

Aprendizaje Automático No Supervisado

Alberto Barbado González

Presentación - Actividad 2

Actividad 2

- ▶ **Objetivo:** Familiarizarse con las **técnicas no supervisadas de extracción de features**
- ▶ Entender los métodos de t-SNE y PCA.
- ▶ Realizar la reducción de dimensionalidad utilizando t-SNE y PCA.
- ▶ Comparar los resultados obtenidos con ambos métodos.
- ▶ Aplicar LDA para reducir el dataset a una bolsa de palabras por cada etiqueta.

Dataset

- Textos (comentarios) de RRSS (Facebook, X...), páginas web de periódicos...
- Incluye una clasificación de esos comentarios en 2 categorías: Odio y No odio.



Detalle del Proyecto del que viene el dataset

<https://github.com/esaidh266/Algorithm-for-classifying-hate-expressions-by-intensities-in-Spanish?tab=readme-ov-file>

Dataset

Odio / No odio Preprocesamiento

Proceso de
limpieza

- Eliminación de nulos
- URLs
- Emojis
- Filas vacías

Homogenización

- Minúsculas
- Números
- Stopwords
- Tokenización
- Lematización...

Extracción de
características

- Palabras positivas
- Palabras negativas
- Bigramas comunes
- Mención usuarios
- pysentimiento

Estandarización

- StandardScaler

Creación de
características
sintéticas

- AB AC AD AE
- BC BD BE
- CD CE
- DE

Dataset

	A	B	C	D	E	comentario	label	A_t	B_t	C_t	...	Valor_1	Valor_2	Valor_3	Valor_4	Valor_5	Valor_6	Valor_7	Valor_8	Valor_9	Valor_10	
	0	2	64	30	0	2	pandemia,originado,covid,cierto,incidencia,aba...	0.0	1.851102	2.759647	7.145831	...	5.108388	13.227660	-0.771127	0.815665	19.719970	-1.149606	1.216004	-2.976790	3.148722	-0.183560
	1	4	70	21	0	0	ser,mes,larga,espera,llegar,momento,siempre,pr...	0.0	3.990202	3.054765	4.877255	...	12.189130	19.461233	-1.662227	-2.119191	14.898871	-1.272546	-1.622382	-2.031754	-2.590305	0.221244
	2	4	88	50	0	0	cartagena,san,sebastiar,fuengirola,irun,orense...	0.0	3.990202	3.940120	12.187108	...	15.721875	48.629021	-1.662227	-2.119191	48.018675	-1.641365	-2.092593	-5.076872	-6.472559	0.221244
	3	3	38	21	0	0	pleno,dia,verano,calor,plan,mas,apetecibl,disf...	0.0	2.920652	1.480801	4.877255	...	4.324903	14.244765	-1.216677	-1.551155	7.222244	-0.616868	-0.786452	-2.031754	-2.590305	0.221244
	4	0	59	17	0	0	pasado,junio,celebro,dia,luchar,frente,leishma...	0.0	-0.287998	2.513715	3.869000	...	-0.723945	-1.114264	0.119973	0.152955	9.725563	-1.047156	-1.335031	-1.611737	-2.054821	0.221244
	
9995	0	1	2	0	6	idiota,noticia,solo,idiota,poder,publicar,	1.0	-0.287998	-0.339096	0.088042	...	0.097659	-0.025356	0.119973	-0.686619	-0.029855	0.141260	-0.808442	-0.036676	0.209901	-0.993166	
9996	0	0	2	0	6	idiota,util,criticar,poder,dar,lujo,dar,critic...	1.0	-0.287998	-0.388282	0.088042	...	0.111824	-0.025356	0.119973	-0.686619	-0.034185	0.161750	-0.925708	-0.036676	0.209901	-0.993166	
9997	0	0	4	0	6	izaguirri,pobre,diablo,querrar,lucir,bolso,po...	1.0	-0.287998	-0.388282	0.592169	...	0.111824	-0.170544	0.119973	-0.686619	-0.229929	0.161750	-0.925708	-0.246684	1.411797	-0.993166	
9998	0	1	1	0	0	jajajajaj,dejar,tirar,mugre,subir,ahora,parasi...	1.0	-0.287998	-0.339096	-0.164022	...	0.097659	0.047238	0.119973	0.152955	0.055619	0.141260	0.180093	0.068328	0.087112	0.221244	
9999	0	2	2	0	0	envidio,corroe,comunista,trabajen,emprender,cr...	1.0	-0.287998	-0.289909	0.088042	...	0.083493	-0.025356	0.119973	0.152955	-0.025524	0.120770	0.153971	-0.036676	-0.046759	0.221244	
10000 rows × 22 columns																						

- Se dispone del comentario original (Limpiado + Homogeneizado).
- Se dispone de la Extracción de Características:

Proceso de extracción de características

- Conteo de palabras positivas (A)
- Conteo de palabras negativas (B)
- Conteo del número de bigrams más comunes (C)
- Conteo del número de menciones a otros usuarios (D)
- Categoría del sentimiento según librería 'pysentimiento' en español (E)
- Estandarización de las características (A_t,..E_t)
- Combinación de características f1*fi (iA..iE) (Valor1,..Valor10).

Dataset

Proceso de extracción de características

- Conteo de palabras positivas (A)
- Conteo de palabras negativas (B)
- Conteo del número de bigrams más comunes (C)
- Conteo del número de menciones a otros usuarios (D)
- Categoría del sentimiento según librería 'pysentimiento' en español (E)
- Estandarización de las características (A_t,..E_t)
- Combinación de características $t1*ti$ (iA..iE) (Valor1,..Valor10).

- Se proporciona para todas esas características el valor estandarizado
- Por tanto, NO ES NECESARIO usar de manera conjunta A, B... con A_t, B_t....
- Recomendación: Usad las variables originales (A, B,...) y así podéis elegir vosotros la técnica de estandarización que queráis

Objetivos

- Usar PCA y t-SNE para **visualizar** los datos
- Usar PCA y t-SNE **como parte de un pipeline** donde luego se entrene un modelo supervisado para predecir odio/no odio (Nota: t-SNE comentamos que NO ES habitual usarlo de esta manera, pero en la tarea lo usaremos así a modo de prueba).
- **Aplicar LDA** (Topic Modeling) sobre este conjunto de datos

Calificación

Tareas que realizar

- ▶ Reducción de dimensionalidad con t-SNE y PCA:
 - Completa el código proporcionado en el Notebook.
- ▶ Punto adicional:
 - Aplicación de LDA: utiliza la técnica de LDA para generar una bolsa de palabras para cada una de las etiquetas.

Entrega: Notebook Python.

Rúbrica

Análisis de reducción de dimensionalidad: PCA y t-SNE	Descripción	Puntuación máxima (puntos)	Peso %
Criterio 1	Realiza correctamente el código solicitado para reducir la dimensionalidad con t-SNE	3	30 %
Criterio 2	Realiza correctamente el código solicitado para reducir la dimensionalidad con PCA	3	30 %
Criterio 3	Analiza y responde correctamente a las preguntas realizadas	4	40 %
Criterio adicional	Puede reemplazar el criterio 2 por este. Realiza correctamente el código solicitado para reducir la dimensionalidad con LDA	3	30 %
		10	100 %

- **Opción 1:** No hacer LDA, pero sí PCA (LDA es optativo, se puede seguir sacando 10).
- **Opción 2:** Hacer LDA en lugar de PCA (se puede seguir sacando 10).
- **Opción 3:** Hacerlo todo (LDA, PCA y t-SNE). En este caso LDA da un punto extra como máximo (de manera que os ayudaría a sacar mayor nota. Siempre sumaría, nunca restaría)
- **Recomendación:** Intentad hacerlo todo, que siempre va a sumar más 😊

UNIVERSIDAD
INTERNACIONAL
DE LA RIOJA

unir

www.unir.net