

Un Estudio Comparativo de PCA y t-SNE en Textos Procesados con NLP

Carlos Rodrigo Pascual

Daniel Vélez Serrano



Índice

Introducción



Algoritmo t-SNE



Caso práctico



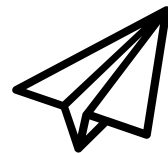
Conclusiones



Futuros trabajos



Introducción



Introducción - Motivaciones



Comparar PCA y t-SNE



Trabajar con corpus textuales



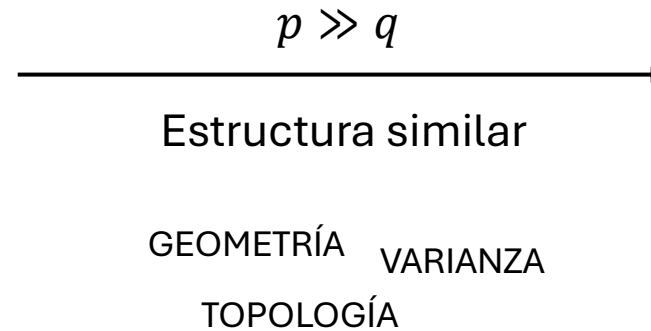
Revisión de la literatura



Un Estudio Comparativo de PCA y t-SNE en Textos Procesados con NLP

Introducción – Reducción de dimensionalidad

	V_1	V_2	V_3	V_4	...	V_p
x_1						
x_2						
x_3						
\vdots						
x_n						

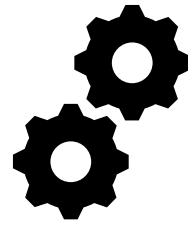


	V_1	V_2	...	V_q
x_1				
x_2				
x_3				
\vdots				
x_n				

¿POR QUÉ?

- Por aumentar la simplicidad de los datos.
- Por aumentar la interpretabilidad de los datos.
- Por reducir costes de procesamiento y/o entrenamiento.

Algoritmo t-SNE



t-SNE - Modelización

1. Modelización de entornos originales con Normal:

$$P_i = \left\{ p_{ij} = \frac{e^{-\frac{\|x_j - x_i\|^2}{2\sigma_i^2}}}{\sum_{k=1}^n e^{-\frac{\|x_k - x_i\|^2}{2\sigma_i^2}}} \mid j = 1, \dots, n, j \neq i \right\} \cup \{ p_{ii} = 0 \}$$



Calibración:

$$\log_2(k) = H(P_i) = - \sum_{j=1}^n p_{ij} \log_2(p_{ij})$$

Entropía de Shannon

2. Modelización de entornos proyectados con t-Student:

$$Q_i = \left\{ q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k=1}^n (1 + \|y_k - y_i\|^2)^{-1}} \mid j = 1, \dots, n, j \neq i \right\} \cup \{ q_{ii} = 0 \}$$


3. Simetrización:

$$P'_i = \left\{ p'_{ij} = \frac{p_{ij} + p_{ji}}{2n} \mid j = 1, \dots, n \right\}$$

$$Q'_i = \left\{ q'_{ij} = \frac{q_{ij} + q_{ji}}{2n} \mid j = 1, \dots, n \right\}$$

t-SNE - Optimización

Función objetivo:

$$F_{obj} = \sum_{i=1}^n D_{KL}(P'_i || Q'_i) = \sum_{i=1}^n \sum_{j=1}^n p'_{ij} \cdot \log\left(\frac{p'_{ij}}{q'_{ij}}\right)$$


Divergencia de Kullback-Leibler

Gradiente:

$$\nabla F_{obj} = \left(\frac{\partial F_{obj}}{\partial y_i} \right)_{i=1}^n = \left(4 \sum_{j=1}^n (p'_{ij} - q'_{ij}) (1 + \|y_i - y_j\|^2)^{-1} (y_i - y_j) \right)_{i=1}^n$$

Descenso de gradiente con momento:

$$\{y_1^t, \dots, y_n^t\} = \mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} - \alpha \nabla F_{obj}(\mathcal{Y}^{(t-1)}) + \beta(t)(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)})$$

Caso práctico



Caso práctico – Construcción del corpus

1	'Ucrania guerra', 'Ucrania Rusia', 'Putin guerra', 'Putin nuclear', 'Ucrania invasión', 'sanciones Rusia', 'OTAN Rusia', 'Zelenski', 'Ucrania fosas', 'Rusia crímenes guerra', 'bomba nuclear', 'Trump guerra Rusia', 'Rusia movilización', 'tropas Rusia', 'movilización Rusia', 'avance Ucrania', 'Nord Stream'
2	'Qatar', 'mundial Qatar', 'mundial 2022', 'futbol Qatar', 'homosexualidad Qatar', 'construcción estadio mundial', 'brazalete selección mundial', 'LGTBI Qatar', 'FIFA corrupción', 'polémica Qatar muertos', 'gais Qatar', 'gays Qatar', 'esposas Qatar', 'cárceles Qatar', 'Derechos Humanos mundial', 'Derechos humanos Qatar', 'DDHH Qatar', 'homosexual daño mental Qatar', 'falsos aficionados', '#Qatar2022', 'Qatar cerveza'
3	'coronavirus', 'COVID', 'vacuna', 'Pfizer vacuna', 'AstraZeneca', 'COVID gripe', 'pandemia', 'variantes covid', 'confinamiento', 'mascarillas', 'restricciones covid', 'vacunación', 'primera ola', 'segunda ola', 'tercera ola', 'cuarta ola', 'quinta ola', 'sexta ola', 'residencias COVID'
4	'crisis energética', 'crisis energía', 'corbata Sánchez', 'inflación', 'recesión', 'precio gas', 'IPC', 'precio energía', 'independencia energética', 'dependencia gas', 'energía nuclear crisis', 'UE gas', 'Alemania gas', 'encarecimiento energía', 'encarecimiento gas', 'invierno Europa', 'precio calefacción'
5	'inmigración España', 'menas', 'Melilla BBC', 'valla Melilla', 'inmigrantes ilegales España', 'inmigración Europa', 'pateras', 'Marlaska Melilla', 'inmigrantes Melilla', 'tragedia Melilla', 'Melilla BBC', 'Ceuta Marruecos', 'Melilla Marruecos', 'delitos extranjeros', 'España Marruecos valla', 'ministerio interior Melilla', 'migrantes muertos', 'mafias inmigración', 'Ceuta Melilla', 'inmigración Barcelona', 'inmigrantes España', 'inmigrantes Barcelona', 'inmigrantes Valencia', 'menas Batán', 'inmigrantes Europa', 'inmigración marroquí', 'imágenes tragedia Melilla', 'videos Melilla'
6	'mujeres iraníes', 'mujeres Irán', 'machismo Irán', 'revolución hijab', 'protestas Irán', 'Mahsa Amini', 'política Irán', 'sanciones Irán', 'represión Irán', 'DDHH Irán', '#IranProtests2022', '#IranRevolution2022', 'clérigos Irán', 'condena muerte Irán', 'Derechos Humanos Irán', 'manifestante Irán', 'dictadura Irán', 'velo Irán', 'pena de muerte Irán', 'Islam Irán', 'feministas Irán', 'activistas iraníes', 'mujeres velo', 'gobierno Irán', '#IranRevolution', 'solidaridad Irán', 'turbante Irán', 'líderes religiosos Irán'
7	'LGTBI SEPE', 'Ley trans', 'ley sí es sí', 'justicia machista', 'ministerio igualdad', 'niños trans', 'médicos trans', 'trans deporte femenino', 'hormonación trans', 'registro trans', 'Irene Montero', 'lenguaje inclusivo Irene Montero', 'fascistas con toga', 'cambio de sexo', 'reducción condena ley', 'ideología de género', 'hazte oír', 'rebajas pena sí es sí'
8	'correos comunista', 'sello PCE', 'sello comunista', 'abogados cristianos sello', 'abogados cristianos correos', 'sello partido comunista', 'juez sello PCE', 'centenario PCE', 'centenario comunista', 'correos neutralidad PCE'
9	'sanidad Madrid', 'sanidad publica Ayuso', 'sanidad profesionales huelga', '#SanidadPublica', '#MadridSeLevantaEl13', 'sanitarios Madrid', 'sanitarios Ayuso', 'huelga sanidad', 'ambulatorios Madrid', 'atención primaria Madrid', 'marea blanca Madrid', 'Ayuso sanidad', 'recortes sanidad PP', 'sanidad madrileña'
10	'delito sedición', 'independentismo', 'lideres independentistas', 'reforma sedición', 'castellano aulas Cataluña', 'Cataluña castellano', 'Cataluña referendum', 'sedición 1-O', 'Cataluña 2017', 'Puigdemont', 'Junqueras', 'delito malversación', 'reforma malversación'

Caso práctico – Estructuración del corpus



Eliminación de *stopwords*

Lematización

Stemming

Detección de bigramas

PoS

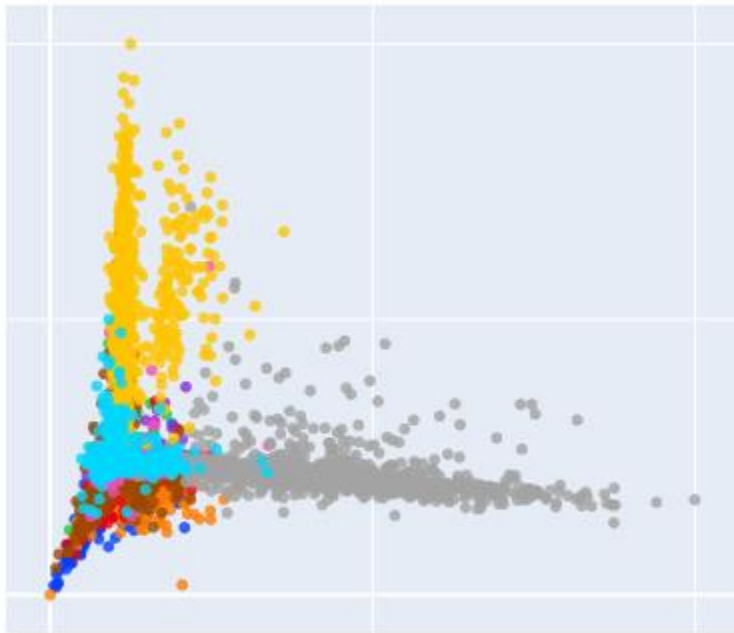
$$TF_{td} = \frac{\text{Número de veces que aparece } d \text{ en } t}{\text{Número total de términos de } d}$$

$$IDF_t = \log \left(\frac{\text{Número total de documentos}}{\text{Número de documentos en los que aparece } t} \right)$$

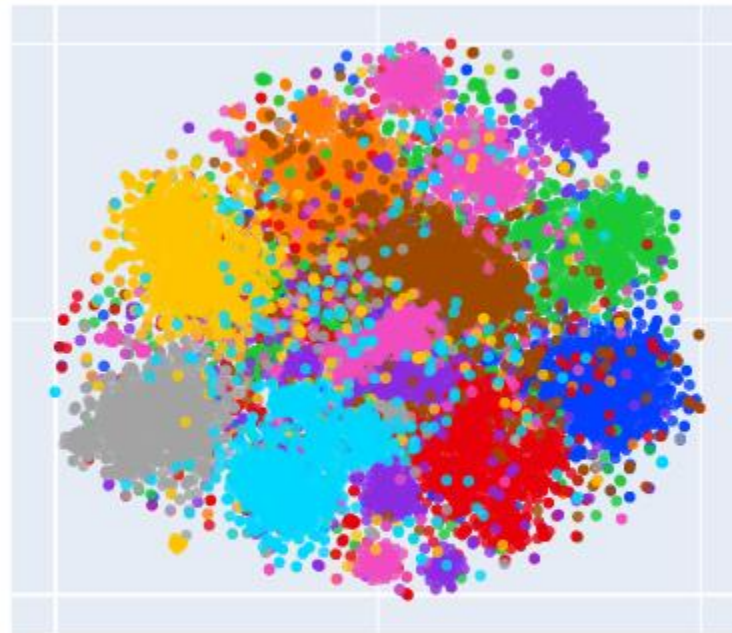
$$TFIDF_{td} = TF_{td} \cdot IDF_t$$

Caso práctico - Proyecciones

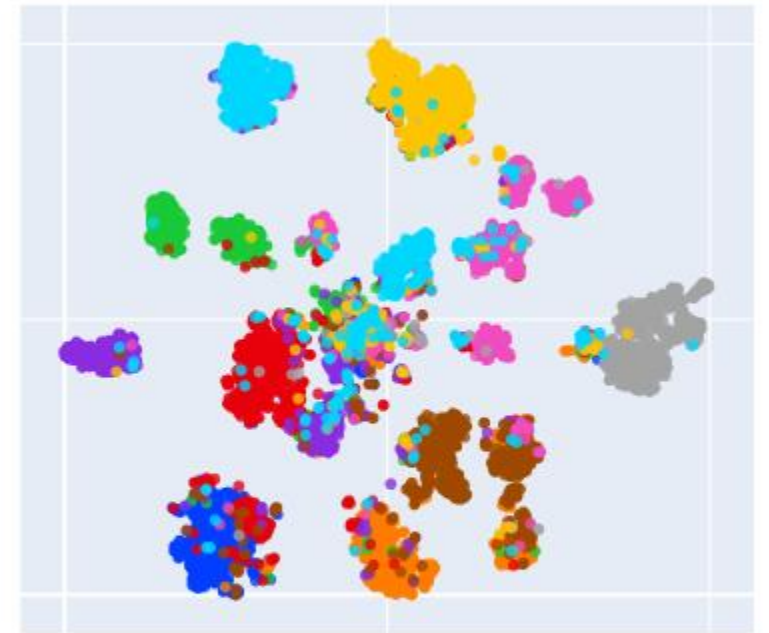
Tweets in embedded spaces by CLUSTER_REAL



PCA



TSNE



PCA_TSNE



Caso práctico – Evaluación de resultados

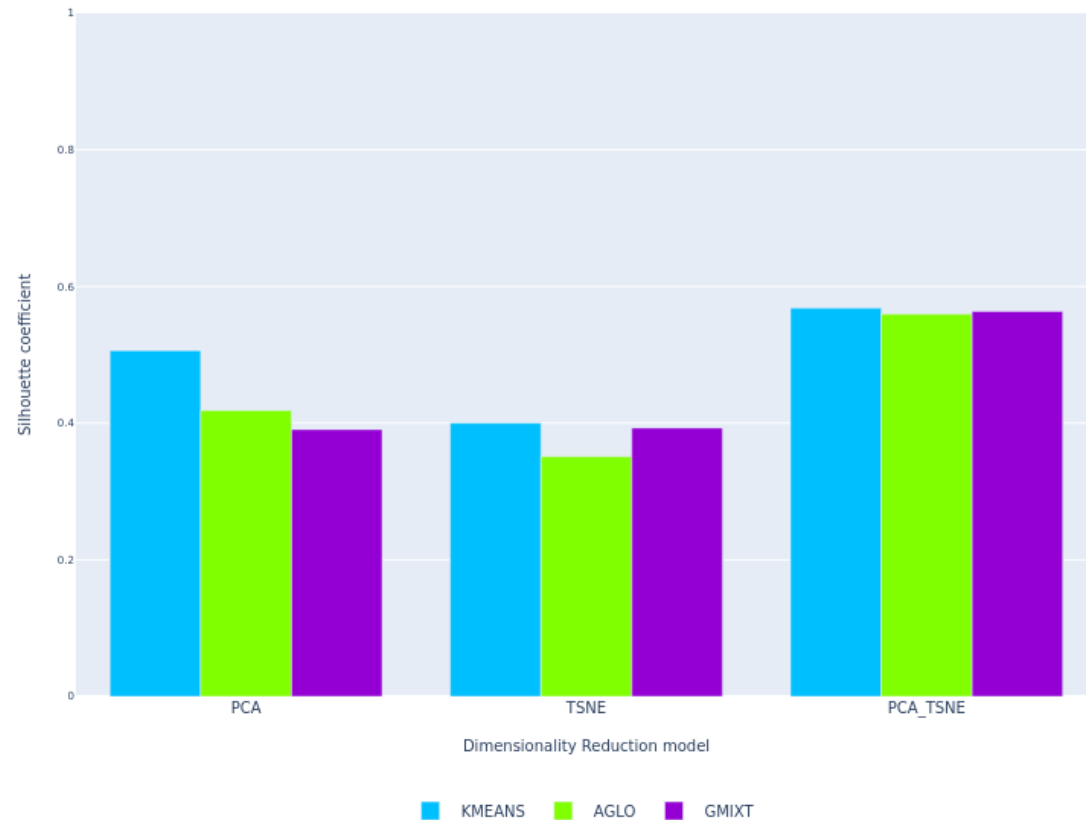
k-Means

Jerárquico
Aglomerativo

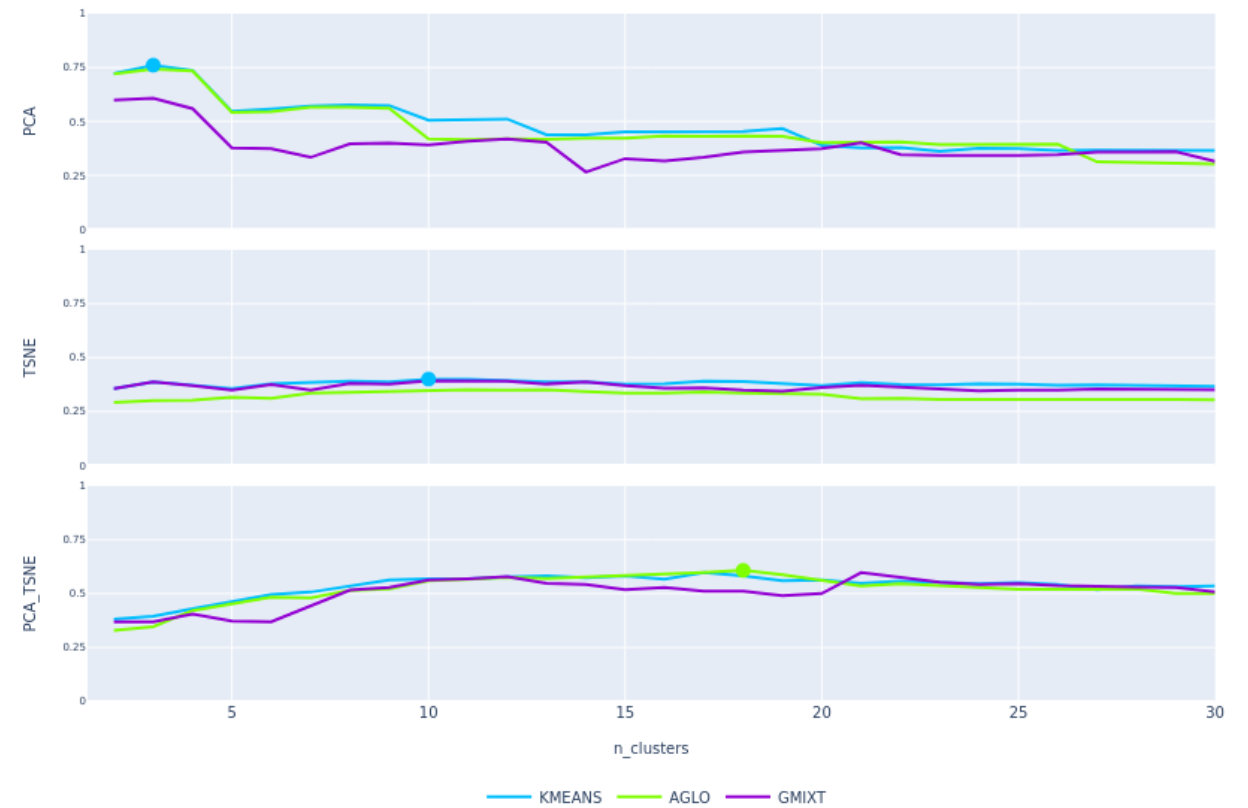
GMM

DBSCAN

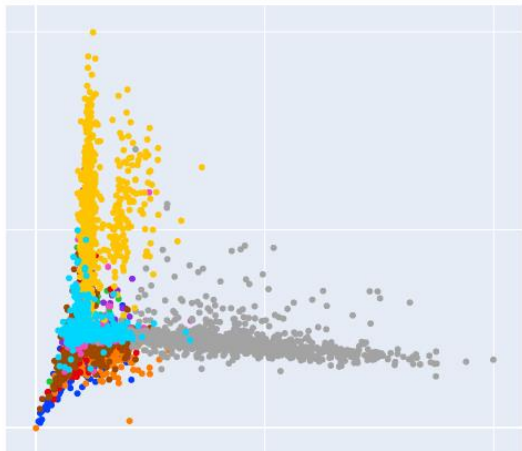
Silhouette coefficients for K-fixed classifications



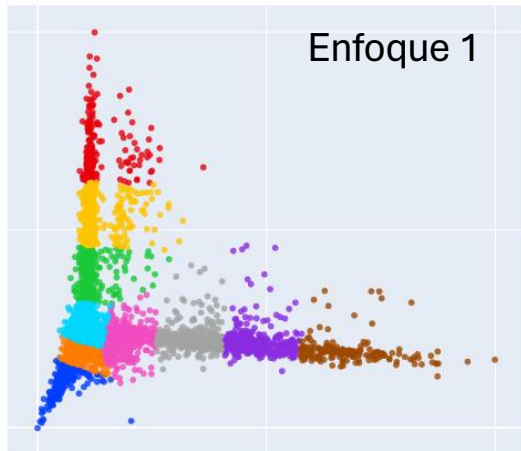
Silhouette coefficients for K-optimized classifications



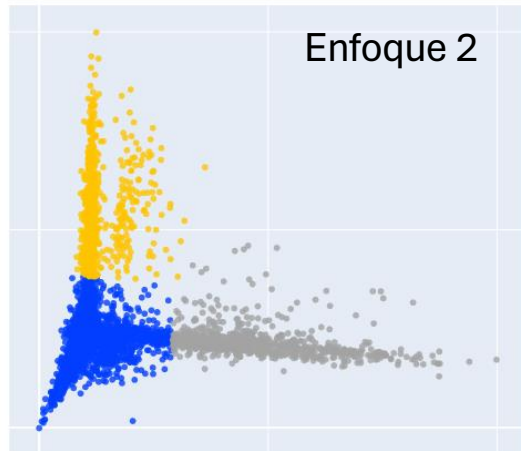
PCA



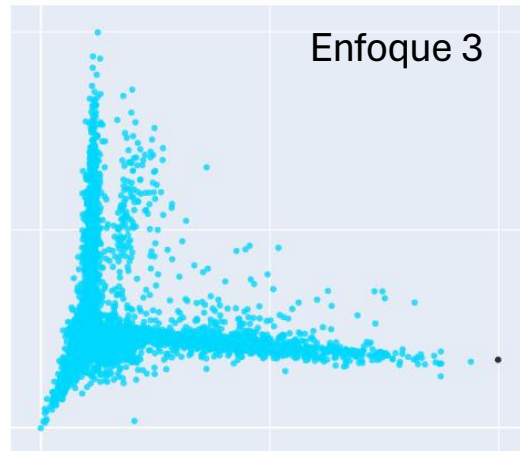
Enfoque 1



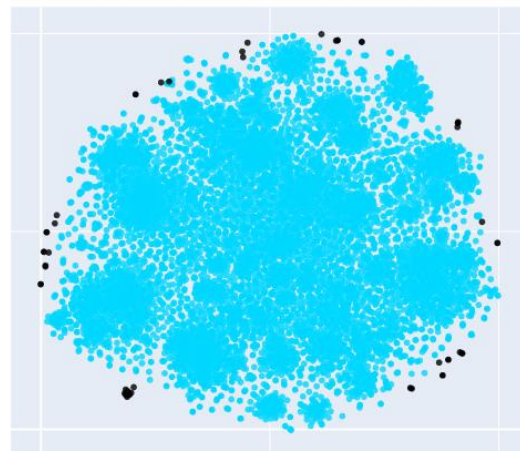
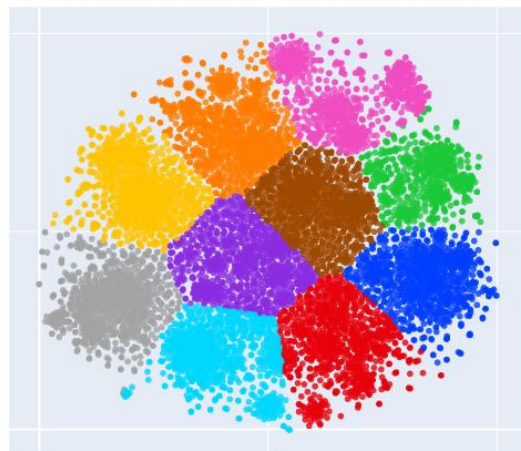
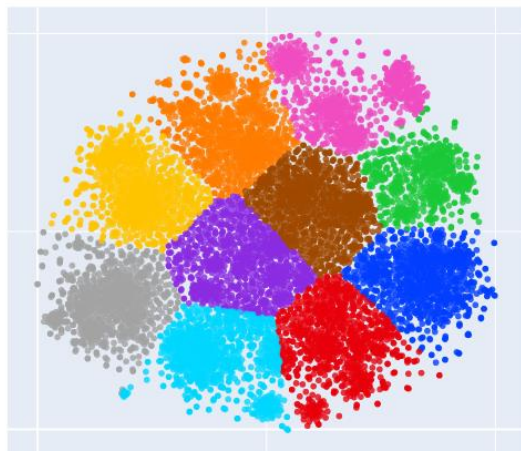
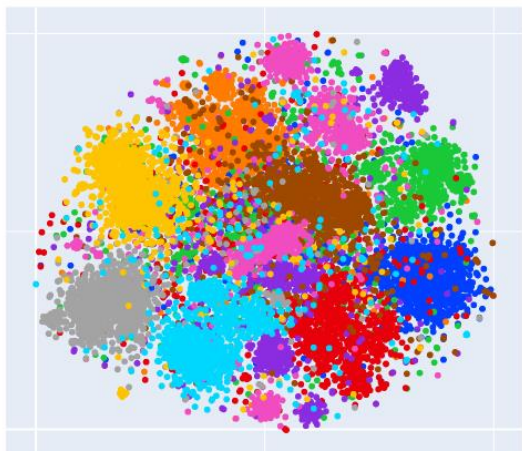
Enfoque 2



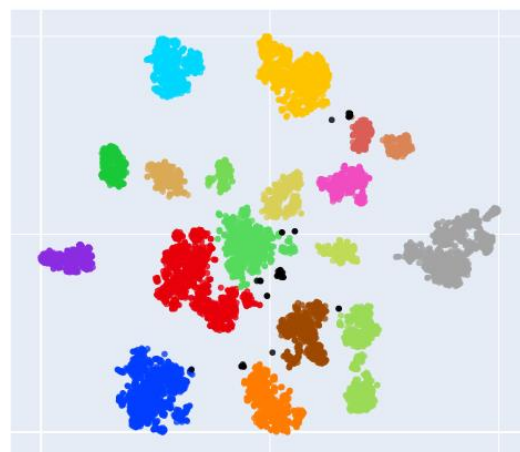
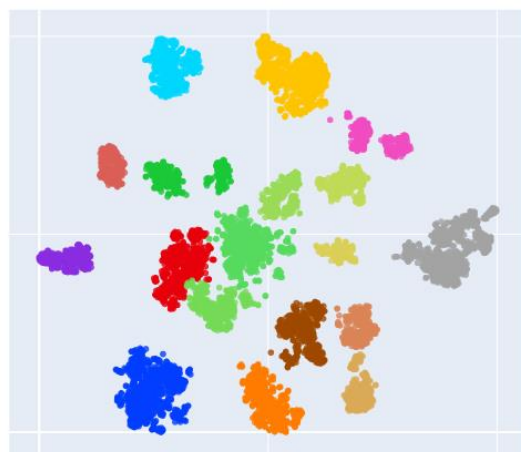
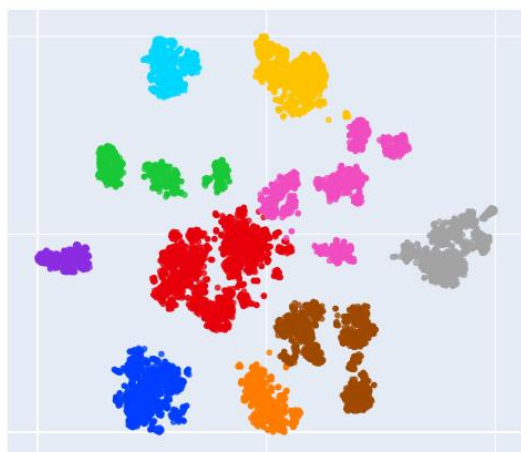
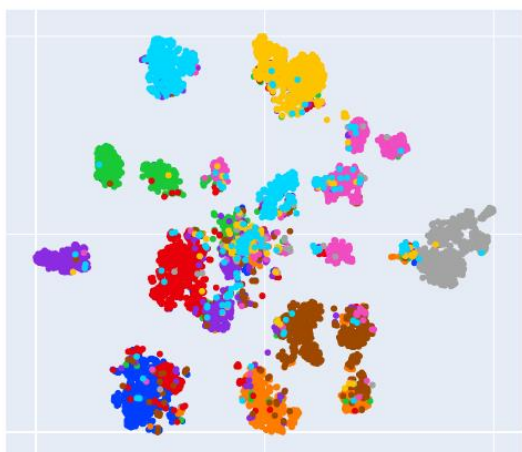
Enfoque 3



t-SNE



PCA + t-SNE



Conclusiones



Conclusiones – Análisis de resultados

Modelo	Enfoque 1		Enfoque 2		Enfoque 3	
	<i>Clusters</i>	<i>Accuracy</i>	<i>Clusters</i>	<i>Accuracy</i>	<i>Clusters</i>	<i>Accuracy</i>
PCA	10	28.52	3	25.99	1 ★	10.00
t-SNE	10	67.64	10	67.64	1 ★	9.98
PCA + t-SNE	10	73.13	18	59.83	18 ★	60.57



Más un cluster de outliers

- El PCA no ha conseguido resultados satisfactorios bajo ninguno de los enfoques.
- El t-SNE ha obtenido resultados satisfactorios y bajo el segundo enfoque ha conseguido que el número de clusters coincida con el esperado.
- El t-SNE no ha conseguido separar los clusters, haciendo que obtenga un mal resultado por el tercer enfoque.
- El PCA + t-SNE ha obtenido resultados satisfactorios bajo los tres enfoques.

Conclusiones

1

El algoritmo t-SNE supera al PCA clásico en la tarea de mantener las estructuras locales y globales de los datos.

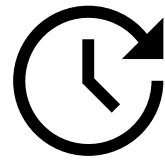
2

La eliminación de ruido previa, permite que el t-SNE capte mejor la estructura de los datos, o al menos, que lo haga más rápido.

3

La combinación de t-SNE y *clustering* constituye una potente opción para abordar el reto del *Topic Modelling* (segmentación de corpus por temáticas).

Futuros trabajos



Futuros trabajos – Profundizar en el análisis

- 1 Probar con diferentes procesos de NLP.
- 2 Probar con diferentes modelos de estructuración de textos.
- 3 Probar con diferentes corpus textuales de diferentes tipos.
- 4 Continuar la búsqueda de mejores calibraciones de los modelos.
- 5 Probar un modelo en el que se concatenen varios t-SNE con diferentes calibraciones.
- 6 Probar a combinar t-SNE con un modelo diferente al PCA.

Un Estudio Comparativo de PCA y t-SNE en Textos Procesados con NLP

Carlos Rodrigo Pascual

Daniel Vélez Serrano

