

```
In [1]: import pandas as pd
import numpy as np
import re

import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
stopwords = stopwords.words('spanish')

from wordcloud import WordCloud, ImageColorGenerator
import matplotlib.pyplot as plt
import PIL.Image
from PIL import Image
from IPython.display import display
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

```
In [2]: # Ruta de archivo
post = pd.read_excel("solicitudes.xlsx")

# Muestra las primeras filas del DataFrame
post.head()
```

Out[2]:

	tema
0	Traslado Derecho de Petición Situación Departam...
1	Solicitud subsidios zona de frontera
2	Emergencia social y económica
3	Seguimiento a temas de energía
4	H.R Modesto Aguilera-

```
In [3]: post.tema
palabra = post.tema.str.cat(sep=" ")
stopwords.extend(["solicitud","petición","información","cuáles","cuánto","informe","indique","fecha",
"ministerio",

"entidad","traslado","formato","excel","existen","existe","siguiente","vigente","pregunta","preguntas",
",
"vigencia","cargo","indicar","informar","sírvasse","cómo","realizado","remitir","colombia","ministro",
"respuesta","respecto","solicito","incluyendo","frente","número","remitido","indicando","presidente",
"indicando","copia","dentro","solicita","detallar","presente","cuales","actualmente","congreso",
"preguntas","colombiano","actual","documento","cuántos","cuestionario","informacion","comisión",
"último","correspondiente","últimos","manera","encuentran","derecho","petición","radicado","favor",
"peticion","urgente"
])
```

In [5]:

```
def pre_procesado(texto):
    texto = texto.lower()
    texto = re.sub(r"[\W\']+"," ", texto)
    texto = [palabra for palabra in texto.split() if palabra not in stopwords and len(palabra) >= 5]
    texto = " ".join(texto)
    return texto

post['pp'] = post['tema'].apply(lambda texto: pre_procesado(texto))
post
```

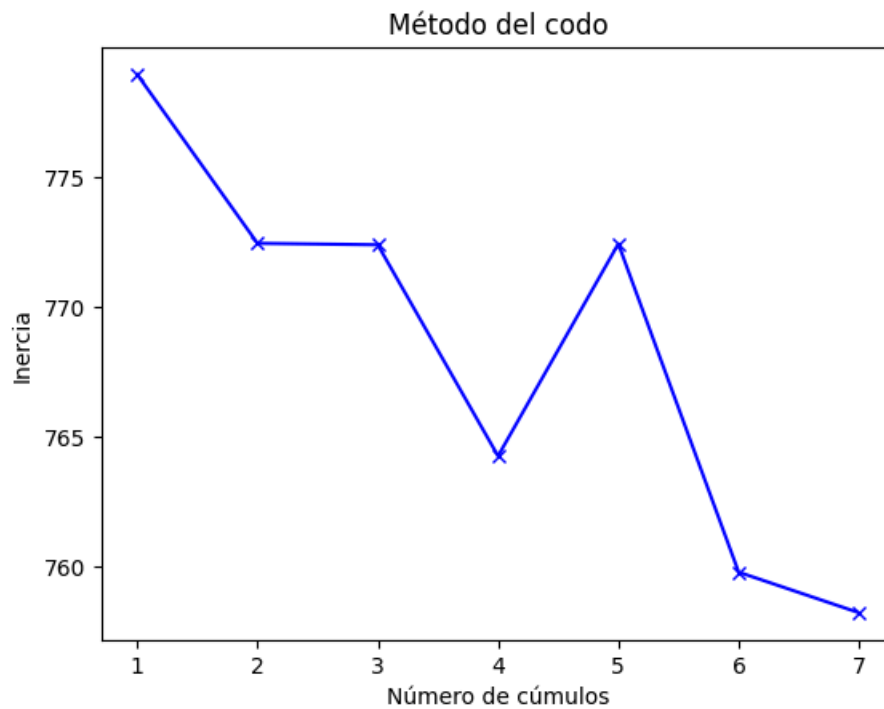
Out[5]:

	tema	pp
0	Traslado Derecho de Petición Situación Departam...	situación departamento chocó
1	Solicitud subsidios zona de frontera	subsidios frontera
2	Emergencia social y económica	emergencia social económica
3	Seguimiento a temas de energía	seguimiento temas energía
4	H.R Modesto Aguilera-	modesto aguilera
...	...	...
791	El H.R Alexander Guarin solicita se brinde res...	alexander guarin brinde solución urgente probl...
792	Traslado de la Agencia de Renovacion del Terri...	agencia renovacion territorio diego caicedo fi...
793	La Presidencia de la Camara de Representantes ...	presidencia camara representantes remite petic...
794	El H.R Hugo Archila traslada derecho de petici...	archila traslada peticion ciudadana elevado as...
795	El H.S Ciro Ramírez hace solicitud de 5 pregun...	ramírez recursos reservas publicado retraso pl...

796 rows × 2 columns


```
In [10]: # Método del codo
plt.xlabel("Número de cúmulos")
plt.ylabel("Inercia")
plt.title("Método del codo")
plt.plot(ks, k_inertias, 'bx-')
```

Out[10]: [<matplotlib.lines.Line2D at 0x7c8367e18230>]



```
In [11]: #Entrenamiento del modelo
k = 6
model = KMeans(n_clusters=k)
model.fit(tfidf_matrix)
```

Out[11]:

KMeans ⓘ ?

KMeans(n_clusters=6)

```
In [12]: post['cluster'] = model.labels_
```

```

In [13]: #los vecinos más cercanos al centroide
#entrenamiento con matriz tfidf
nbrs = NearestNeighbors(n_neighbors=3, metric='euclidean').fit(tfidf_matrix.values)

#porcentaje de los datos de cada clase
clust_cnt = post['cluster'].value_counts()
clust_cnt_pct = post['cluster'].value_counts(normalize=True)

centroids = model.cluster_centers_

terms = tfidf_vect.get_feature_names_out()
order_centroids = centroids.argsort()[::-1]

for i in range(k):
    print(f"CLUSTER #{i}")
    print(f"Tiene {clust_cnt[i]} registros ({clust_cnt_pct[i]:.2%} de los datos)")

    print("Palabras más importantes: ")
    for ind in order_centroids[i][:5]:
        print(terms[ind], end=" ")
    print()

    print(f"Comentarios más Representativos: ")
    for vecino in post.iloc[nbrs.kneighbors([centroids[i]])[1][0]].pp.values:
        print(f" {vecino}")

    print("*****")
    print()

```

```

CLUSTER #0
Tiene 130 registros (16.33% de los datos)
Palabras más importantes:
ejecución contratos valor recursos gobierno
Comentarios más Representativos:
*
*
*
*****

```

```

CLUSTER #1
Tiene 169 registros (21.23% de los datos)
Palabras más importantes:
peticion insistencia antioquia urgente ciudadana
Comentarios más Representativos:
*
*
*
*****

```

```

CLUSTER #2
Tiene 181 registros (22.74% de los datos)
Palabras más importantes:
proyectos departamento energética nacional ambiental
Comentarios más Representativos:
*
*
*
*****

```

```

CLUSTER #3
Tiene 42 registros (5.28% de los datos)
Palabras más importantes:
cauca seguimiento territorial norte proyectos
Comentarios más Representativos:
*
*

```

```

*
*****

CLUSTER #4
Tiene 101 registros (12.69% de los datos)
Palabras más importantes:
energía combustible natural eléctrica servicio
Comentarios más Representativos:
*
*
*
*****

CLUSTER #5
Tiene 173 registros (21.73% de los datos)
Palabras más importantes:
proyecto presupuesto programa proyectos recursos
Comentarios más Representativos:
*
*
*
*****

```

```

In [14]: clusters = {0: 'Ejecución presupuestal',
                    1: 'Peticiones ciudadanas',
                    2: 'Proyectos con afectación ambiental',
                    3: 'Entidades territoriales',
                    4: 'Matriz energética',
                    5: 'Seguimiento a planeación' # Agregamos la etiqueta para el cluster 5
                    }

post['nombres_clusters'] = post['cluster'].apply(lambda val: clusters[val])
post.sample(5)

```

Out[14]:

	tema	pp	cluster	nombres_clusters
379	Traslado del CORPOCESAR (Corporacion Autonoma ...	corpocesar corporacion autonoma regional cesar...	2	Proyectos con afectación ambiental
465	Solicitud de 4 preguntas de la H.S Sandra Lobo...	sandra detallada títulos mineros otorgados dep...	2	Proyectos con afectación ambiental
368	Devolucion de MIN DEFENSA a solicitud de infor...	devolucion defensa trasladadas acuerdo conside...	5	Seguimiento a planeación
61	H.S Ana Paola Agudelo. H.S Carlos Eduardo Guev...	paola agudelo carlos eduardo guevara manuel vi...	5	Seguimiento a planeación
530	1) ¿Cuál ha sido la inversión total del Minist...	inversión total entidades adscritas vinculadas...	0	Ejecución presupuestal

```

In [15]: from sklearn.feature_extraction.text import TfidfVectorizer
          from sklearn.cluster import DBSCAN
          from sklearn.preprocessing import StandardScaler

```

```
In [16]: from sklearn.model_selection import train_test_split

X = post['pp']
y = post['cluster'] # Se cambió 'label' a 'cluster'

# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
In [17]: import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.countplot(data=post, x='nombres_clusters', hue='nombres_clusters', palette='viridis',
legend=False)
plt.xlabel('Temas')
plt.ylabel('Solicitudes')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

