eda

October 12, 2025

1 EDA Dataset Análisis de las Sociedades Argentinas

[67]:
```
import os
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sqlalchemy import create_engine
from dotenv import load_dotenv
```

Creamos la conexión a la BD.

[10]:
```
load_dotenv()

DB_USER = os.getenv("MYSQL_USER")
DB_PASSWORD = os.getenv("MYSQL_PASSWORD")
DB_HOST = os.getenv("DB_HOST")
DB_NAME = os.getenv("MYSQL_DATABASE")

DATABASE_URL = f"mysql+mysqlconnector://{DB_USER}:{DB_PASSWORD}@{DB_HOST}/
 ↪{DB_NAME}"

engine = create_engine(DATABASE_URL)
connection = engine.connect()
```

Leemos todos los avisos disponibles.

[11]:
```
query = "SELECT * from avisos;"
df = pd.read_sql(query, connection)
df.head()
```

[11]:

	id	aviso_id	seccion
0	1	A1	Segunda sección
1	2	A2	Segunda sección
2	3	A3	Segunda sección
3	4	A4	Segunda sección
4	5	A5	Segunda sección

```
                                          sociedad  \
0         JUZGADOS NACIONALES \n EN LO CIVIL\n N° 42
1  JUZGADO NACIONAL EN LO \n CIVIL NRO. 3 \n SECR…
2         JUZGADOS NACIONALES \n EN LO CIVIL\n N° 19
3                                           N° 82
4                                           Nº 19

                                          rubro id_rubro  \
0  CITACIONES Y NOTIFICACIONES. CONCURSOS Y QUIEB…     3100
1  CITACIONES Y NOTIFICACIONES. CONCURSOS Y QUIEB…     3100
2  CITACIONES Y NOTIFICACIONES. CONCURSOS Y QUIEB…     3100
3  CITACIONES Y NOTIFICACIONES. CONCURSOS Y QUIEB…     3100
4  CITACIONES Y NOTIFICACIONES. CONCURSOS Y QUIEB…     3100

  fecha_publicacion                                    detalle_aviso  \
0        2011-01-03  El Juzgado Nacional de Primera Instancia N° 42…
1        2011-01-06  El Juzgado Nacional de Primera Instancia en lo…
2        2011-01-14  Juzgado Nacional de Primera Instancia en lo Ci…
3        2011-01-03  La Señora Juez Subrogante a cargo del Juzgado …
4        2010-12-03  El Juzgado Nacional de Primera Instancia en lo…

            crawled_at
0 2025-09-20 02:31:54
1 2025-09-20 02:31:56
2 2025-09-20 02:31:57
3 2025-09-20 02:31:58
4 2025-09-20 02:31:59
```

[12]:
```
print("Columnas: ", df.columns.to_list())
print("# de avisos:", len(df))
print("Tipos de Datos::\n", df.dtypes)
print("\n# de Valores únicos:\n", df.nunique())
```

```
Columnas:  ['id', 'aviso_id', 'seccion', 'sociedad', 'rubro', 'id_rubro',
'fecha_publicacion', 'detalle_aviso', 'crawled_at']
# de avisos: 1342069
Tipos de Datos::
 id                            int64
aviso_id                     object
seccion                      object
sociedad                     object
rubro                        object
id_rubro                     object
fecha_publicacion            object
detalle_aviso                object
crawled_at           datetime64[ns]
dtype: object
```

```
# de Valores únicos:
 id                   1342069
aviso_id              1342069
seccion                     1
sociedad               308939
rubro                      19
id_rubro                   19
fecha_publicacion        3544
detalle_aviso          940086
crawled_at              72586
dtype: int64
```

Convertimos la columna `id_rubro` a integer.

[13]:
```
print((df['id_rubro'] + ' - ' + df['rubro']).value_counts())
df['id_rubro'] = df['id_rubro'].astype(int)
```

```
2300 - AVISOS COMERCIALES                                          352521
3300 - SUCESIONES                                                  330253
2100 - CONVOCATORIAS                                               167840
3100 - CITACIONES Y NOTIFICACIONES. CONCURSOS Y QUIEBRAS. OTROS     157610
1210 - CONTRATO SRL                                                 93952
1110 - CONSTITUCION SA                                              60512
1120 - REFORMA SA                                                   60434
1220 - MODIFICACIONES SRL                                           43105
3400 - REMATES JUDICIALES                                           26228
4000 - PARTIDOS POLITICOS                                           14987
1130 - CONSTITUCION SAS                                             14620
2200 - TRANSF. FONDO DE COMERCIO                                    12657
2400 - REMATES COMERCIALES                                           3056
1420 - REFORMA OTRAS SOCIEDADES                                      1853
1320 - REFORMA SCA                                                   1068
2500 - BALANCES                                                       697
5000 - INFORMACION Y CULTURA                                          474
1410 - ESTATUTO OTRAS SOCIEDADES                                      190
1310 - ESTATUTO SCA                                                    12
Name: count, dtype: int64
```

Los **rubros clave** para la constitución de las Sociedades son: - 1110 - CONSTITUCION SA (Sociedad Anónima) - 1130 - CONSTITUCION SAS (Sociedad Anónima Simplificada)

La S.A. es una estructura tradicional, más rígida, costosa y compleja de constituir y administrar (requiere escritura pública, un mínimo de dos socios y órganos de gestión y fiscalización formales), la S.A.S. es un vehículo moderno diseñado para emprendedores que permite una creación rápida, económica y digital, puede ser constituida por un único socio y ofrece una enorme libertad para organizar su administración y funcionamiento de manera más ágil y con menos formalidades.

Definimos el método `plot_fecha_publicacion` para graficar las publicaciones por año.

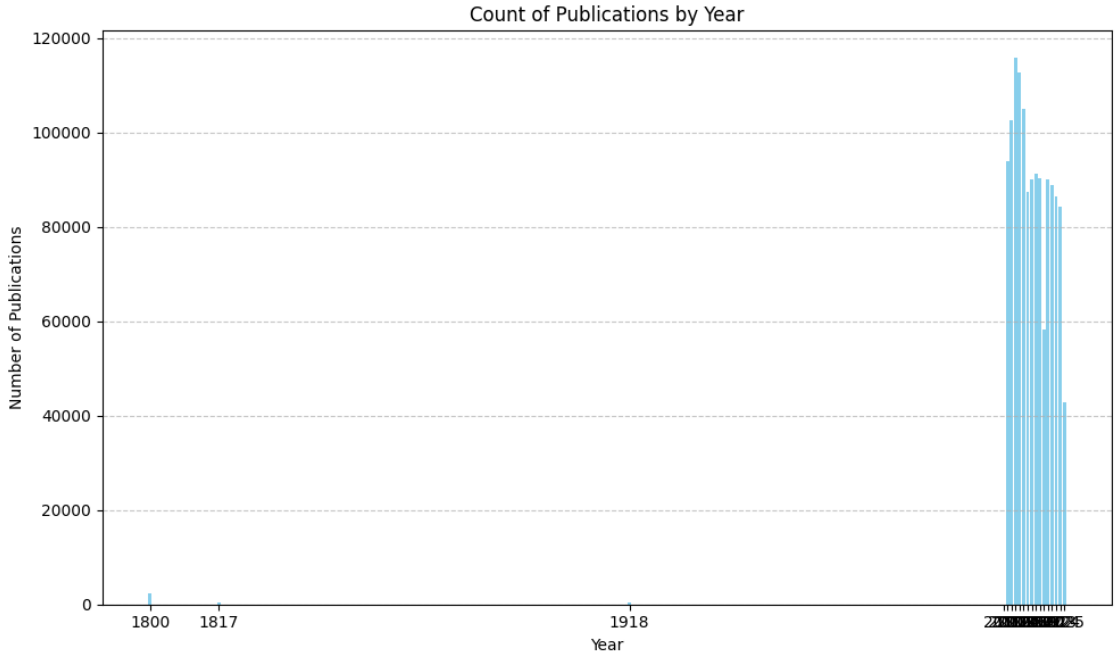
[14]:
```
def plot_fecha_publicacion(df):
    yearly_counts = df['year'].value_counts().sort_index()
    plt.figure(figsize=(10,6))
    plt.bar(yearly_counts.index, yearly_counts.values, color='skyblue')
    plt.xlabel('Year')
    plt.ylabel('Number of Publications')
    plt.title('Count of Publications by Year')
    plt.xticks(yearly_counts.index.astype(int)) # Ensure x-axis ticks are␣
 ↪integers
    plt.grid(axis='y', linestyle='--', alpha=0.7)
    plt.tight_layout()
```

Convertimos la columna `fecha_publicacion` a tipo datetime.

Al graficar las publicaciones por año, se detectaron 3011 registros con fecha incorrecta.

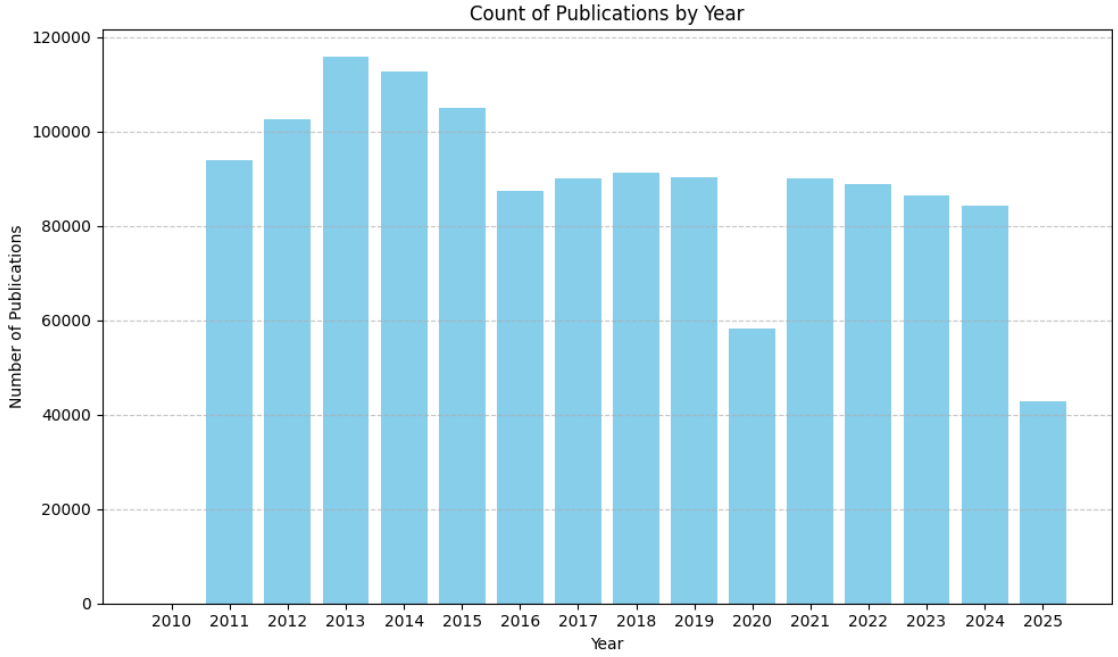
[15]:
```
df['fecha_publicacion'] = pd.to_datetime(df['fecha_publicacion'],␣
 ↪errors='coerce')
df.dropna(subset=['fecha_publicacion'], inplace=True)

df['year'] = df['fecha_publicacion'].dt.year

plot_fecha_publicacion(df)
print("Cantidad de entradas erróneas: ", len(df[df['year'] <␣
 ↪2000]['fecha_publicacion']))
print(df[df['year'] < 2000]['fecha_publicacion'].value_counts())
```

```
Cantidad de entradas erróneas:  3011
fecha_publicacion
1800-06-24    872
1800-10-24    466
1817-03-22    397
1800-01-01    389
1800-12-14    339
1918-06-12    338
1800-01-16    210
Name: count, dtype: int64
```



Eliminamos las entradas con año anterior al 2000.

[16]:
```
df = df[df['year'] > 2000]
plot_fecha_publicacion(df)
```

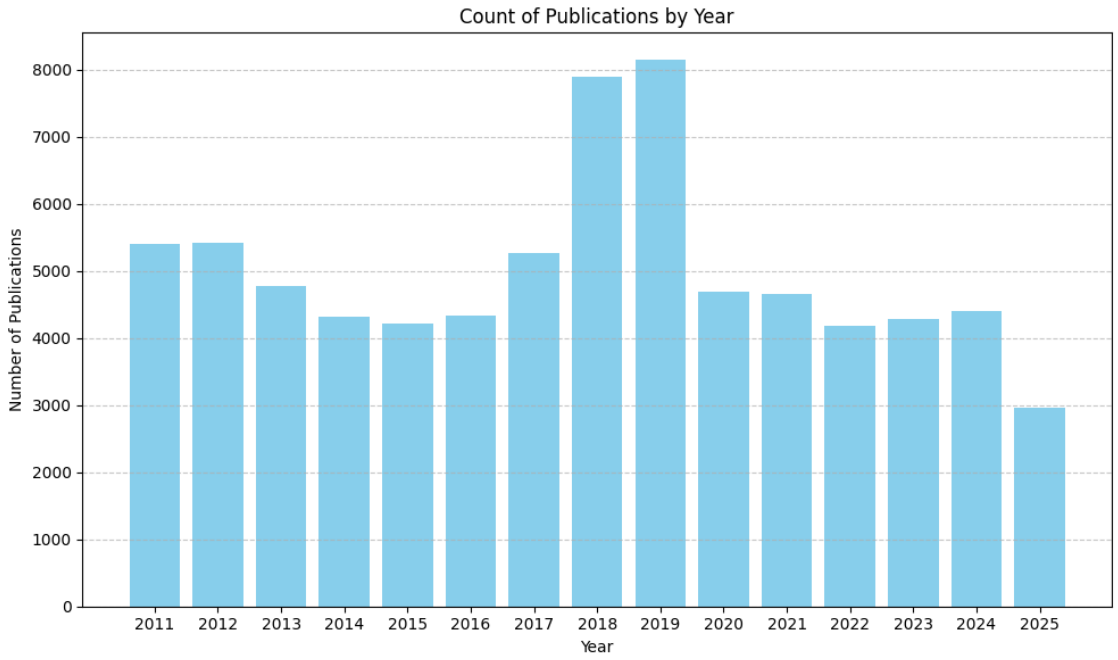


Definimos un nuevo DataFrame `df_constituciones` solo con las constituciones de Sociedades.

[17]:
```
df_constituciones = df[df['id_rubro'].isin([1110, 1130])]
print("# de Sociedades Constituidas: ", len(df_constituciones))
```

```
# de Sociedades Constituidas:  74930
```

[18]:
```
plot_fecha_publicacion(df_constituciones)
```



[19]:
```
df_constituciones.head()
```

[19]:

	id	aviso_id	seccion	sociedad	rubro	id_rubro	fecha_publicacion	detalle_aviso	crawled_at
320	321	A322	Segunda sección	TECNOLOGIA Y CABLEADOS	CONSTITUCION SA	1110	2011-01-03		
322	323	A324	Segunda sección	BOMBAS DE HORMIGON	CONSTITUCION SA	1110	2011-01-03		
324	325	A326	Segunda sección	BHRISA	CONSTITUCION SA	1110	2011-01-03		
343	344	A345	Segunda sección	BELFIL	CONSTITUCION SA	1110	2011-01-03		
366	367	A368	Segunda sección	GRUPO SAMIRA	CONSTITUCION SA	1110	2011-01-03		

```
320  Por Esc. 255 del 21/12.10, Carlos Alberto Koga… 2025-09-20 02:36:44
322  Por Esc. 247 del 16/12/10, Rodolfo Gontek, 27/… 2025-09-20 02:36:45
324  Agustín Rodolfo Spotorno, con DNI 22.501.187, … 2025-09-20 02:36:45
343  Carlos Enrique Bilevich, dni: 17029874 y Gabri… 2025-09-20 02:36:49
366  1) Rosela Beatriz Diaz, 18-7-79, DNI 27368918,… 2025-09-20 02:36:54

     year
320  2011
322  2011
324  2011
343  2011
366  2011
```

La columna `detalle_aviso` contiene la información más relevante para el análisis. Para poder encontrar sus características, es necesario procesar la columna para convertirla a un formato estructurado.

Todas las sociedades deben cumplir con el Artículo 11 de la Ley 19.550.

Este es un ejemplo de una constitución, pero pueden escribirse de manera no estructurada:

[20]:
```
print(df_constituciones.sample(1)['detalle_aviso'].iloc[0])
```

```
Por escritura del 02/11/2018 se constituyo la sociedad. Socios: Fabian Enrique
ARES, 19/6/67, DNI 18.414.218, Ruta 58, Km 10, Lote 6, Fracción 16, Club de
Campo el Lauquen San Vicente Provincia de Buenos Aires; y Claudio Adrian ATANCE,
5/9/68, DNI 20.507.025, Darwin 327 Piso 14 Departamento A, CABA, ambos
argentinos, casados, empresarios, Plazo: 99 años; Objeto: a) Toda actividad
destinada al diseño de servicios, diseño de software, diseño de imagen,
transformación digital, y experiencia de usuario asi como el diseño, producción,
creación, explotación y mantenimiento de páginas web, y de bases de datos
orientadas al marketing y la publicidad ya sea tradicional o mediante internet y
de toda otra forma electrónica. b) Creación, diseño y desarrollo de sistemas
operativos destinados al desarrollo del diseño gráfico, artes gráficas y a la
imprenta en general; Capital: $ 100.000; Cierre de ejercicio: 31/12; Presidente:
Fabian Enrique Ares, y Director Suplente: Claudio Adrian Atance, ambos con
domicilio especial en la sede; Sede: Catulo Castillo 2630 Piso 1 Oficina A,
CABA. Autorizado según instrumento público Esc. Nº 701 de fecha 02/11/2018 Reg.
Nº 536 Gerardo Daniel Ricoso - T°: 95 F°: 2 C.P.A.C.F.
e. 08/11/2018 N° 84583/18 v. 08/11/2018
```

En base al **Artículo 11 de la Ley 19.550**, todas las constituciones deben contener la siguiente información: 1) El nombre, edad, estado civil, nacionalidad, profesión, domicilio y número de documento de identidad de los socios; 2) La razón social o la denominación, y el domicilio de la sociedad. Si en el contrato constare solamente el domicilio, la dirección de su sede deberá inscribirse mediante petición por separado suscripta por el órgano de administración. Se tendrán por válidas y vinculantes para la sociedad todas las notificaciones efectuadas en la sede inscripta; 3) La designación de su objeto, que debe ser preciso y determinado; 4) El capital social, que deberá ser expresado en moneda argentina, y la mención del aporte de cada socio. En el caso de las sociedades unipersonales, el capital deberá ser integrado totalmente en el acto constitutivo; 5) El plazo de

duración, que debe ser determinado; 6) La organización de la administración, de su fiscalización y de las reuniones de socios; 7) Las reglas para distribuir las utilidades y soportar las pérdidas. En caso de silencio, será en proporción de los aportes. Si se prevé sólo la forma de distribución de utilidades, se aplicará para soportar las pérdidas y viceversa; 8) Las cláusulas necesarias para que puedan establecerse con precisión los derechos y obligaciones de los socios entre sí y respecto de terceros; 9) Las cláusulas atinentes al funcionamiento, disolución y liquidación de la sociedad.

A partir de este punto, el objetivo es obtener la mejor muestra para construir un dataset diverso, que permita hacer un finetuning de Gemma 3 4B.

Este LM va a recibir de input una constitución, y su output será un JSON estructurado con la metadata de la sociedad, capital social, plazos y socios.

En el conjunto tenemos ~22 millones de palabras.

[21]:
```
total_tokens = df_constituciones['detalle_aviso'].str.split().str.len().sum()
print(f'Cantidad de palabras: {total_tokens:,}')
```

```
Cantidad de palabras: 22,512,430
```

Para intentar clasificar los detalles, filtramos con una regular expression, los textos que contengan del 1 al 9, en alguno de los formatos: 1., (1), [1]

[22]:
```
regex_pattern = r'(?<!\d)(?:[1-9][\.\-]|\([1-9]\)|\[[1-9]\])'
filtered_df = df_constituciones[df_constituciones['detalle_aviso'].str.
 ↪contains(regex_pattern, na=False)]

print("# de Sociedades Constituidas: ", len(df_constituciones))
print("# de Sociedades Constituidas, con formato ordenado: ", len(filtered_df))

ratio = len(filtered_df) / len(df_constituciones)
print(f"% De Sociedades constituidas con formato ordenado: {ratio:.2%}")
```

```
# de Sociedades Constituidas:  74930
# de Sociedades Constituidas, con formato ordenado:  39928
% De Sociedades constituidas con formato ordenado: 53.29%
```

Alrededor del 53.29% de los avisos de sociedades mantienen numerado cada punto requerido por el Artículo 11. Esto no es un valor exacto, ya que existen documentos que usan los separadores para delimitar la cantidad de acciones, o accionistas.

[23]:
```
print(filtered_df.sample(1)['detalle_aviso'].iloc[0])
```

```
Constitución: Esc. 688 del 17-10-22 Registro 200 CABA. Socios: Cinthya Elizabeth
Olazarri, argentina, nacida el 27-5-92, soltera, DNI 36.749.976, CUIT
27-36749976-7, empresaria, domiciliada en Av.Corrientes 1584 piso 4 CABA, Rocio
Buffolo, argentina, nacida el 2-8-97, soltera, DNI 40.614.832, CUIT
27-40614832-2, empresaria, con domicilio en Salguero 1063 piso 3 depto.A, CABA;
Darío Antonio Carambula Galeano, paraguayo, nacido el 14-8-73, soltero, DBU
92.329.152, CUIL 20-92329152-1, empresario, con domicilio en Paraná 39 piso 10
depto.A, CABA; y Gabriel Andrés Juricich, argentino, nacido el 21-3-61,
divorciado, DNI 14.508.394, CUIT 20-14508394-0, abogado, con domicilio en
```

Catamarca 159 piso 1 depto.B, CABA. Duración: 30 años. Objeto: realizar por cuenta propia, de terceros, o asociada a ellos las siguientes actividades: A) Producción, realización, organización, comercialización, importación, exportación y contratación en cualquiera de sus formas de espectáculos artísticos, obras teatrales, cinematográficas, de televisión o video, espectáculos culturales, realización de negocios cinematográficos, televisivos, publicitarios en cualquier ámbito y a través de cualquier medio de difusión existente o a crearse en lo sucesivo. B) Promoción, selección y contratación de autores, actores y compositores, directores y productores; artistas e intérpretes, nacionales y extranjeros de los más variados géneros, incluyendo representaciones artísticas. C) Promoción, edición, distribución y difusión del material artístico escrito, grabado o filmado en cualquiera de sus formas. D) Promoción y organización de estudios sobre obras y medios audiovisuales; realización de shows, muestras, ferias, y eventos artísticos; celebración de concursos y certámenes; otorgamiento de premios, distinciones y becas, todo vinculado con el rubro artístico. E) Producción de programas y sistemas publicitarios, publicidad gráfica, radial, televisiva, cinematográfica, periodística, computarizada, y cualquier otro tipo de recurso publicitario, en cualquier ámbito y a través de cualquier medio de difusión existente o a crearse en lo sucesivo. Asimismo podrá realizar anuncios, campañas, auspicios, locaciones de espacios publicitarios, producción, realización, promoción y organización de reuniones, encuentros, congresos y cualquier clase de eventos publicitarios y artísticos. F) Producción, confección, diseño, publicación, compra, venta, edición, comunicación, difusión y distribución de libros, folletos, catálogos, revistas y periódicos culturales, artísticos, o publicitarios y grabaciones en general, nacionales y/o extranjeros, en el país y en el exterior, sobre soportes físicos y/o magnéticos; J) Comercialización, importación y exportación de material gráfico, fílmico e interactivo y de todos aquellos productos relacionados con las actividades anteriores. Capital: $ 100.000 representado por 100.000 acciones ordinarias, nominativas no endosables de $ 1 valor nominal cada una con derecho a un voto por acción. Integración: 25% en dinero efectivo. Plazo para integrar el saldo: 2 años. Suscripción: Cinthya Elizabeth Olazarri suscribe 25.000 acciones e integra $ 6.250; Rocío Buffolo suscribe 25.000 acciones e integra $ 6.250; Darío Antonio Carambula Galeano suscribe 25.000 acciones e integra $ 6.250; y Gabriel Andrés Juricich suscribe 25.000 acciones e integra $ 6.250. Administración: Directorio 1 a 5 titulares por 3 ejercicios. Representación legal: Presidente del Directorio y Vicepresidente, en caso de ausencia o impedimento del Presidente. Fiscalización: Se prescinde. Cierre de ejercicio: 31 de Diciembre. Presidente: Rocío Buffolo. Directora Suplente: Cinthya Elizabeth Olazarri. Sede social y domicilio especial de los directores: Catamarca 159 piso 1 depto.B, CABA. Autorizado según instrumento privado NOTA de fecha 18/10/2022 facundo javier amundarain - Matrícula: 5510 C.E.C.B.A.
e. 01/11/2022 N° 88370/22 v. 01/11/2022

Este método no es confiable, ya que los filtrados pueden fallar, por lo que vamos a aplicar otras técnicas para comprender mejor los avisos de constitución disponibles.

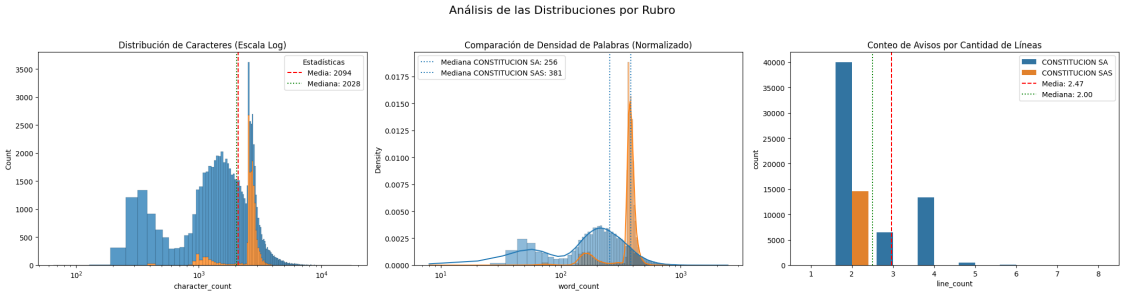
A continuación, se van a calcular tres métricas: - `character_count`: Cantidad total de caracteres.

- `word_count`: Cantidad total de palabras. - `line_count`: Cantidad total de líneas.

[74]:
```
df_constituciones.loc[:, 'character_count'] =␣
 ↪df_constituciones['detalle_aviso'].str.len()
df_constituciones.loc[:, 'word_count'] = df_constituciones['detalle_aviso'].str.
 ↪split().str.len()
df_constituciones.loc[:, 'line_count'] = df_constituciones['detalle_aviso'].str.
 ↪split('\n').str.len()
```

[25]:
```
fig, axes = plt.subplots(1, 3, figsize=(22, 6))
fig.suptitle('Análisis de las Distribuciones por Rubro', fontsize=16)

# --- 1. Caracteres (Escala Logarítmica y Mediana) ---
sns.histplot(data=df_constituciones, x='character_count', hue='rubro',␣
 ↪multiple='stack', ax=axes[0])
axes[0].set_title('Distribución de Caracteres (Escala Log)')
axes[0].set_xscale('log') # Aplicar escala logarítmica

# Añadir líneas de media y mediana
mean_char = df_constituciones['character_count'].mean()
median_char = df_constituciones['character_count'].median()
axes[0].axvline(mean_char, color='r', linestyle='--', label=f'Media: {mean_char:
 ↪.0f}')
axes[0].axvline(median_char, color='g', linestyle=':', label=f'Mediana:␣
 ↪{median_char:.0f}')
axes[0].legend(title='Estadísticas')

# --- 2. Palabras (Comparación de Densidad Normalizada) ---
sns.histplot(
    data=df_constituciones,
    x='word_count',
    hue='rubro',
    multiple='layer',
    stat='density',
    common_norm=False,
    kde=True,
    ax=axes[1]
)
axes[1].set_title('Comparación de Densidad de Palabras (Normalizado)')
axes[1].set_xscale('log')

# Calcular y mostrar la mediana para cada rubro
for rubro_name in df_constituciones['rubro'].unique():
    median_val = df_constituciones[df_constituciones['rubro'] ==␣
 ↪rubro_name]['word_count'].median()
    axes[1].axvline(median_val, linestyle=':', label=f'Mediana {rubro_name}:␣
 ↪{median_val:.0f}')
```

```
axes[1].legend()


# --- 3. Líneas (Usando un Bar Plot para datos discretos) ---
# Para datos con pocos valores únicos como 'line_count', un bar plot es a␣
 ↪menudo más claro.
sns.countplot(data=df_constituciones, x='line_count', hue='rubro', ax=axes[2])
axes[2].set_title('Conteo de Avisos por Cantidad de Líneas')
mean_line = df_constituciones['line_count'].mean()
median_line = df_constituciones['line_count'].median()
axes[2].axvline(mean_line - 0.5, color='r', linestyle='--', label=f'Media:␣
 ↪{mean_line:.2f}') # El -0.5 es para centrar la línea entre las barras
axes[2].axvline(median_line - 0.5, color='g', linestyle=':', label=f'Mediana:␣
 ↪{median_line:.2f}')
axes[2].legend()


plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```



Utilizando estos gráficos, se pueden notar algunas conclusiones: - Los avisos de CONSTITUCION SAS (naranja) son muy homogéneos. El gráfico de densidad muestra que tienen una longitud muy predecible y consistente, con una mediana de 381 palabras. - Por el contrario, los avisos de CONSTITUCION SA (azul) son mucho más variables. Su distribución de palabras es más ancha y plana, indicando una gran diversidad en la longitud de los avisos, a pesar de que su mediana es menor (256 palabras). - El primer gráfico (distribución de caracteres) muestra dos "grupos" o modelos de longitud. Las SAS se concentran casi exclusivamente en el modelo de avisos más cortos (alrededor de 2000 caracteres), mientras que las SA están presentes en ambos, dominando el grupo de avisos más largos. Esto refuerza la idea de mayor variabilidad en las SA.

En resumen, los avisos de constitución de SAS siguen un formato muy estandarizado y predecible en longitud y estructura, mientras que los de SA presentan una diversidad mucho mayor.

A continuación, vamos a utilizar BERT para vectorizar los anuncios, y luego poder utilizar K-Means para hacer un clustering automático de los anuncios en base a su estructura y contenido.

[26]:
```
import torch
from sentence_transformers import SentenceTransformer
from sklearn.cluster import KMeans
from sklearn.manifold import TSNE
```

Primero, definimos el device a utilizar para ejecutar el modelo. En esta caso vamos a trabajar en GPU con CUDA.

[27]:
```
device = 'cuda' if torch.cuda.is_available() else 'cpu'
print(f"Using device: {device}")
```

```
Using device: cuda
```

Cargamos un modelo de SentenceTransformer, en este caso paraphrase-multilingual-MiniLM-L12-v2, un modelo de 384 dimensiones, que mapea oraciones y párrafos a un espacio vectorial, para poder clusterizarlo.

[28]:
```
print("Loading SentenceTransformer model...")
model = SentenceTransformer('paraphrase-multilingual-MiniLM-L12-v2',␣
 ↪device=device)
print("Model loaded.")
```

```
Loading SentenceTransformer model…
Model loaded.
```

En este paso, vamos a generar los embeddings para cada aviso. Este es el proceso de mayor costo, en CPU puede tardar mucho, por eso es preferible ejecutar en GPU.

[75]:
```
print(f"Generating embeddings for {len(df_constituciones)} documents...")
df_constituciones.loc[:, 'detalle_aviso'] = df_constituciones['detalle_aviso'].
 ↪astype(str)
embeddings = model.encode(df_constituciones['detalle_aviso'].tolist(),␣
 ↪show_progress_bar=True)
print("Embeddings generated with shape:", embeddings.shape)
```

```
Generating embeddings for 74930 documents…
```

```
Batches:   0%|          | 0/2342 [00:00<?, ?it/s]
```

```
Embeddings generated with shape: (74930, 384)
```

Luego, vamos a agrupar los embeddings calculados en ocho clusters, utilizando KMeans.

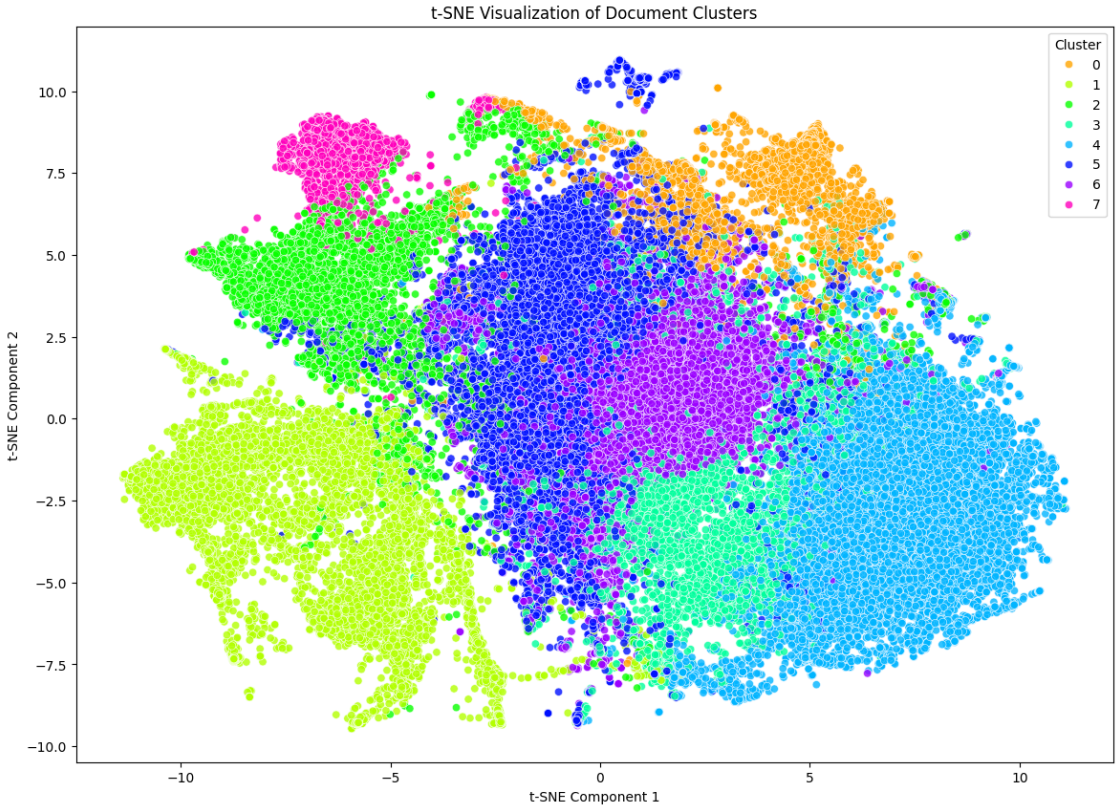
[76]:
```
num_clusters = 8
print(f"Running K-Means clustering with k={num_clusters}...")
kmeans = KMeans(n_clusters=num_clusters, random_state=42, n_init=10)
kmeans.fit(embeddings)

# Assign the cluster label to each document in the DataFrame
df_constituciones.loc[:, 'cluster'] = kmeans.labels_
print("Clustering complete.")
```

```
Running K-Means clustering with k=8…
Clustering complete.
```

Los embeddings, al tener 384 dimensiones, necesitamos reducirlos a un formato bi-dimensional para poder graficarlo. Por esto utilizamos t-SNE, que asigna a cada data point, una posición en un espacio bi-dimensional.

[77]:
```python
print("Reducing dimensions with t-SNE for visualization...")

tsne = TSNE(n_components=2, perplexity=30, random_state=42, max_iter=300)
embeddings_2d = tsne.fit_transform(embeddings)
df_constituciones.loc[:, 'tsne_x'] = embeddings_2d[:, 0]
df_constituciones.loc[:, 'tsne_y'] = embeddings_2d[:, 1]
print("t-SNE complete.")

# Create the scatter plot
plt.figure(figsize=(14, 10))
sns.scatterplot(
    x='tsne_x', y='tsne_y',
    hue='cluster',
    palette=sns.color_palette("hsv", n_colors=num_clusters),
    data=df_constituciones,
    legend="full",
    alpha=0.8
)
plt.title('t-SNE Visualization of Document Clusters')
plt.xlabel('t-SNE Component 1')
plt.ylabel('t-SNE Component 2')
plt.legend(title='Cluster')
plt.show()

# Display the size of each cluster
print("\nDistribution of documents per cluster:")
print(df_constituciones['cluster'].value_counts().sort_index())
```

```
Reducing dimensions with t-SNE for visualization…
t-SNE complete.
```




```
Distribution of documents per cluster:
cluster
0     5036
1    14451
2     6901
3     6736
4    14767
5    14804
6     9863
7     2372
Name: count, dtype: int64
```

Cada número de cluster representa un formato o plantilla particular de constitución. Algunas conclusiones del análisis son las siguientes: - Los clusters más grandes (como el 1, 4, 5 y 6) son, muy probablemente, los formatos más comunes y estandarizados en el conjunto de datos. - Los clusters más pequeños (como el 7) son los más interesantes: representan formatos raros, casos especiales o documentos atípicos.

A continuación, podemos analizar algunas de las muestras de avisos societarios por cada cluster.

[79]:
```python
def inspect_cluster(df, cluster_id, n_samples=3):
    """Prints a few random text samples from a specified cluster."""
    print(f"--- Inespeccionando Cluster {cluster_id} ---")
    samples = df[df['cluster'] == cluster_id].sample(n=min(n_samples,␣
 ↪len(df[df['cluster'] == cluster_id])))
    for i, row in samples.iterrows():
        # Print a snippet of the text to avoid flooding the screen
        print(f"  Índice del Documento: {i}")
        print(f"  Snippet del Aviso: {row['detalle_aviso'][:300]}...\n")

# Get the unique cluster IDs from your DataFrame
cluster_ids = sorted(df_constituciones['cluster'].unique())

# Loop through each cluster and print some examples
print("Muestras aleatorias para cada cluster generado:")
for cid in cluster_ids:
    inspect_cluster(df_constituciones, cid)
```

```
Muestras aleatorias para cada cluster generado:
--- Inespeccionando Cluster 0 ---
  Índice del Documento: 70661
  Snippet del Aviso: Escrituras Nº 143 del 05/08/11 y Nº 213 del 25/10/11.
Accionistas: Claudio Alberto Boccardo, 44 años, casado, DNI: 17.720.355,
Balcarce 362, General Deheza, Provincia de Córdoba; Rosana. Maria Somale, 43
años, casada, DNI: 18.515.529, Belgrano 234 General Deheza, Provincia de
Córdoba; Leandro Esteba…

  Índice del Documento: 1162431
  Snippet del Aviso: Constitucion: 2/6/2023. Socios: Walter Pedro CARRERA,
13/4/1964, DNI: 16822523, CUIT: 20-16822523-8, Montiel 775, Planta baja, CABA,
suscribe 25000 acciones; Lautaro Ezequiel CARRERA, 11/4/1997, DNI: 40239619,
CUIT: 23-40239619-9, Yatay 54, Piso 5, Depto A, CABA, suscribe 25000 acciones;
Miguel Ange…

  Índice del Documento: 938991
  Snippet del Aviso: Por Escritura Pública 199 del 11/11/2020 se constituyó la
sociedad; accionistas: Jose Manuel RODRIGUEZ HENRIQUEZ, DNI 16827721, 20/2/1964,
Abogado, domicilio real en Calle 58 N° 1269, La Plata, Provincia de Buenos
Aires, 85.000 acciones; Alejandro Teodoro SABBINO, DNI 14318513, 20/6/1961,
comerciant…

--- Inespeccionando Cluster 1 ---
  Índice del Documento: 801405
  Snippet del Aviso: CONSTITUCIÓN: 19/12/2018. 1.- GERMAN MANUEL WERNICKE
ZAPIOLA, 16/08/1947, Casado/a, Argentina, SERVICIOS INMOBILIARIOS REALIZADOS A
CAMBIO DE UNA RETRIBUCIÓN O POR CONTRATA N.C.P., MONTEVIDEO 1306 piso 7° B
CIUDAD_DE_BUENOS_AIRES, LE Nº 7598861, CUIL/CUIT/CDI Nº 20075988614, . 2.-
"Mercantil Rural S…
```

Índice del Documento: 719266
Snippet del Aviso: CONSTITUCIÓN: 27/02/2018. 1.- JUAN ELPIDIO BOSCH, 07/01/1950, Casado/a, Argentina, SERVICIOS RELACIONADOS CON LA SALUD HUMANA N.C.P., JUANA MANSO 1830 piso 8 808 CIUDAD_DE_BUENOS_AIRES, LE Nº 7706668, CUIL/CUIT/CDI Nº 20077066684, MARIA PILAR BOSCH, 13/05/1992, Soltero/a, Argentina, TRABAJADOR RELAC…

Índice del Documento: 870310
Snippet del Aviso: CONSTITUCIÓN: 07/10/2019. 1.- CARLOS SEBASTIAN GONZALEZ, 22/09/1980, Soltero/a, Argentina, REPARACIÓN DE CÁMARAS Y CUBIERTAS, RONDEAU 2768 piso GENERAL_SAN_MARTíN, DNI Nº 28440024, CUIL/CUIT/CDI Nº 20284400241, JAVIER ALEJANDRO ZANUSSI, 01/06/1970, Soltero/a, Argentina, SERVICIO DE TRANSPORTE AUTOMO…

--- Inespeccionando Cluster 2 ---
Índice del Documento: 523086
Snippet del Aviso: El siguiente aviso complementa el publicado el 07/09/2015 T N° 142769/15. Director suplente: Jorge Roberto Presutto Autorizado según instrumento privado autorizacion de fecha 02/09/2015 AGUSTINA MEOLI - T°: 119 F°: 704 C.P.A.C.F.
e. 02/12/2015 N° 172060/15 v. 02/12/2015…

Índice del Documento: 499110
Snippet del Aviso: Rectifica edicto del 30/6/15 T.I. N° 114382/15. Donde dice "6) $ 170000", debe decir "6) $ 100000", siendo que el Capital Social es de $ 100000 Autorizado según instrumento público Esc. Nº 383 de fecha 29/05/2015 Reg. Nº 172
Silvia Epelbaum - Habilitado D.N.R.O. N° 3369
e. 09/09/2015 N° 143438/15 v…

Índice del Documento: 973551
Snippet del Aviso: Rectificatorio publicación del 08/04/2021, No. 21302/21, la sede social es Tucuman 1455, piso 11, Of "F" Caba y no como por error se consignó.- Autorizado según instrumento público Esc. Nº 93 de fecha 05/04/2021 Reg. Nº 1393 Marcelo Fabian Schillaci - Matrícula: 4868 C.E.C.B.A.
e. 27/04/2021 N° 2664…

--- Inespeccionando Cluster 3 ---
Índice del Documento: 1167290
Snippet del Aviso: Escritura 76 del 8/6/2023 registro 474 CABA. Socios Andrés Rodolfo ZAZZINI nació 8/10/1968 DNI 20404270, María Laura GALETTI nació 9/6/1974 DNI 23767874, ambos domicilio Lavalle 1290 piso 7 "710" CABA; Ariel Osvaldo ROPERTI nació 22/4/1968 DNI 20027616, Jéssica Alexandra PÉREZ nació 22/9/1994 DNI 38…

Índice del Documento: 784887
Snippet del Aviso: Constitución: Esc. 775 del 26/10/18 F° 2018 Registro 2001

C.A.B.A. Socios: Claudia Patricia MALAGRINO, nac. 15/9/61, DNI 14976103, CUIT 27-14976103-4, domiciliada en Camacua 719, piso 3, CABA; y Horacio Daniel TOMASULO, nac. 9/4/61, DNI 14.728.169, CUIT 20-14728169-3, domiciliado Carabobo 790, piso …

  Índice del Documento: 87933
  Snippet del Aviso: Esc. Pub. 125 pasada al folio 273 del 21/11/2011 Registro Notarial 808 de María Isabel Ugarteche. 1) Graciela Sandra Meira soltera, argentina, empresaria, 21/05/73, DNI N° 23.170.202, México 1482 2° piso Dpto 1 C.A.B.A y Andrea Yanina Berta, soltera, argentina, 18/04/88, empresaria, DNI N° 33.554.56…

--- Inespeccionando Cluster 4 ---
  Índice del Documento: 691022
  Snippet del Aviso: Instrumento Público del 18/10/2017. Socios: Daniel Horacio TRIDICO, argentino, casado, nacido el 5 de julio de 1964, comerciante, titular del Documento Nacional de Identidad número 17.199.101, C.U.I.T. 20-17199101-4, con domicilio real en la calle Juan Agustín Garcia 3939, y Ana María DA SILVA RAMAL…

  Índice del Documento: 161416
  Snippet del Aviso: Constitución: Esc.81 del 2/8/12 Fº 221 Registro 1855 C.A.B.A. Socios: Kwon Jin LEE, DNI 18.883.281, nacido el 1/12/71, empresario, con domicilio en Avenida Rivadavia 1965 piso 1, Comodoro Rivadavia, Provincia de Chubut; y Kwon Mi LEE, DNI 18.766.954, nacida el 12/7/67, abogada, con domicilio en Pres…

  Índice del Documento: 932876
  Snippet del Aviso: Esc. 113, del 22/10/2020, Esc. Maria Paula Vega Reg. 1525, Socios: Flavio Alejandro IGLESIAS, argentino, nacido el 28/01/1977, D.N.I. 25.787.311, CUIT 20-25787311-1, abogado, domicilio Arcos Nº 2761, 9º, depto."D",C.A.B.A. y Fernando Javier IGLESIAS, argentino, nacido el 20/08/1974, D.N.I. 24.111.27…

--- Inespeccionando Cluster 5 ---
  Índice del Documento: 567885
  Snippet del Aviso: Por Escritura Nº 48 del 15/06/2016, se constituye la sociedad: 1) DESDEPRIX S.A. 2) SERGIO GABRIEL RAO, argentino, DNI 26.240.323, casado, diseñador gráfico, domiciliado en Conesa 2885, Piso 2º departamento E, CABA, PRESIDENTE; y CRISTINA VERONICA SALGADO, argentina, DNI 26.967.220, casada, profesor…

  Índice del Documento: 787405
  Snippet del Aviso: "CENTRO ARGENTINO DE UROLOGÍA S.A." Por instrumento Público del 22/10/2018. 1) Socios: Vanesa Soledad GAUDE, argentina, casada, CUIT 27-29167868-3, 15/07/1981, instrumentadora quirúrgica, DNI 29.167.868, y Norberto Osvaldo Bernardo, argentino, casado, CUIT 20-16606774-0, 20/05/1964, médico, DNI 16.6…

Índice del Documento: 183226
Snippet del Aviso: Escritura 24/10/12. 1) Susana Ines Mariani, DNI 4975996, 22/8/45, soltera, Av. Pueyrredón 154 piso 6° departamento D, Córdoba, Pcia. Córdoba; y Maria Magdalena de Olmos, DNI 23825165, 4/5/74, casada, Cipriano Perello 4401, Córdoba, Pcia. Córdoba; ambas argentinas, comerciantes. 2) PRECOB S.A. 3) Av…

--- Inespeccionando Cluster 6 ---
Índice del Documento: 1228699
Snippet del Aviso: ESCRITURA 35 de fecha 15/03/24. 1) Víctor Fabian FLORES, 18/4/1976, DNI 25235581, CUIT 20-25235581-3, casado en primeras nupcias con Leonidas Pereira Ocampo, domicilio en Av San Martín 4001, edificio 5, puerta 1 s/n, piso 1, departamento A, Rafael Calzada, Almirante Brown, Provincia de Buenos Aires…

Índice del Documento: 426449
Snippet del Aviso: 1) Alejandro Omar MOLERO, 3/9/1967, casado, DNI 18.393.883, domiciliado en Salta 3038, Olivos, Provincia de Bs. As., y Gustavo Marcelo CASTELNUOVO, 25/9/1973, soltero, DNI 23.450.536, domiciliado en la calle Marcelo T. de Alvear 1308, Piso 21, Departamento "B", C.A.B.A., ambos argentinos y empresari…

Índice del Documento: 933641
Snippet del Aviso: Por escritura del 20/10/2020 1) Marcos Ariel CENTURIÓN, divorciado, comerciante, 18/11/1984, DNI 31.292.036, Pasaje Renan 1173, CABA, Osvaldo Rubén BELIZAN, soltero, comerciante, 02/10/1973, DNI 23.532.310, Nicasio Oroño 676, Planta Baja, CABA; Sergio Rubén BORTZ, divorciado, comerciante, 24/06/1965…

--- Inespeccionando Cluster 7 ---
Índice del Documento: 1185489
Snippet del Aviso: Se rectifica edicto: 18/8/23 N° 65029/23- Por acta complementaria de constitución del 4/9/23 se reformo el objeto social (art 3): La sociedad tiene por objeto exclusivo realizar las actividades previstas para las categorías de Agente Productor, Agente de Negociación, Agente Asesor Global de Inversio…

Índice del Documento: 166840
Snippet del Aviso: Se hace saber por un día la constitución de la sociedad GVA SOCIEDAD ANONIMA. Instrumento público de constitución: Escrituras Números 9 y 10 de fecha 9.01.2012 y 467 de fecha 24.07.2012. Razón Social: GVA SOCIEDAD ANONIMA. Sede social: Maipú Nº 741, Piso 1° Departamento "B", Ciudad Autónoma de Bueno…

Índice del Documento: 1246444
Snippet del Aviso: RECTIFICA publicación del 18/04/2024 Nº 21913/24 Por Escritura Complementaria nº 215 Folio 443 del 29/05/2024 se modifica el objeto

```
social el que queda redactado del siguiente modo: "ARTICULO TERCERO: La sociedad
tendrá por objeto realizar por cuenta propia o de terceros y/o asociada a
terceros, ya …
```

Viendo los clusters identificados, el Cluster 7 no son constituciones realmente, sino que son modificaciones mal clasificadas! Deberían ser del rubro **1120 - REFORMAS**. Por esto, vamos a excluirlas para construir el dataset de finetuning.

[93]:
```python
df_filtered = df_constituciones[df_constituciones['cluster'] != 7].copy()
print(f"\nOriginal document count: {len(df_constituciones)}")
print(f"Document count after excluding cluster 7: {len(df_filtered)}")
```

```
Original document count: 74930
Document count after excluding cluster 7: 72558
```

Vamos a tomar una muestra de 500 avisos, con un sampling estratificado, tomando muestras en relación a la cantidad total de avisos por cluster.

[ ]:
```python
total_sample_size = 500

# Perform the stratified sampling
df_sample = df_filtered.groupby('cluster').apply(
    lambda x: x.sample(int(np.rint(total_sample_size * len(x) /␣
 ↪len(df_filtered)))),
    include_groups=False
).sample(frac=1, random_state=42).reset_index() # Use reset_index() to turn␣
 ↪cluster index into a column

print(f"\n--- Created a stratified sample of {len(df_sample)} documents ---\n")
print("Distribution of samples per cluster:")

print(df_sample['cluster'].value_counts().sort_index())
```

```
--- Created a stratified sample of 501 documents ---

Distribution of samples per cluster:
cluster
0     35
1    100
2     48
3     46
4    102
5    102
6     68
Name: count, dtype: int64
```

Finalmente, exportamos las muestras para empezar a hacer el labeling del dataset. únicamente

tomamos el ID de aviso y el Detalle.

[95]:
```
df_to_label = df_sample[['aviso_id','detalle_aviso']]

output_filename = 'finetuning_samples_for_labeling.csv'
df_to_label.to_csv(output_filename, index=False)

df_to_label.head()
```

[95]:

	aviso_id	detalle_aviso
0	A849431	1) 12/6/19 2) Horacio Ezequiel SEGOVIA, DNI 38…
1	A858048	CONSTITUCIÓN: 15/07/2019. 1.- EXEQUIEL DAVID C…
2	A844068	Escritura de constitución "VATUIT S.A." del 30…
3	A952759	Se rectifica aviso T.I. Nº 47398/20 del 19-10-…
4	A970210	CONSTITUCIÓN: 03/11/2020. 1.- PABLO JORGE CERN…