

A stylized sunburst graphic in shades of purple and blue, located in the top-left corner of the slide.

# Minería de textos

Máster Online en Ciencia de Datos

**UCO**  
ONLINE

Four horizontal bars of varying lengths in yellow and red, located at the bottom of the slide.

## Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.





# Agrupamiento de textos

*Introducción al agrupamiento de documentos*

# Introducción

- El agrupamiento (*clustering*) determina cómo se deben reunir documentos en grupos que compartan características
- Es habitual hacer uso de los **algoritmos de agrupamiento más habituales**:
  - **Modelos jerárquico**. Se basan en el concepto de que objetos similares estarán más cercanos a objetos relacionados en el espacio del vector que los no relacionados. Permite la visualización en un diagrama de árbol (*dendrogram*) mostrando una jerarquía exhaustiva de clústeres
  - **Modelos basados en centroide**. Construyen grupos que tienen un elemento central representativo para cada clúster, de forma que exhibe las características distintivas que lo distingue del resto de grupos. k-means y k-medoids son ejemplos representativos

# Introducción

- **Modelos basados en distribución.** Aplican distribuciones de probabilidad de forma que objetos con distribuciones similares estén en el mismo grupo. Estos modelos permiten capturar dependencias y correlaciones entre características y atributos, aunque son propensos al *over-fitting*
- **Modelos basados en densidad.** Generan clústeres a partir de puntos de datos que se agrupan juntos en áreas de alta densidad, lo que podría ocurrir de forma aleatoria en zonas dispersas del espacio del vector. Estas zonas dispersas se tratan como ruido o sirven de frontera.
- El agrupamiento puede tener similitudes con la obtención de n-gramas

# n-gramas Vs. clustering de documentos

- El agrupamiento puede verse con **similitudes con la obtención de n-gramas**
- La **ventana deslizando** se utiliza para determinar si 2 o más palabras son similares
  - Cuenta las apariciones conjuntas de un número de palabras y se mueve por el texto
  - El conteo se refleja en una tabla o matriz que muestra la cercanía en la relación entre palabras
  - Trata todo el texto como una única unidad, sin diferenciar inicios, rupturas, paradas, etc. → solo importa la secuencia de palabras → **modelos secuenciales**
- El método de la **bolsa de palabras** (*bag of words*) utiliza el vector de palabras
  - También realiza un conteo por filas (documento) como bloques de texto
  - En este caso, la secuencia no es determinante
  - Una bolsa de palabras es la metáfora de un espacio físico alojando palabras relacionadas
  - Muy criticado por no considerar aspectos clave relacionados con el significado de las palabras

# n-gramas Vs. clustering de documentos

Se puede considerar que los métodos de obtención de n-gramas realizan un agrupamiento de palabras (*word clustering*) frente a métodos requeridos para poder agrupar documentos (*document clustering*)

# Clustering de palabras

- ¿**Por qué deberíamos agrupar (clustering) si ya hemos creado una matriz TDM?**  
Esta matriz ya está agrupando variables en forma de tabla
- El **análisis factorial** es el método para generación de n-gramas más utilizado cuando el texto está dispuesto en forma tabular
  - Los factores ofrecen una **vista numérica**, simplificando los datos al conjunto de relaciones más simple
  - El análisis factorial es un **método de reducción de datos**





# Clustering de palabras

- Los **métodos de clustering** ofrecen patrones de similitud en los datos más sutiles
  - Los métodos de **clustering trabajan sobre las filas de la matriz**, es decir, los documentos
  - Es **importante escoger adecuadamente el método** porque tienden a tener estructuras “preferidas”, esto es, tienden a estructuras en los datos que otros métodos podrían describir mejor
  - En texto es **complicado definir la “similitud”**, por lo que distintos métodos pueden llegar a distintas conclusiones sobre el mismo documento
- El **objetivo subyacente** a los métodos de analítica de textos es extraer información de forma concisa y condensada
  - Todo proceso de **reducción implica pérdidas de información** respecto al original → en texto, **no podemos cuantificar** cuánta información perdemos ni qué información necesitamos mantener
  - El **problema reside en la transformación de texto a variables** → no hay métricas de pérdida
  - Esto hace **difícil decidir qué método de clustering de palabras es mejor** que otro

A stylized sunburst graphic in shades of purple and blue, located in the top-left corner of the slide. It features a semi-circle on the left with several rays extending outwards to the right.

# Agrupamiento de textos

*Clustering de documentos*

# Clustering de documentos

- Puede interesarnos **categorizar el documento completo** en base a la información aportada por los términos que lo contienen
- Recordemos que un documento puede venir representado por algo muy breve (tweet) o algo muy extenso (libro completo)

# Ejemplo práctico

- Vamos a realizar clustering sobre una serie de tweets de Elon Musk
- Lo primero que haremos, una vez hayamos cargado los tweets, será preprocesarlos siguiendo el pipeline explicado en los temas anteriores

index	Tweet
14	Prev ideas for paying ~\$10B dev cost incl. Kickstarter & collecting underpants, which turned out to be um... less lucrative than expected
36	Tesla Semi truck unveil & test ride tentatively scheduled for Oct 26th in Hawthorne. Worth seeing this beast in person. It's unreal.
52	China, Russia, soon all countries w strong computer science. Competition for AI superiority at national level most likely cause of WW3 imo.
55	Putting together SpaceX rocket landing blooper reel. We messed up a lot before it finally worked, but there's some epic explosion footage ?
67	To be clear, a Hyperloop passenger version wouldn't have intense light strobe effect (just for testing), nor uncomfortable acceleration
70	Will run the SpaceX pusher sled later this week and see what it can do
79	Btw, high accel only needed because tube is short. For passenger transport, this can be spread over 20+ miles, so no speed limit
82	Might be possible to go supersonic in our test Hyperloop tube, even though it's only 0.8 miles long. Very high accel/decel
102	Max recovered booster velocity was Mach 7.9 (BulgarianSat). Energy is velocity squared, so this is a bigger difference than you think
103	Max velocity: Mach 6.9 Max altitude: 247 km Highest so far, but velocity matters much more
108	Touchdown: Vertical Velocity (m/s): -1.47 Lateral Velocity (m/s): -0.15 Tilt (deg): 0.40 ? Lateral position: 0.7m from target
124	Watching eclipse with sunglasses on through the Model S glass roof. Wow!
129	Pics of SpaceX spacesuit developed for NASA commercial crew program coming out next week. Undergoing ocean landing tests
140	Wigs me out too much. Deleting ...
152	Would like to express our appreciation to Microsoft for use of their Azure cloud computing platform. This required massive infrastructure

*Cada tweet se corresponde con un documento y, por lo tanto, va a ser preprocesado de manera independiente.*

```

29 def preprocess_tweet(tweet_text):
30     # pasamos todo el texto a minúscula
31     tweet_text = tweet_text.lower()
32     # tokenizamos
33     words = [word_tokenize(sentence) for sentence in sent_tokenize(tweet_text)]
34     words = [x for xs in words for x in xs] # flatten the list
35     # eliminamos las stop words y los signos de puntuación
36     stop_words = nltk.corpus.stopwords.words("english")
37     stop_words.extend(list(punctuation))
38     words = [word for word in words if word not in stop_words]
39     # realizamos el stemming
40     words = [PorterStemmer().stem(word) for word in words]
41     return " ".join(words)

```

# Ejemplo práctico

- Utilizaremos el TF-IDF de los términos de los tweets como entrada del algoritmo de clustering
- El objetivo es que cada clúster quede caracterizado por una serie de términos o n-gramas

	capitol	capitol build	captain	captain hercul	car
nobodi like regul everyth car plane food drug etc 's danger public regul ai	0.0	0.0	0.0	0.0	0.102583
openai first ever defeat world 's best player competit esport vastli complex tradit board game like chess go	0.0	0.0	0.0	0.0	0.000000
could n't believ incred inspir creativ	0.0	0.0	0.0	0.0	0.000000
project loveday winner ...	0.0	0.0	0.0	0.0	0.000000
want happen fast pleas let local feder elect repres know make big differ hear	0.0	0.0	0.0	0.0	0.000000



Esta figura es un ejemplo reducido con solo 5 tweets y n-gramas. Cabe destacar que se obtiene una matriz muy sparse

## Clúster #1 – Key terms

tesla, car, model, like,  
autopilot



Elon Musk ✓  
@elonmusk

En respuesta a @elonmusk

No signature series for Model 3. Same reservation amount for all.

## Clúster #2 – Key terms

rocket, launch, falcon,  
good, land



Elon Musk ✓  
@elonmusk

Rocket is extra toasty and hit the deck hard (used almost all of the emergency crush core), but otherwise good

```
28 # ejecutamos k-means
29 km = KMeans(n_clusters=2)
30 km.fit(feature_matrix)
31 clusters = km.labels_
```



Hemos fijado  
el número de  
clústers a 2

A stylized sunburst graphic in shades of purple and blue, located in the top-left corner of the slide. It features a semi-circle on the left with several rays extending outwards to the right.

¡Gracias!

**UCO**  
ONLINE

A horizontal bar at the bottom of the slide consisting of three equal-width rectangles of yellow, red, and blue, representing the colors of the Spanish flag.