Aprendizaje no supervisado

Agustín D. Delgado

Reading Group Doctorandos NLP&IR@UNED
15 de noviembre de 2017, Madrid

Índice de contenidos

- **Introducción**
- **Algoritmos de *clustering***
- **Técnicas de *topic modelling***
- **Técnicas de reducción de dimensionalidad**
- **Conclusiones**

Índice de contenidos

Introducción

- **Aprendizaje supervisado**
  - Se recibe un conjunto de pares $(\bar{\imath}, \bar{o}) \in I \times O$ y se aprende/infiere una función $f: I \rightarrow O$.
  - **Ventajas:** generalización para nuevos datos de entrada, métodos eficientes.
  - **Desventajas:** coste y dependencia de la anotación de datos, sesgos de los datos de entrenamiento, *overfitting*.
  - Asociado a técnicas de **clasificación** (clases discretas) y **regresión** (clases continuas).

- **Aprendizaje no supervisado**
  - Se recibe únicamente un conjunto de datos de entrada I y se aprende/infiere su estructura.
  - **Ventajas:** evita la necesidad y el coste de la anotación de datos.
  - **Desventajas:** dependencia del problema, métodos más costosos computacionalmente.
  - Asociado a **agrupamiento (*clustering*)**, ***topic modelling*** y **reducción de dimensionalidad**.

Introducción

- **Aprendizaje supervisado**
  - Se recibe un conjunto de pares $(\bar{\imath}, \bar{o}) \in I \times O$ y se aprende/infiere una función $f: I \rightarrow O$.
  - **Ventajas:** generalización para nuevos datos de entrada, métodos eficientes.
  - **Desventajas:** coste y dependencia de la anotación de datos, sesgos de los datos de entrenamiento, *overfitting*.
  - Asociado a técnicas de **clasificación** (clases discretas) y **regresión** (clases continuas).

- **Aprendizaje no supervisado**
  - Se recibe únicamente un conjunto de datos de entrada I y se aprende/infiere su estructura.
  - **Ventajas:** evita la necesidad y el coste de la anotación de datos.
  - **Desventajas:** dependencia del problema, métodos más costosos computacionalmente.
  - Asociado a **agrupamiento (*clustering*)**, ***topic modelling*** y **reducción de dimensionalidad**.

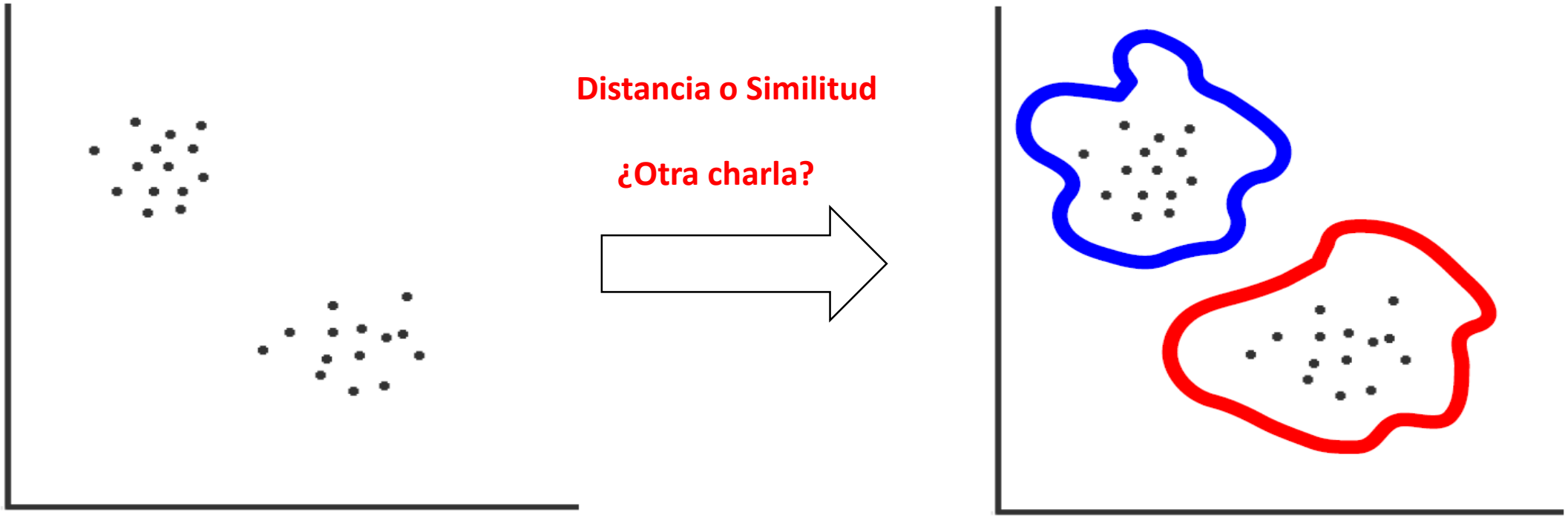
Índice de contenidos

- **Introducción**
- **Algoritmos de *clustering***
- **Técnicas de *topic modelling***
- **Técnicas de reducción de dimensionalidad**
- **Conclusiones**

Índice de contenidos

- Introducción
- **Algoritmos de *clustering***
- Técnicas de *topic modelling*
- Técnicas de reducción de dimensionalidad
- Conclusiones

Agrupamiento (*Clustering*)



Diferencia c.r.a. la clasificación: no se conoce necesariamente el número de clases (*clusters*).

Clasificación de algoritmos de *clustering*

- **Algoritmos jerárquicos**
- **Algoritmos de partición (o planos)**
- **Otros tipos:**

Clasificación de algoritmos de *clustering*

- **Algoritmos jerárquicos**
  - Presentan los *clusters* generados a partir de una estructura jerárquica dividida en niveles de especialización.
  - **Tipos:**
    - **Algoritmos aglomerativos:** consideran que inicialmente cada objeto conforma un *cluster*.
    - **Algoritmos divisivos:** consideran que inicialmente hay un único *cluster* que contiene a todos los objetos.
  - **Ventajas:** no requieren conocer el número de clusters.
  - **Desventajas:** coste computacional elevado (cuadrático o superior c.r.a. número de objetos).
  - **Ejemplos:** Algoritmo Jerárquico Aglomerativo (HAC), Minimal Spanning Tree.
- **Algoritmos de partición (o planos)**
- **Otros tipos:**

Clasificación de algoritmos de *clustering*

- **Algoritmos jerárquicos**

- **Algoritmos de partición (o planos)**
  - Los *clusters* generados consisten en una partición matemática del conjunto de objetos.
  - **Tipos:**
    - **Requieren el número de *clusters*:** por ej. k-means, k-medoids, etc.
    - **No requieren el número de *clusters*:** por ej. X-means, SPC, QT.
  - **Ventajas:** eficientes (lineales c.r.a número de objetos).
  - **Desventajas:** indeterminismo y/o requieren información a priori.
- **Otros tipos:**

Clasificación de algoritmos de *clustering*

- **Algoritmos jerárquicos**

- **Algoritmos de partición (o planos)**

- **Otros tipos:**
  - Basados en densidad (por ej. DBSCAN).
  - *Spectral clustering*.
  - Estadísticos (por ej. Expectation-Maximization, EM).
  - *Fuzzy clustering*. Basados en lógica/reglas *fuzzy*.
  - Basados en grafos (por ej. *Correlation Clustering*).
  - Basados en redes neuronales (por ej. Mapas auto-organizativos, SOM).
  - etc.

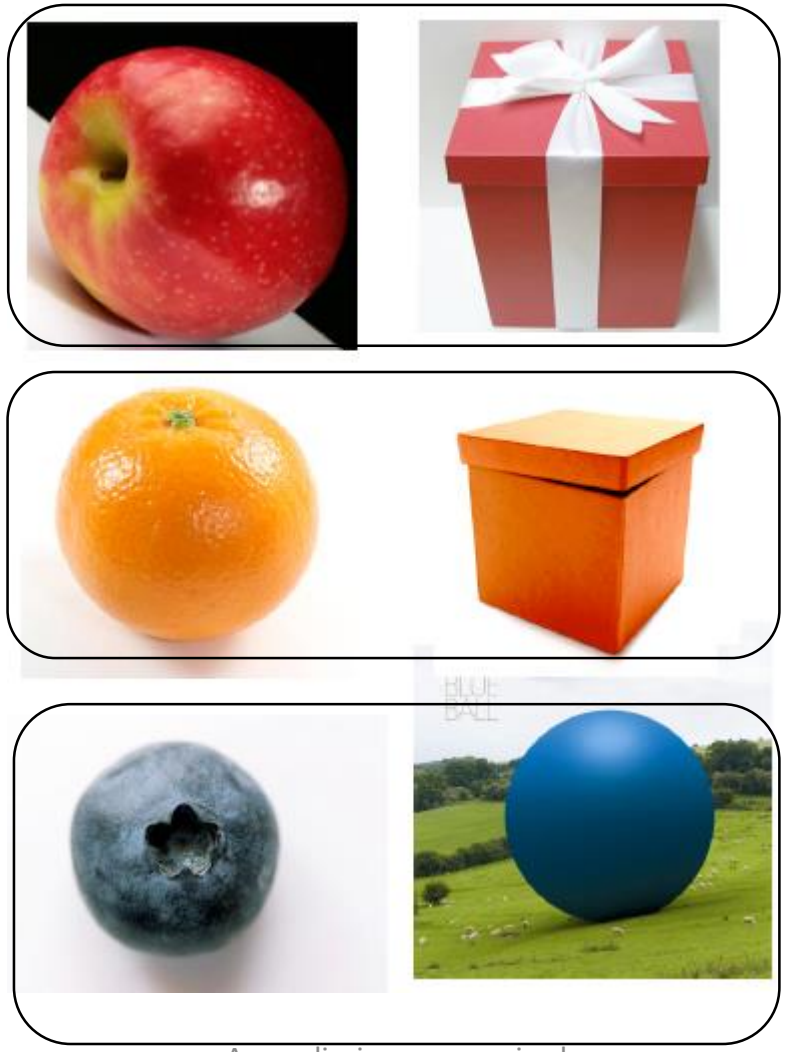
Subjetividad del *clustering*

Subjetividad del *clustering*



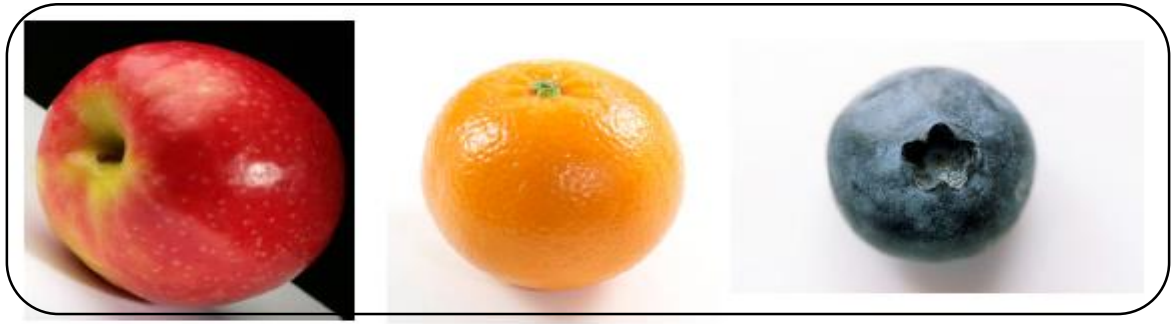
Forma


Subjetividad del *clustering*



Color

Subjetividad del *clustering*



Frutas

Cajas

Tipo de objeto

Balones

Índice de contenidos

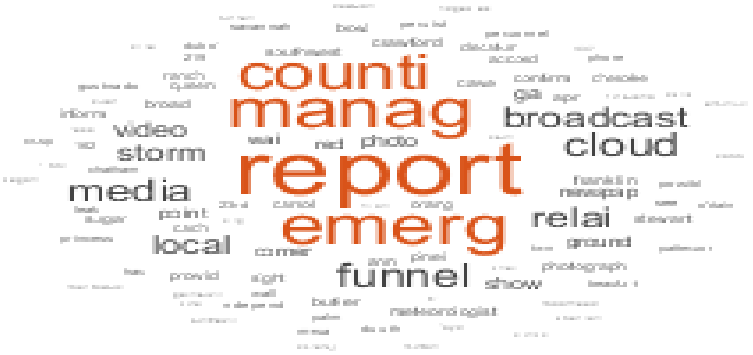
- **Introducción**
- **Algoritmos de *clustering***
- **Técnicas de *topic modelling***
- **Técnicas de reducción de dimensionalidad**
- **Conclusiones**

Índice de contenidos

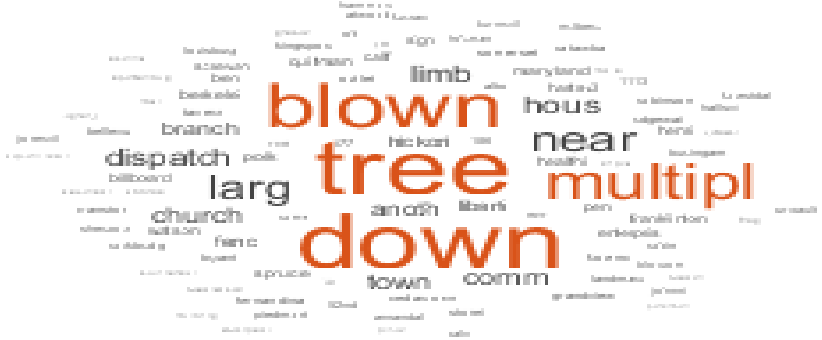
Topic models

Extracción de las temáticas (*topics*) tratadas en una colección de documentos.

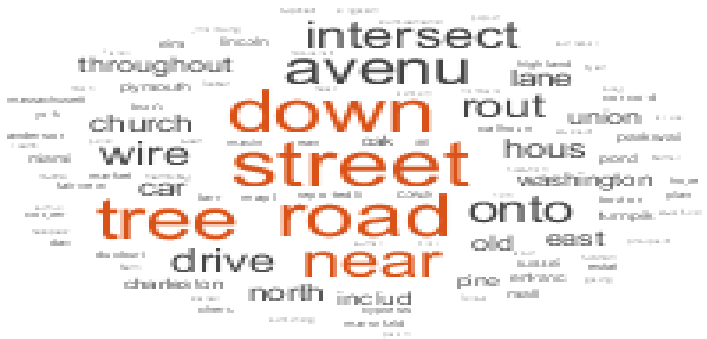
Topic: 1



Topic: 2



Topic: 3



Topic: 4

Topic models

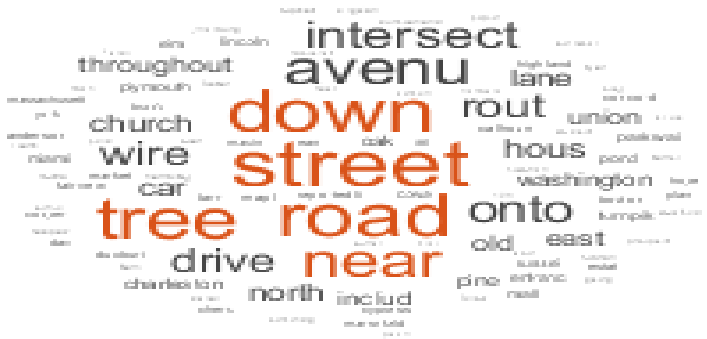
Extracción de las temáticas (*topics*) tratadas en una colección de documentos.

Topic: 1

Topic: 2

Topic = conjunto de palabras (se asume *bolsa de palabras*)

Topic: 3



Topic: 4

Hipótesis de las técnicas de *topic modelling*

- Una colección de documentos trata de un conjunto de k *topics*.
- Un documento puede referirse a varios *topics*.
- Un *topic* se compone de un conjunto de palabras.
- Una palabra pertenece a cada *topic* con una cierta probabilidad.

Representación mediante *topic models*

- Una colección de documentos trata de un conjunto de k *topics*.
  - El número k de *topics* se establece de antemano o se estima mediante alguna técnica.
- Un documento puede referirse a varios *topics*.
- Un *topic* se compone de un conjunto de palabras.
- Una palabra pertenece a cada *topic* con una cierta probabilidad.

Representación mediante *topic models*

- Una colección de documentos trata de un conjunto de k *topics*.
- Un documento puede referirse a varios *topics*.
  - Documento ↔ distribución de probabilidad multinomial de varios *topics*.
- Un *topic* se compone de un conjunto de palabras.
- Una palabra pertenece a cada *topic* con una cierta probabilidad.

Representación mediante *topic models*

- Una colección de documentos trata de un conjunto de k *topics*.
- Un documento puede referirse a varios *topics*.
- Un *topic* se compone de un conjunto de palabras.
  - *Topic* ↔ distribución de probabilidad multinomial de varias palabras.
- Una palabra pertenece a cada *topic* con una cierta probabilidad.

Métodos de *topic modelling*

- Una colección de documentos trata de un conjunto de k *topics*.
- Un documento puede referirse a varios *topics*.
  - Documento ↔ **distribución de probabilidad multinomial** de varios *topics*.
- Un *topic* se compone de un conjunto de palabras.
  - *Topic* ↔ **distribución de probabilidad multinomial** de varias palabras.
- Una palabra pertenece a cada *topic* con una cierta probabilidad.

Métodos:

pLSA (Analisis de Semántica Latente Probabilístico): distribución de probabilidad multinomial.

LDA (Latent Dirichlet Allocation): distrib. probabilidad multinomial y multivariable (distrib. Dirichlet).
Generalización de pLSA.
Multivariable: permite estudiar el comportamiento de varias variables al mismo tiempo.

hLDA (LDA Jerárquico): versión de LDA que permite detectar *subtopics* y/o relaciones entre *topics*.

Otros: extensiones y/o generalizaciones de LDA.

Aplicaciones NLP/IR de las técnicas de *topic modelling*

- Agrupación de documentos (*clustering*) por temática.
- Anotación automática de grandes colecciones de documentos.
- Organización, relación y búsqueda de documentos (*problema IR*).
- Resumen de documentos.

Índice de contenidos

- **Introducción**
- **Algoritmos de *clustering***
- **Técnicas de *topic modelling***
- **Técnicas de reducción de dimensionalidad**
- **Conclusiones**

Índice de contenidos

- Introducción
- Algoritmos de *clustering*
- Técnicas de *topic modelling*
- **Técnicas de reducción de dimensionalidad**
- Conclusiones

Índice de contenidos

- **Introducción**
- **Algoritmos de *clustering***
- **Técnicas de *topic modelling***
- **Técnicas de reducción de dimensionalidad**
- **Conclusiones**

Representación de grandes cantidades de info.

Vocabulario de gran tamaño en grandes cantidades de documentos

1. Problemas de eficiencia c.r.a. tiempo y espacio.

2. Exceso de ruido: impacto negativo en los resultados.

Preprocesamiento en NLP/IR

1. **Eliminar *stopwords* y ‘palabras raras’**

Palabras sin valor significante por ser muy comunes o poco frecuentes.

Preprocesamiento en NLP/IR

1. **Eliminar *stopwords* y 'palabras raras'**

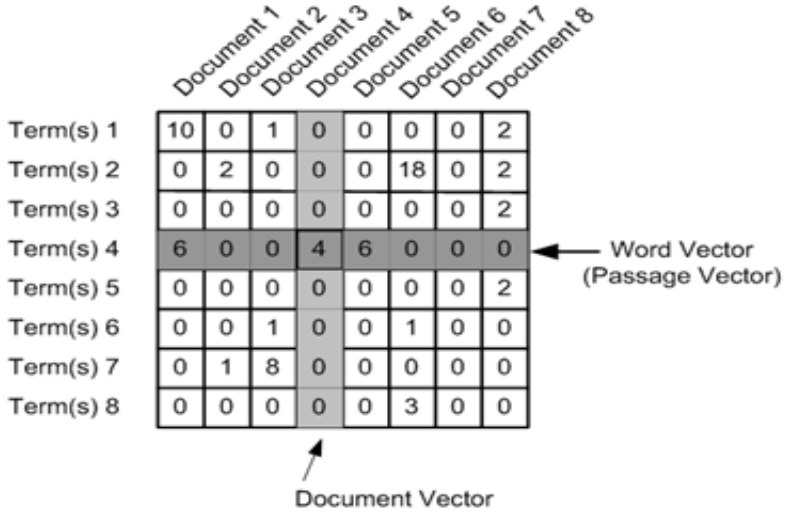
   Palabras sin valor significante por ser muy comunes o poco frecuentes.

2. **Lematizacion o *stemming***

   Representar igual palabras con mismo origen ej. gata, gatos, etc.

   Lema: gato
   Stem: gat-

Factorización de matrices



	I	*like*	*enjoy*	*deep*	*learning*	*NLP*	*flying*	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Métodos

SVD: Descomposición en valores singulares

PCA: Análisis de Componentes Principales

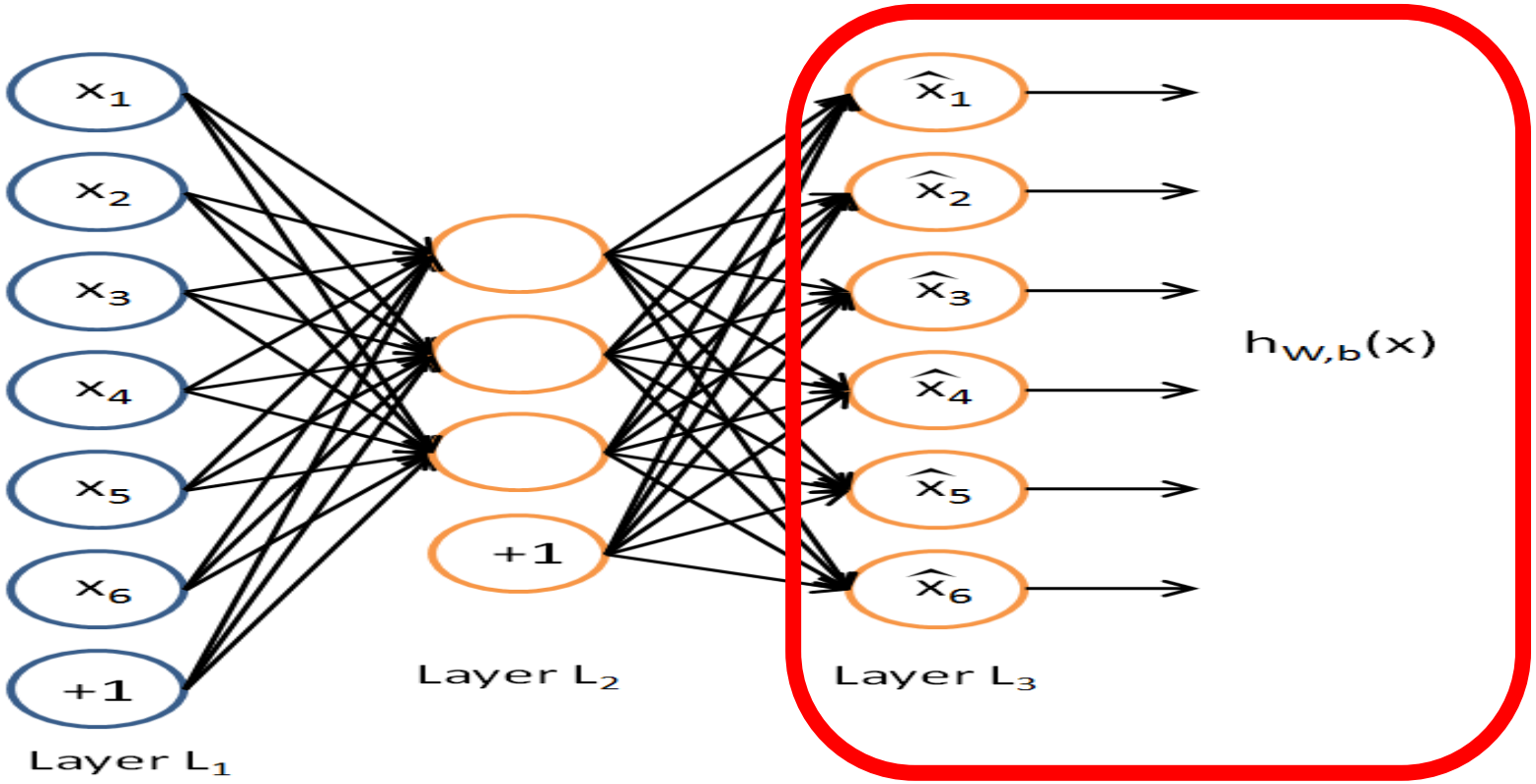
NMF: Factorización no negativa de matrices

Ventajas:

- **Reducción de dimensiones: mayor eficiencia.**
- **Selección de rasgos apropiados.**

Word embeddings

Autoencoders: reducción no lineal



Aprender una aproximación del input comprimida a partir de una capa con un número de neuronas inferior al número de vectores de entrada.

Índice de contenidos

Conclusiones

- Las técnicas de **aprendizaje no supervisado evitan la necesidad y el coste de contar con un conjunto de datos anotados**.

- Los **algoritmos de *clustering* son adecuados cuando no se conoce a priori el número de clases** en los que se quiere dividir un conjunto de datos. No obstante, **el criterio de agrupamiento puede ser subjetivo**.

- Las técnicas de ***topic modelling* permiten extraer las temáticas tratadas en una colección de documentos**.

- Las técnicas de **reducción de dimensionalidad obtienen una representación más óptima de los datos** y pueden emplearse para **identificar aquellos rasgos más representativos**.

¡Gracias!