Visualización de datos en alta dimensión

Guillermo Ruiz

La visualización de datos en alta dimensión es complicada y requiere un análisis especial para cada caso en particular. Un procedimiento que se puede usar es el siguiente:

1. Visualizar los datos en 2 dimensiones usando una técnica como UMAP ajustando los parámetros adecuadamente.
2. Aplicar una técnica de detección de anomalías para descartar valores atípicos.
3. Aplicar un algoritmo de agrupación para separar los datos y analizarlos por separado.
4. Revisar el contenido de los cluster para interpretar su contenido.

Ejemplo

Veremos un ejemplo con un conjunto de datos de vectores obtenidos de textos mediante un modelo de lenguaje. Se lee el conjunto de vectores y se muestran algunos de los textos.

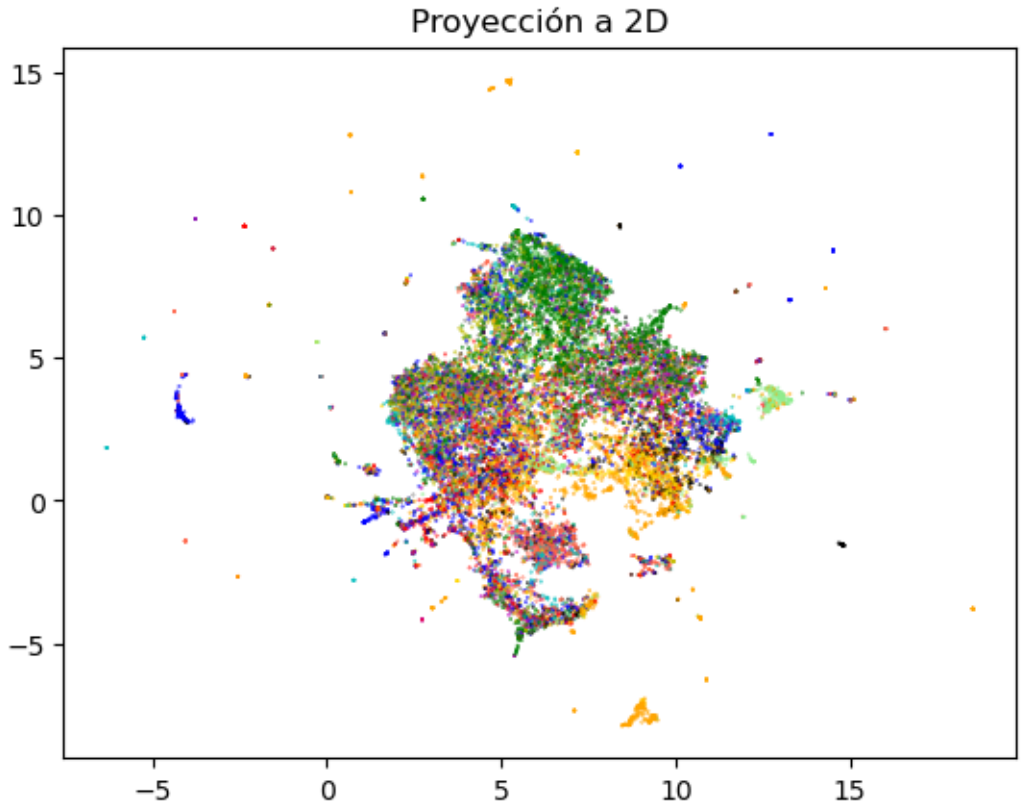
```
X = np.load("encajes.npy")
X.shape
```

```
(20000, 768)
```

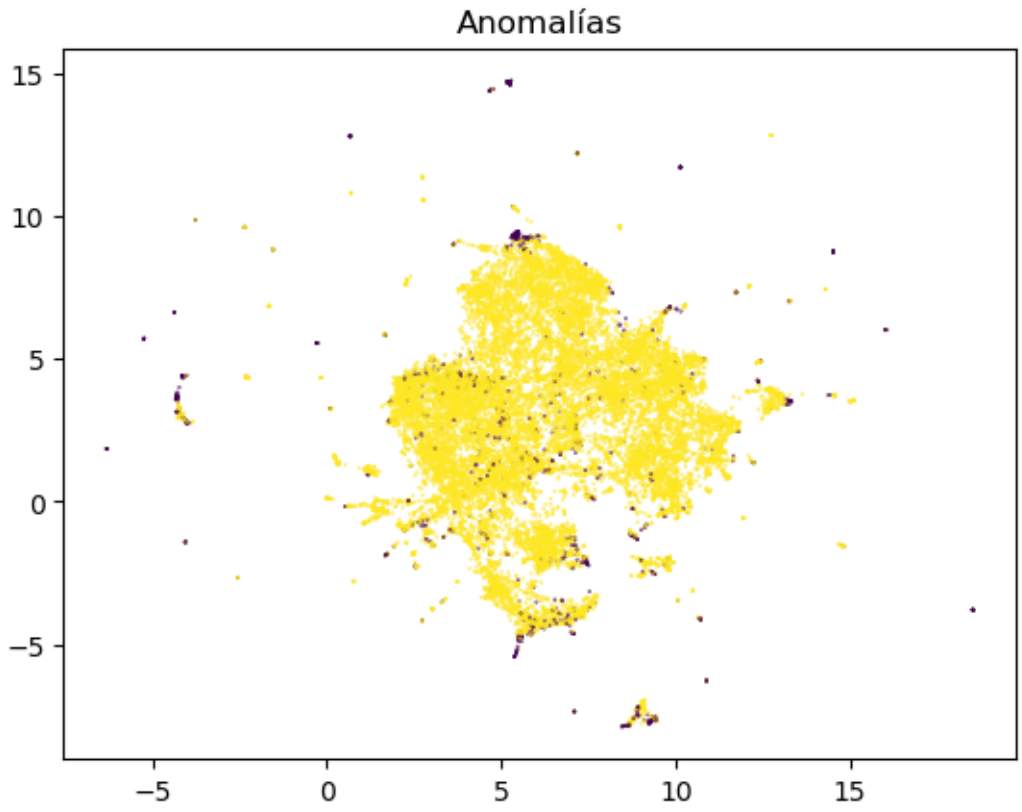
```
df = pd.read_csv("small_df.csv")
df.head()
```

	text	label
0	Mi turbo carnala del alma _USR _emo en La Pl...	
1	Las chicas de *Jimmy's* te esperan para pasar ...	
2	Big Time Rush y Mario Casas me quieren matar a...	
3	Cosas de feria... Ayer _USR nos iba invitar lo...	
4	_USR Ya terminaste ....? _emo_emo_emo _URL	

Se grafican los datos en dos dimenisones. Los colores corresponden con las etiquetas.



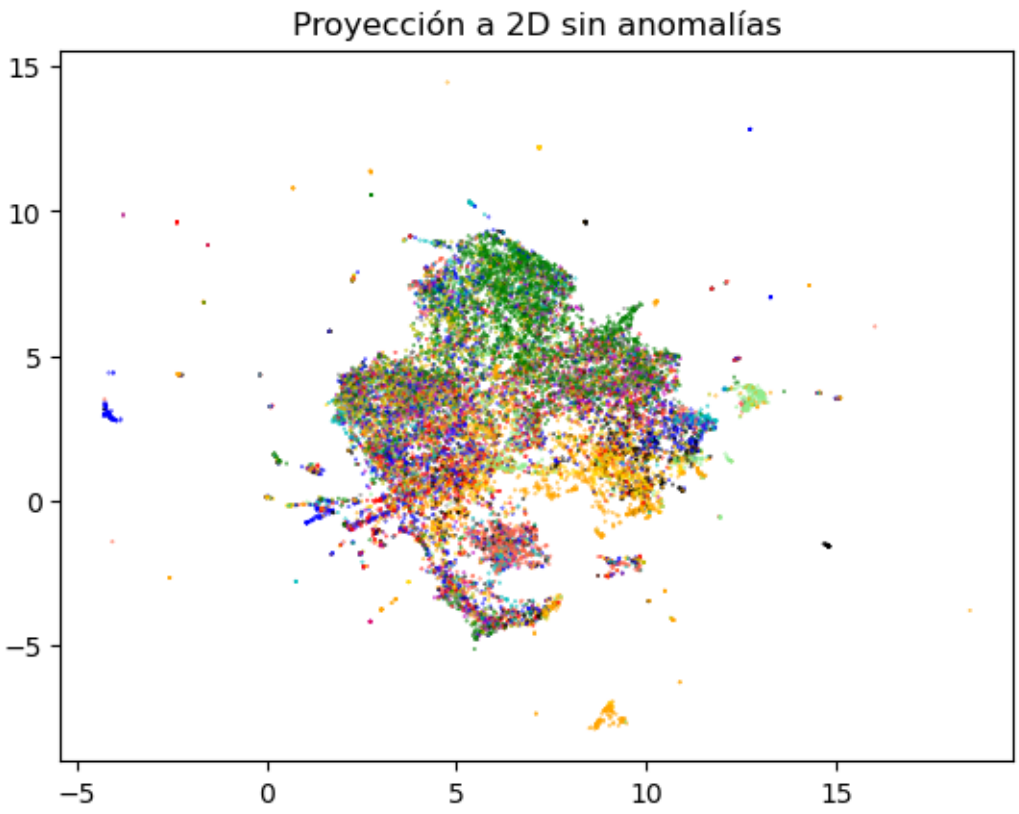
Se identifican las anomalías en los datos.



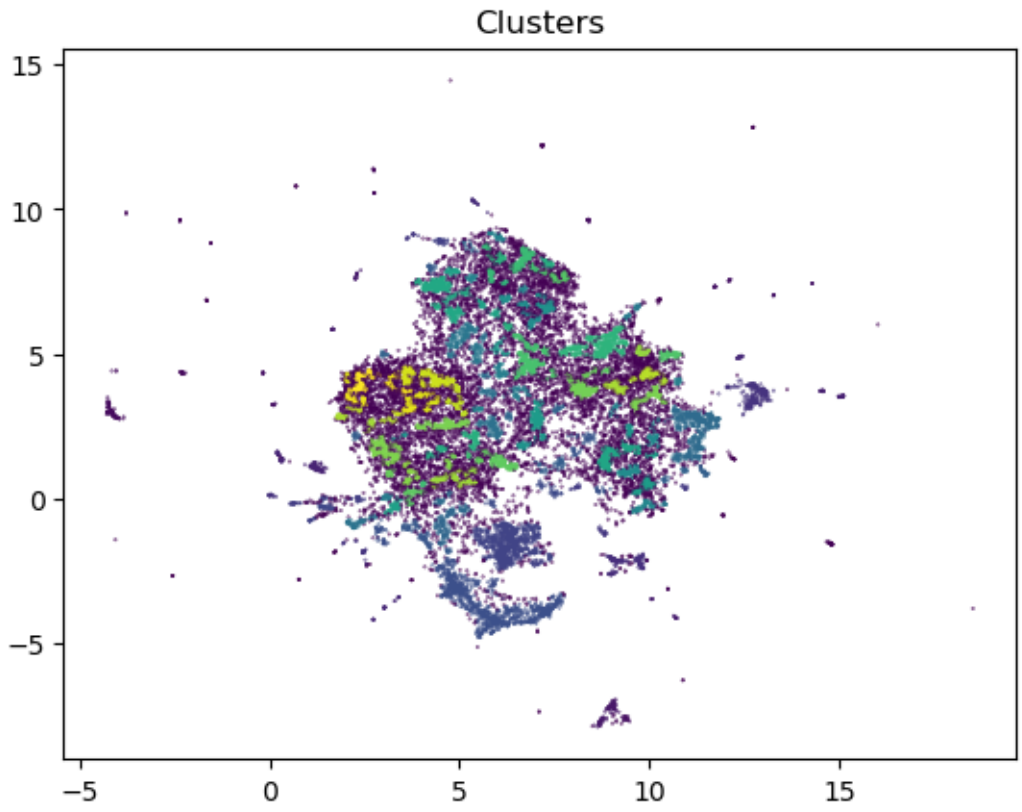
Se muestran los 10 mensajes más alejados (usando `decision_function`).

	text	lab
2155	Pfff que rico se ve todo eso _emo_emo_emo_emo_emo_emo_emo_emo_emo_...	
3291	_USR Entonces novia mía te mando 100 rosas para la más bella flor ...	
3477	_USR _USR ...	
7013	˜ ˜ ! # # # # ...	
15995	˜ ˜ ! # # # # ...	
18930	Busco #cabrón #activo #vergón para #coger MandenMD #HeterosCurioso...	
5448	_emo_emo_emo_emo_emo_emo_emo_emo_emo_emo_emo_emo_emo_emo_emo_emo_e...	
14205	Busco #cabrón #activo #vergón para #coger MandenMD #HeterosCurioso...	
11879	_USR ! _emo	
17345	== ...	

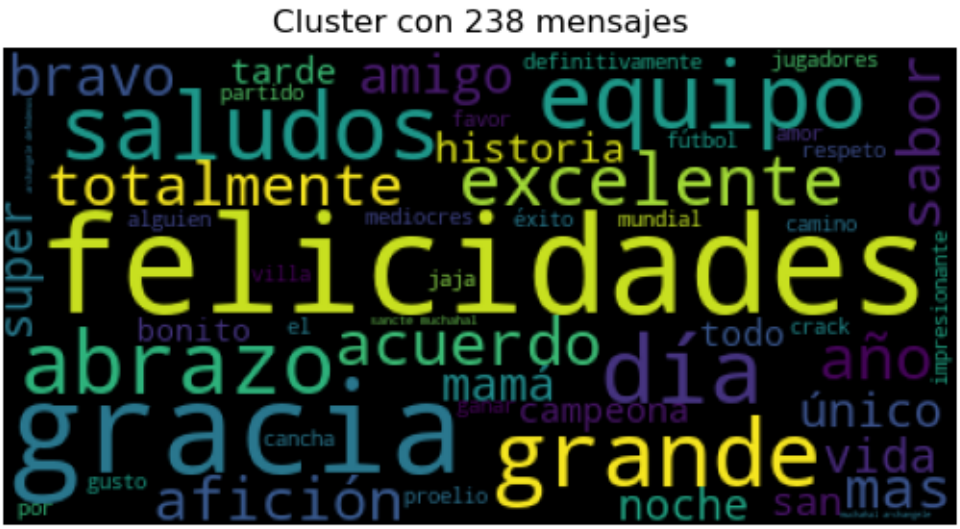
Se eliminan las anomalías y se vuelven a graficar los datos.



Se crean los clusters y se muestran en el gráfico. Los colores indican el cluster.

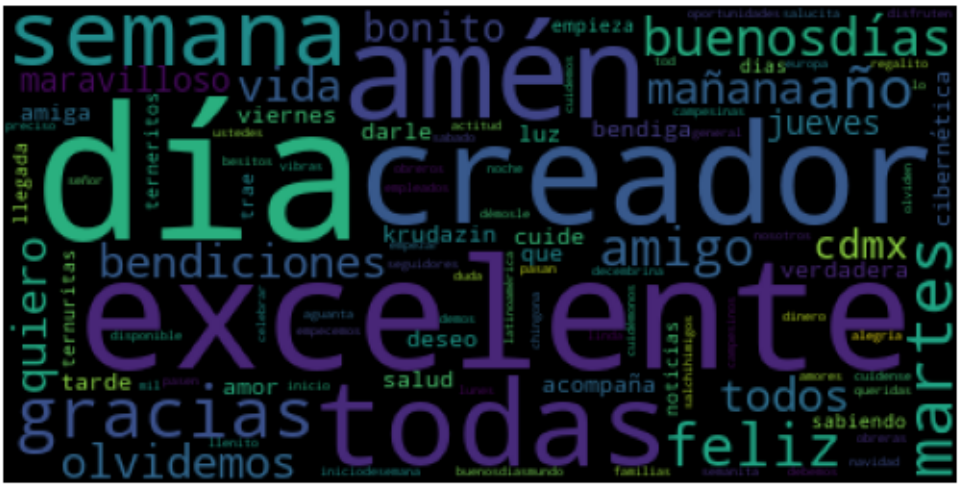


Se analizan los clusters por separado. Para eso usamos la librería wordcloud. El texto se limpió eliminando algunas de las palabras más comunes. Se muestran las nubes de palabras para 4 de los clusters obtenidos.



Cluster con 36 mensajes

Cluster con 26 mensajes



Cluster con 950 mensajes