando dejas de hacer ejercicio?"" con una similitud de ""0.44622981548309326"".

El título con mayor similitud para "" Cómo recuperarte tras un ultra trail"" es: "" ¿Qué le pasa a tu cuerpo cuando dejas de hacer ejercicio?"" con una similitud de ""0.7073679566383362"".
El título con menor similitud para "" Cómo recuperarte tras un ultra trail"" es: "" ¿Cómo crecen los músculos? | Explicación de la hipertrofia muscular"" con una similitud de ""0.4784499406814575"".

El título con mayor similitud para "" Top 5 ejercicios de bajo impacto para entrenar"" es: "" Rutinas para entrenar cuando tienes poco tiempo"" con una similitud de ""0.7863970398902893"".
El título con menor similitud para "" Top 5 ejercicios de bajo impacto para entrenar"" es: "" ¿Qué le pasa a tu cuerpo cuando dejas de hacer ejercicio?"" con una similitud de ""0.3773544132709503"".

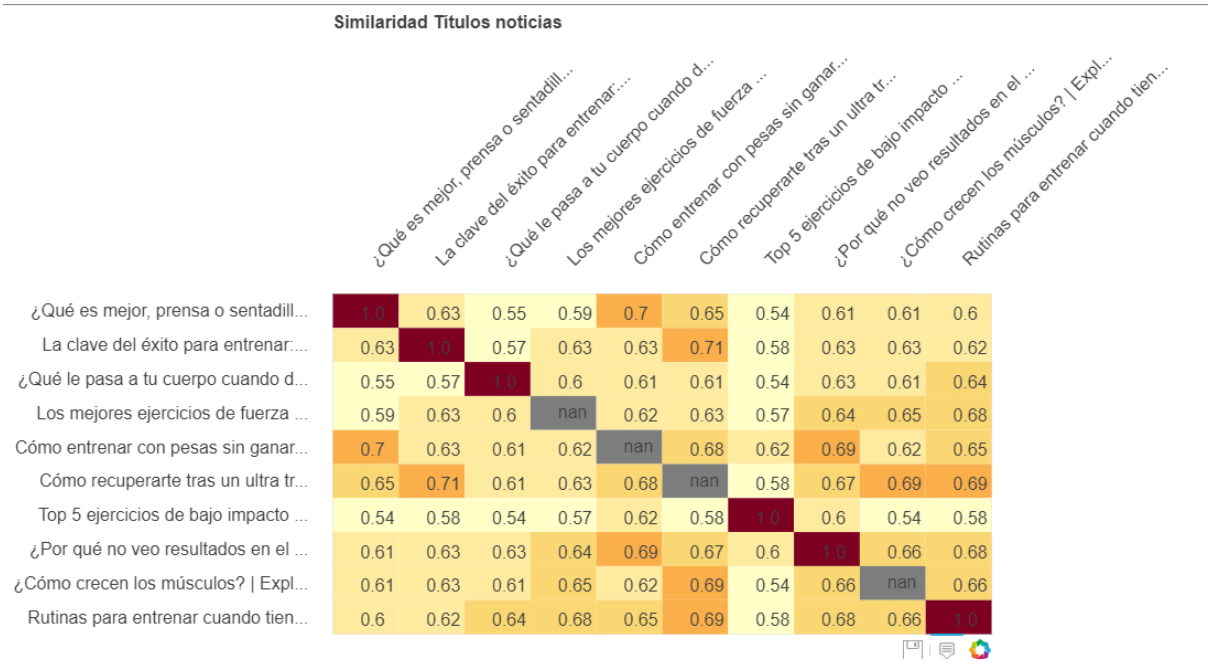
El título con mayor similitud para "" ¿Por qué no veo resultados en el gym? Estas son las razones"" es: "" Rutinas para entrenar cuando tienes poco tiempo"" con una similitud de ""0.7279905676841736"".
El título con menor similitud para "" ¿Por qué no veo resultados en el gym? Estas son las razones"" es: "" La clave del éxito para entrenar: motivación, disciplina y perseverancia"" con una similitud de ""0.31812840700149536"".

El título con mayor similitud para "" ¿Cómo crecen los músculos? | Explicación de la hipertrofia muscular"" es: "" Top 5 ejercicios de bajo impacto para entrenar"" con una similitud de ""0.6552886962890625"".
El título con menor similitud para "" ¿Cómo crecen los músculos? | Explicación de la hipertrofia muscular"" es: "" Cómo recuperarte tras un ultra trail"" con una similitud de ""0.4784500002861023"".

El título con mayor similitud para "" Rutinas para entrenar cuando tienes poco tiempo"" es: "" Top 5 ejercicios de bajo impacto para entrenar"" con una similitud de ""0.7863970994949341"".
El título con menor similitud para "" Rutinas para entrenar cuando tienes poco tiempo"" es: "" ¿Qué es mejor, prensa o sentadilla?"" con una similitud de ""0.5432701110839844"".

5.1. Similaridad Universal Sentence Encoding

Se observarán ahora las similaridades obtenidas entre los títulos utilizando un modelo de encoding que captura el significado semántico de lo que representa. Tal vez, bajo esta consideración, se robustece la interpretación de la similitud entre títulos.



5.3. Observaciones y conclusiones

Se observa que los niveles de mayor similitud hallada con Doc2Vec entre los títulos es razonablemente alta. En algunos casos llegando hasta casi el 80%. Hay títulos que representan la mayor similaridad con más de un título. Se entiende que en el espacio vectorial que los representa, dichos títulos se encuentran muy cercanos con los demás. Es decir, el ángulo entre dichos vectores es de los menores encontrados.
Se esperaba hallar resultados de estas características para las similaridades más altas. No se aspiraba a similaridades que lleguen a 1 ya que para ello los títulos deberían ser exactamente iguales y, si bien corresponden a la misma temática, dentro de ella se pueden hallar diversos subtemas.

En cuanto a las menores similaridades halladas entre los títulos de las noticias se considera que continúan representando una similaridad bastante alta en algunos casos, (mayores a 0.4). El umbral más bajo ha sido 0.31.

Al utilizar Universal Sentence Encoding en esta oportunidad, el modelo ha detectado las similaridades más elevadas en pares de títulos distintos de los encontrados con el modelo

anterior. Se considera que esta diferencia de similitudes encuentra su explicación en que el último modelo empleado captura el significado semántico de las oraciones y el contexto en el que se utilizan las palabras.

Aún así, las similaridades encontradas no difieren grandemente en valor numérico a las halladas con el modelo anterior.

6. Ejercicio 5

Escriba un programa interactivo que, según la categoría seleccionada por el usuario, devuelva un resumen de las noticias incluidas en ella.

Justifique la elección del modelo usado para tal fin.

Opcional: Investigar y programar un bot de Telegram que entregue un resumen de noticias del blog de su elección. Recomendamos el uso de pyTelegramBotAPI.

6.1. Elección de enfoque para el resumen

Dado el contexto, se utilizará un enfoque extractivo para generar resúmenes de las noticias. Las razones que acompañan la elección de dicho enfoque son las siguientes:

- Mayor control sobre la precisión:

Los modelos extractivos tienden a ser más precisos en la selección de la información relevante del texto original. Esto es especialmente importante en este caso porque, al tratarse de noticias para informar al lector, se necesitan resúmenes precisos y fieles al contenido original.

- Facilita la evaluación:

Como los resúmenes extractivos seleccionan directamente fragmentos del texto original, es más fácil evaluar la calidad del resumen, ya que se puede comparar directamente con el texto fuente.

- Conservación de la integridad del texto:

Los resúmenes extractivos mantienen la coherencia y cohesión del texto original, ya que utilizan partes del texto existente. Esto es beneficioso para mantener la fidelidad al contenido original. Dentro de las categorías elegidas se encuentra la de medicina y salud. Si bien conservar la integridad es aplicable a todas las categorías, la mencionada anteriormente se considera más sensible debido a que puede tener repercusiones más pronunciadas en las personas. Por ello, se considera fundamental conservar la integridad.

- Menos probabilidades de generar información incorrecta o engañosa:

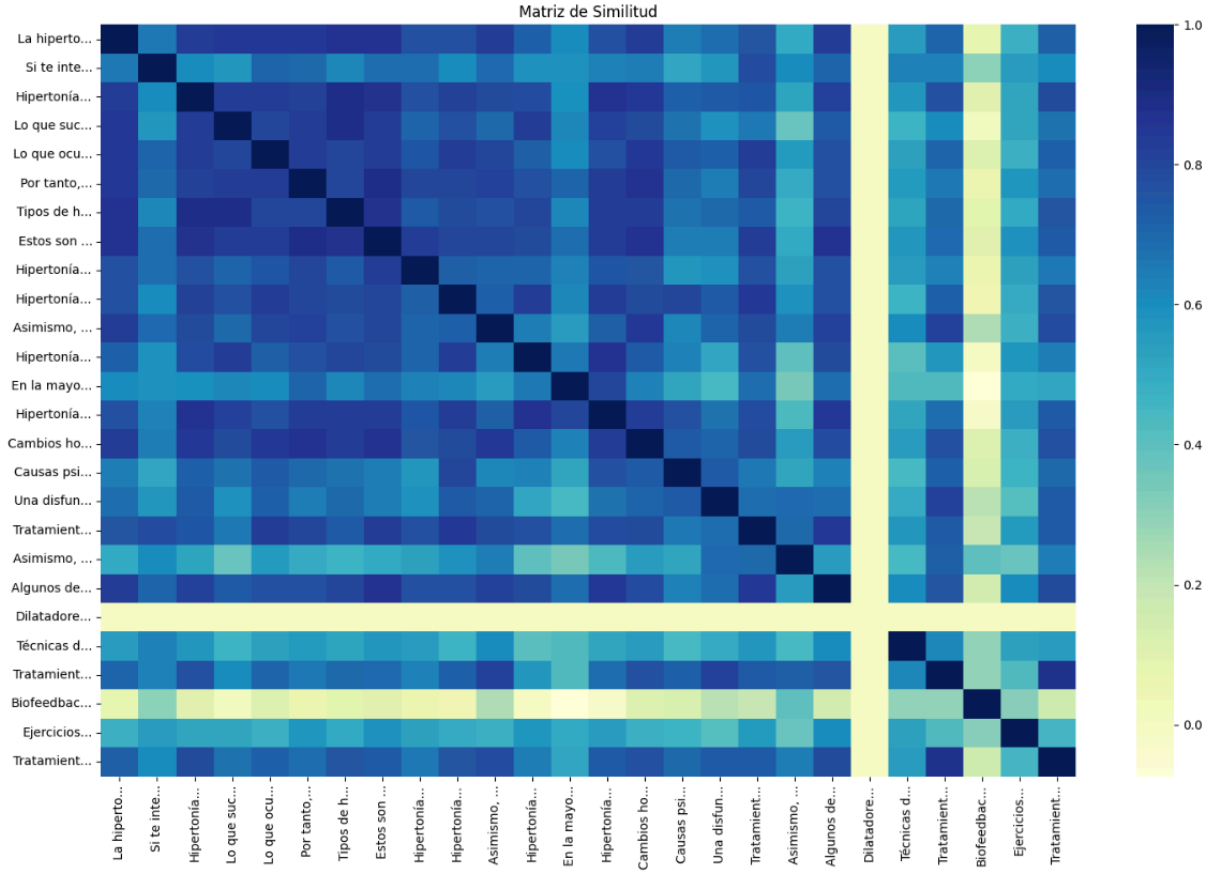
Los modelos abstractivos a veces pueden generar información que no estaba presente en el texto original, lo que podría llevar a interpretaciones incorrectas o engañosas. Los extractivos evitan este riesgo.

Es cierto que los resúmenes extractivos pueden ser menos creativos y no generarán resúmenes completamente nuevos. Sin embargo, en el contexto de noticias, donde la precisión y la fidelidad al contenido original son cruciales, el enfoque extractivo ha sido la opción preferida.

Se realiza un tratamiento pertinente al contenido original de las noticias. El mismo consiste en la remoción de símbolos que se ha observado (por iteraciones previas) introducen una incorrecta segmentación del texto y, por lo tanto, de los resúmenes producidos.

6.2. Matrices de similitud para cada noticia

Dado que la cantidad de noticias es muy elevada para agregar todos los gráficos de matrices en el informe y que, además, las mismas gráficas se hallan en el archivo .ipynb se agregará solo una de ellas a modo de ejemplo para dar una idea de los resultados.



Resumen extractivo obtenido:

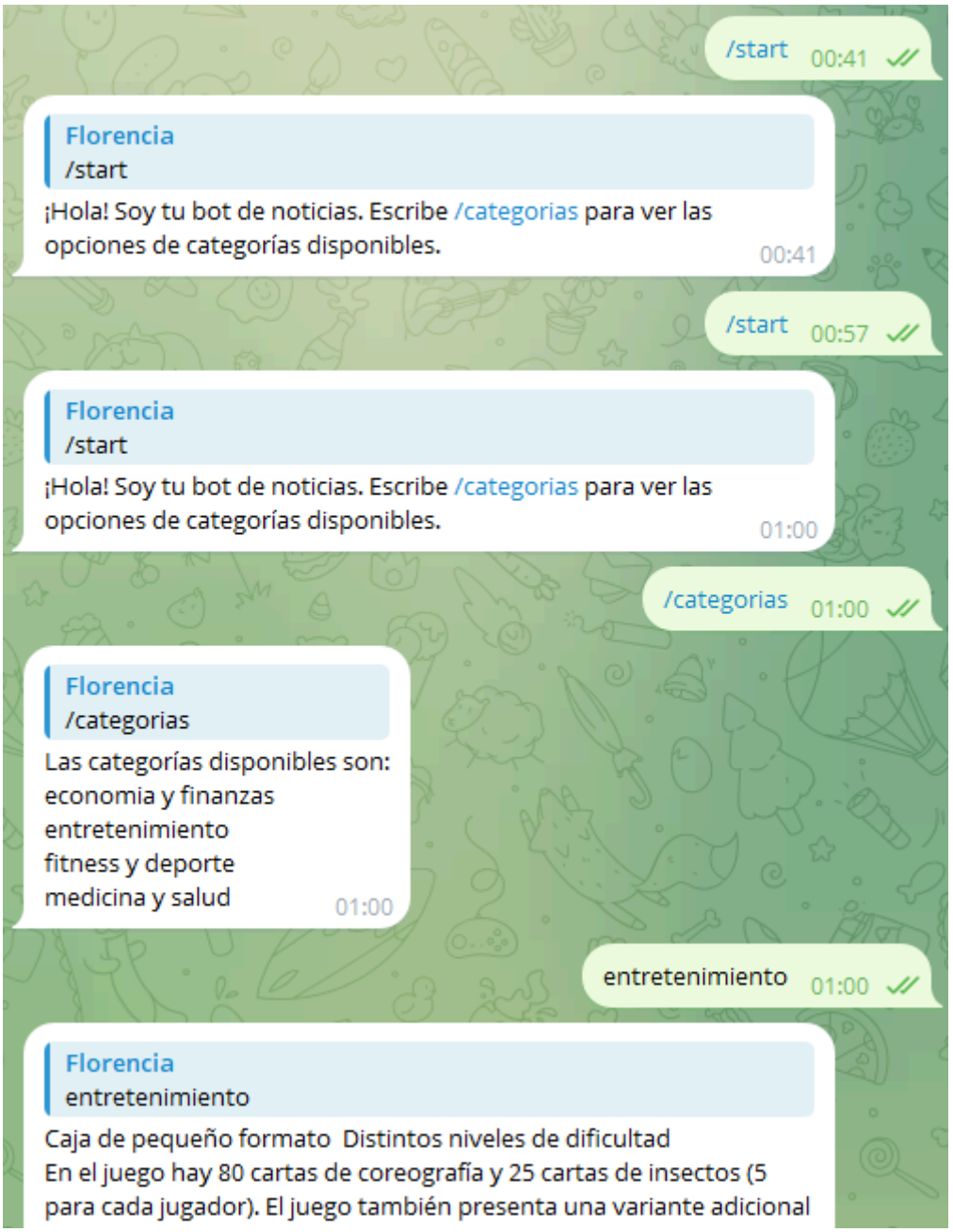
Estos son los tipos de hipertonía muscular que existen:Hipertonía elástica: el músculo movilizado vuelve a su posición de origen cuando se deja de hacer el movimiento de forma manual. Algunos de los tratamientos más recurrentes utilizados para tratar la hipertonía del suelo pélvico encontramos:Masajes perineales. Tratamiento hipertonía suelo pélvicoAl tratarse de una afección que puede tener diferentes orígenes, se puede atender desde diferentes ramas de la medicina, aunque como primera opción, siempre se recomienda acudir al fisioterapeuta especializado en suelo pélvico y aplicar técnicas que no sean invasivas.

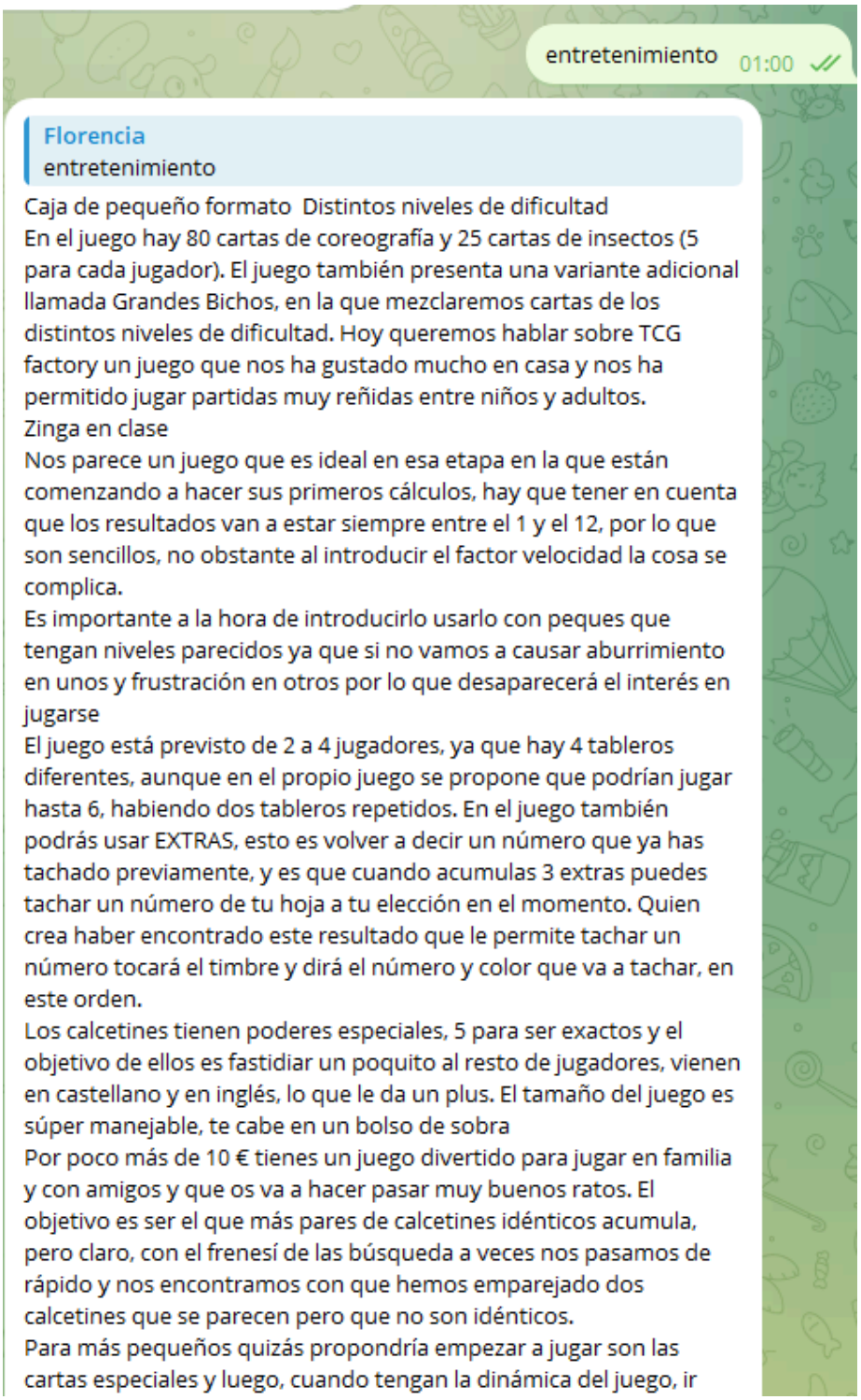
6.3. Observaciones y conclusiones

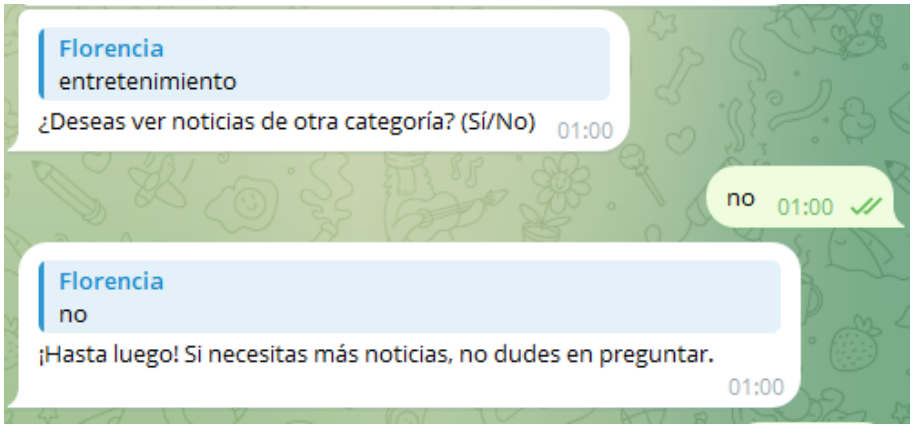
En general, para cada noticia, se observa en las matrices la existencia de oraciones con un elevado registro de similitud. Esto es positivo ya que nos da indicios de la presencia de un texto coherente que puede ser representado destacando unas pocas oraciones. Se han elegido 3 oraciones por noticia para la confección de los resúmenes ya que algunos textos son muy largos y lo que se busca es informar el contenido más destacado de cada noticia y no abrumar al lector con excesivos detalles.

6.4. Opcional: programar un bot de Telegram

Resultados:




entretenimiento 01:00
Florencia
entretenimiento
Caja de pequeño formato Distintos niveles de dificultad
En el juego hay 80 cartas de coreografía y 25 cartas de insectos (5 para cada jugador). El juego también presenta una variante adicional llamada Grandes Bichos, en la que mezclaremos cartas de los distintos niveles de dificultad. Hoy queremos hablar sobre TCG factory un juego que nos ha gustado mucho en casa y nos ha permitido jugar partidas muy reñidas entre niños y adultos.
Zinga en clase
Nos parece un juego que es ideal en esa etapa en la que están comenzando a hacer sus primeros cálculos, hay que tener en cuenta que los resultados van a estar siempre entre el 1 y el 12, por lo que son sencillos, no obstante al introducir el factor velocidad la cosa se complica.
Es importante a la hora de introducirlo usarlo con peques que tengan niveles parecidos ya que si no vamos a causar aburrimiento en unos y frustración en otros por lo que desaparecerá el interés en jugarse
El juego está previsto de 2 a 4 jugadores, ya que hay 4 tableros diferentes, aunque en el propio juego se propone que podrían jugar hasta 6, habiendo dos tableros repetidos. En el juego también podrás usar EXTRAS, esto es volver a decir un número que ya has tachado previamente, y es que cuando acumulas 3 extras puedes tachar un número de tu hoja a tu elección en el momento. Quien crea haber encontrado este resultado que le permite tachar un número tocará el timbre y dirá el número y color que va a tachar, en este orden.
Los calcetines tienen poderes especiales, 5 para ser exactos y el objetivo de ellos es fastidiar un poquito al resto de jugadores, vienen en castellano y en inglés, lo que le da un plus. El tamaño del juego es súper manejable, te cabe en un bolso de sobra
Por poco más de 10 € tienes un juego divertido para jugar en familia y con amigos y que os va a hacer pasar muy buenos ratos. El objetivo es ser el que más pares de calcetines idénticos acumula, pero claro, con el frenesí de las búsqueda a veces nos pasamos de rápido y nos encontramos con que hemos emparejado dos calcetines que se parecen pero que no son idénticos.
Para más pequeños quizás propondría empezar a jugar son las cartas especiales y luego, cuando tengan la dinámica del juego, ir


Florencia
entretenimiento
¿Deseas ver noticias de otra categoría? (Sí/No) 01:00
no 01:00
Florencia
no
¡Hasta luego! Si necesitas más noticias, no dudes en preguntar.
01:00