Introducción al Procesamiento de Lenguaje Natural

Grupo PLN – InCo

Clustering

Clustering

Clustering vs. Clasificación

Clustering

- Es una tarea que tiene como finalidad principal lograr el agrupamiento de conjuntos de objetos no etiquetados
- Esos agrupamientos reciben el nombre de ***clusters***
- Los elementos de cada cluster poseen algunas características similares y que los distinguen de los otros
- Los métodos de clustering intentan encontrar este tipo de clusters de manera automática
- Queda a nuestro criterio darles una interpretación semántica

Clustering

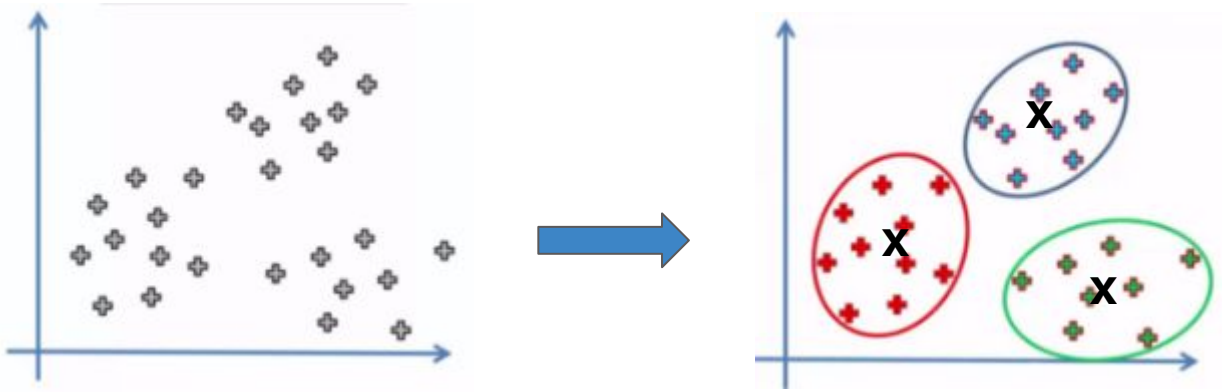
Algunos usos:

- Biología
- Medio Ambiente
- Segmentación de mercado
- Sociología
- Periodismo
- Análisis de redes sociales
- Recuperación de información
- ...

Clustering

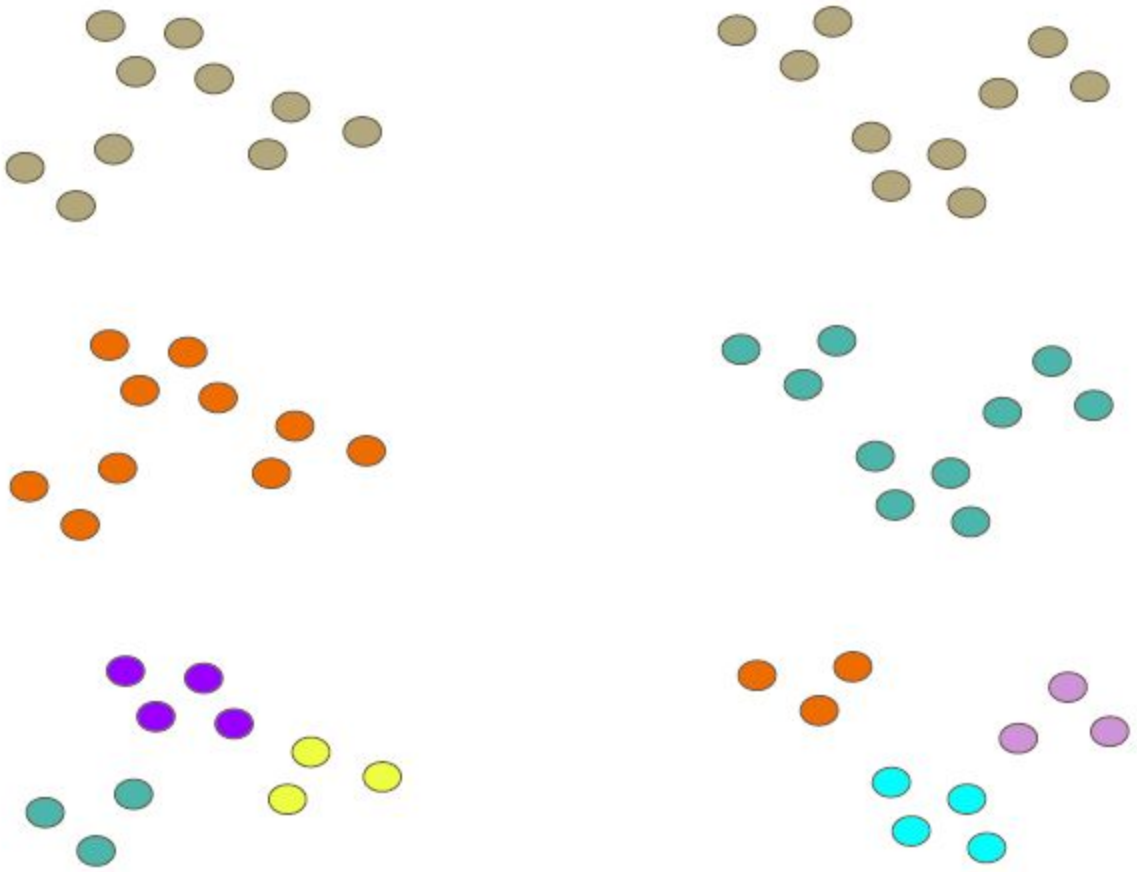
Algoritmos

- k-means



- Cada cluster se representa mediante un punto en el espacio (denominado centroide)
- Tengo *K* de estos puntos
- Los puntos que queden más cerca del centroide c_i que de cualquier otro centroide corresponden al cluster C_i
- Proceso iterativo

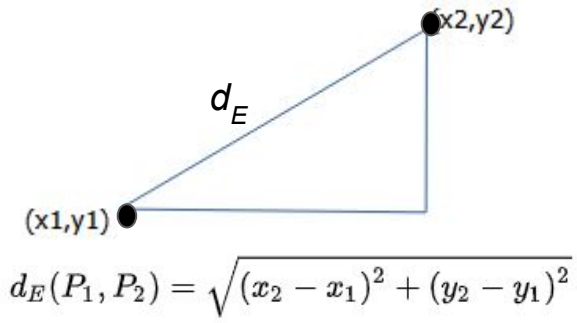
Clustering



Clustering

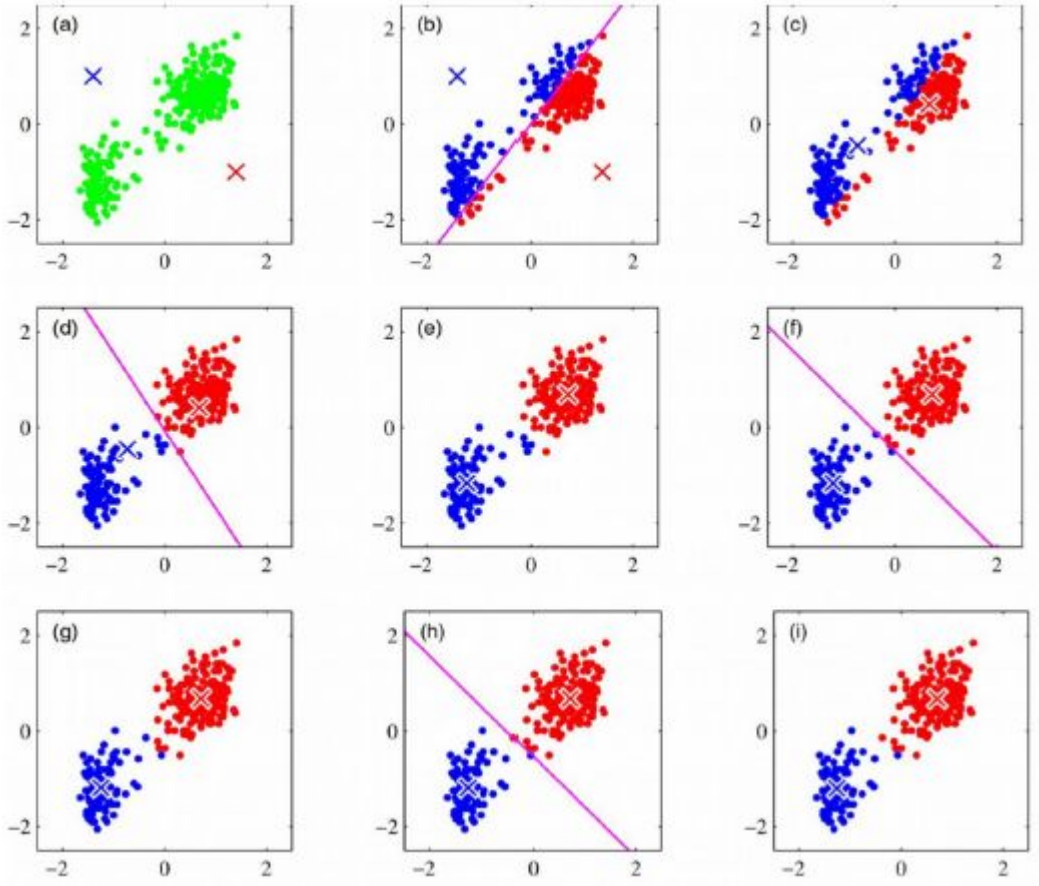
k-means (cont.)

- Para el cálculo de la distancia entre los puntos
  - euclídea
  - coseno

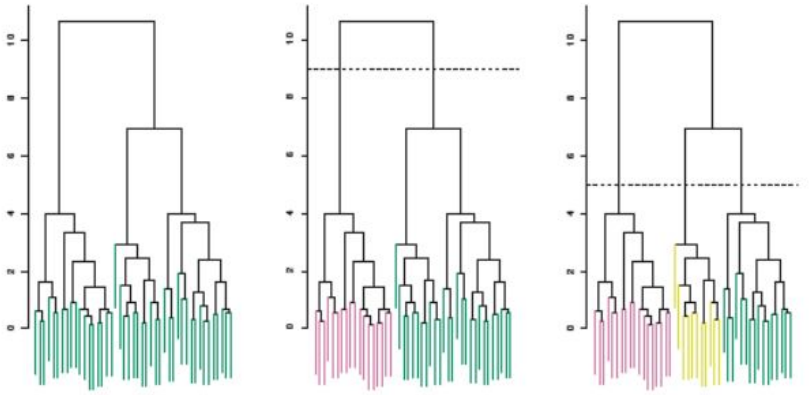


- En general es muy rápido y converge en pocas iteraciones
- No hay solapamiento de objetos de distintos clusters
- Desafío → elegir el "mejor" ***k***

Clustering

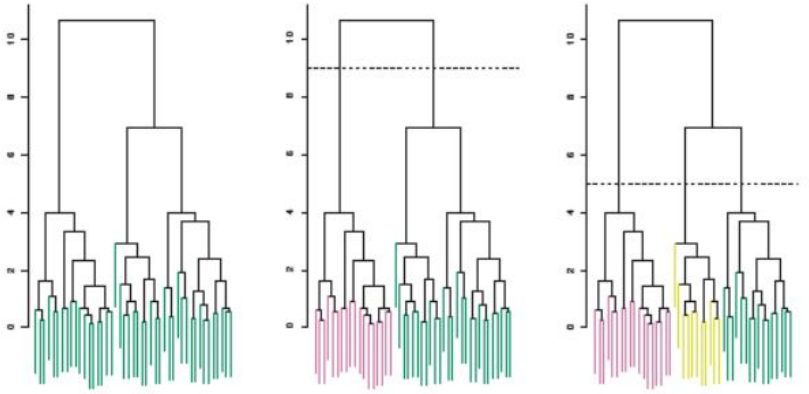


➢ Jerárquico



- No se parte de un número predefinido de clusters
- Representación mediante un dendograma (árbol)
- Enfoque tradicional bottom-up (uniendo por las hojas)

➢ Jerárquico



- Cada hoja representa un elemento u observación
- A medida que se sube, algunas de las hojas se fusionan en ramas
- La clave para interpretar el árbol es centrarse en la altura en la que dos elementos que se unen
- La posición horizontal de cada división da información sobre la distancia entre dos clústeres

Modelado de Tópicos

Tópicos

4. m. *Ling.* Elemento de un enunciado, normalmente aislado entre pausas, que introduce alguno de los elementos de la relación predicativa o bien aporta el marco o el punto de vista pertinente para la enunciación.

5. m. *Ling.* **tema** (parte de un enunciado).

(https://dle.rae.es/topico)

Tópicos

Tópico vs Tema

Que un conjunto de palabras tienden a aparecer juntas no significa que compartan un tema semántico.

el Origen

…las ***colocaciones***

Combinación frecuente/estable de palabras:

cerrar una ventana

cometer un error

…

vs expresiones

meter la pata

tomar el pelo

cortar por lo sano

…

Tópicos

- Es el asunto principal del que se habla, se explica, se predica o se comunica algo

- Dado un documento, no necesariamente es fácil determinarlo

Tópicos

Ejemplo:

"...a partir de este martes cada club sólo podrá sumar nueve puntos, unidades que no sólo definirán el último módulo del Campeonato Uruguayo, sino que también decidirán quiénes se mantienen en Primera…"

"...a partir de este martes cada **estudiante** sólo podrá sumar nueve puntos, unidades que no sólo definirán el último módulo del **curso actual**, sino que también decidirán quiénes se mantienen en **carrera**…"

Modelado de Tópicos

- El modelado de tópicos nos permite organizar, entender y resumir grandes colecciones de texto.
- Intenta detectar patrones de las ocurrencias de las palabras agrupándolas en base a las distribuciones de esas palabras en un conjunto de documentos.
- Es útil para identificar temáticas *implícitas* en un conjunto de documentos y así poder agruparlos

Modelado de Tópicos

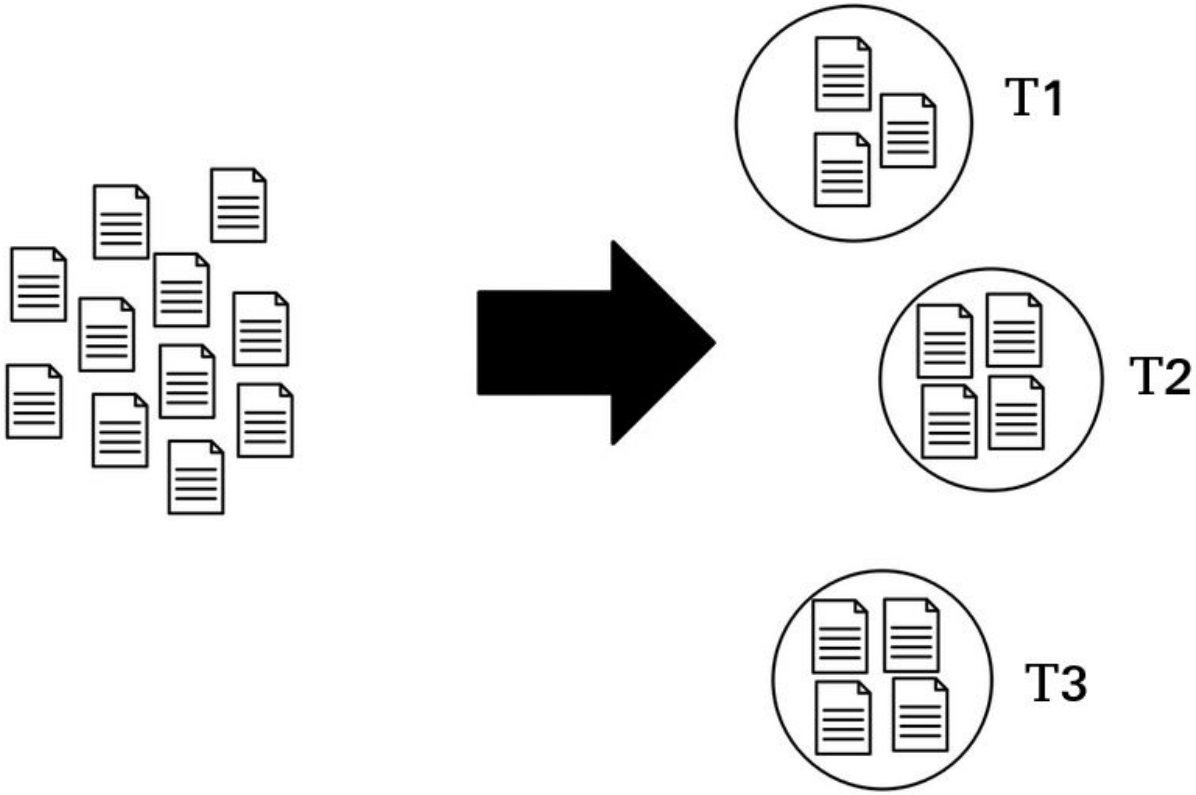
- Consiste en construir un modelo que busque y encuentre palabras que estén relacionadas (de alguna manera) en los textos
- Esas agrupaciones conforman *clusters* y reciben el nombre de **tópicos**
- La estrategia es tratar de que los tópicos sean lo más diferentes entre sí

Modelado de Tópicos

Finalmente, en el modelado de tópicos:

- Cada tópico es una bolsa de palabras
- Cada documento es una mezcla de tópicos
- Cada documento puede tener cierto “porcentaje” de palabras que (con mayor o menor frecuencia) pueden aparecer en más de un tópico

Modelado de Tópicos



Modelado de Tópicos

2 enfoques:

→ lista de palabras clave

→ detectar patrones de las ocurrencias de las palabras agrupándolas en base a las distribuciones de esas palabras en un conjunto de documentos

Modelado de Tópicos

Ejemplo: Lista de palabras clave

Economía	economía, económico, económica, economista, comercio, inflación
Incertidumbre	incertidumbre, incierto, incierta, riesgo país
Política impositiva	impuestos, impuesto, impositivo, gravado, subsidio, subsidios, imesi, iva, irpf

Latent Dirichlet Allocation (LDA)

Tópico: una distribución sobre un vocabulario fijo

(Blei, 2003)

- Este modelo genera *tópicos* proponiendo una cierta distribución de todas las palabras del corpus, y calcula la distribución de estos tópicos en cada documento.

- Cada documento en el corpus es atribuible con cierta probabilidad a cada uno de esos tópicos.

- Cada palabra en el corpus, también es atribuible con cierta probabilidad a pertenecer a un tópico.

LDA

- Cada tópico es una distribución probabilística de palabras

Turismo	Educación	Economía
Ministerio 13%	Ministerio 8%	Economista 18%
Argentinos 17%	Aula 29%	Dólar 22%
Bilateral 9%	Estudiante 31%	Pesos 15%
Blue 21%	Libro 15%	Blue 5%
…	Escuela 12%	
	…	

LDA

En base a cifras del Ministerio de Turismo, el Observatorio liderado por el economista Javier de Haedo señaló que en el primer trimestre de este año "el gasto de los argentinos en Uruguay alcanzó a US$ 431 millones (83% más que un año antes), mientras que el de los uruguayos en Argentina ascendió a US$ 291 millones (680% más que en igual trimestre de 2022)".

En esta línea, se señala que si se consideran los últimos 12 meses los argentinos gastaron US$ 922 millones en Uruguay, mientras que los uruguayos gastaron por US$ 913 millones en el país vecino, donde este martes el dólar "blue" trepó a casi 500 pesos argentinos tras una nueva corrida bancaria.

- Cada documento es una distribución probabilística de tópicos

  Turismo 25% Educación 7% Economía 19%

  (por ejemplo)

LDA

- Se asigna inicialmente una probabilidad $p_{d,t}$ de que el documento *d* pertenezca al tópico *t* siguiendo la distribución de Dirichlet

- LDA permite que un documento sea parte de varios tópicos, cada uno con un peso diferente

- LDA requiere que se le defina de antemano la cantidad de tópicos a buscar (k)

- Métricas:
  - coherencia
  - perplejidad

Modelado de Tópicos

- CTM (Correlated Topic Model)

Variante de LDA, donde se reemplaza la distribución de Dirichlet por una distribución normal logística

- BTM (Biterm Topic Model)

Incluye el concepto de *bitérmino*, que es un par de palabras no necesariamente contiguas que co-ocurren en un contexto corto

Modelado de Tópicos

- ETM (Embedded Topic Modeling)
  - Enriquece LDA con el uso de *word embeddings* (Dieng, 2020)
  - Representación vectorial de ***d*** dimensiones de las palabras de un vocabulario ***V***
  - Utiliza un vector del mismo espacio para representar cada tópico
  - La probabilidad de una palabra en un tópico es proporcional al producto escalar entre sus vectores asociados (el vector del tópico y el de la palabra)

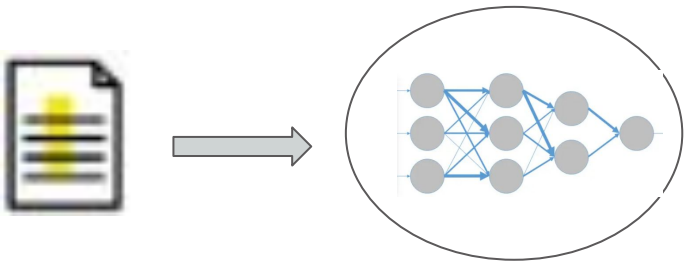
ETM

Hiperparámetros:

- **n**: número de tópicos a inferir
- **d**: dimensión del espacio de embeddings
- **V** : cantidad de palabras del vocabulario.

Además:

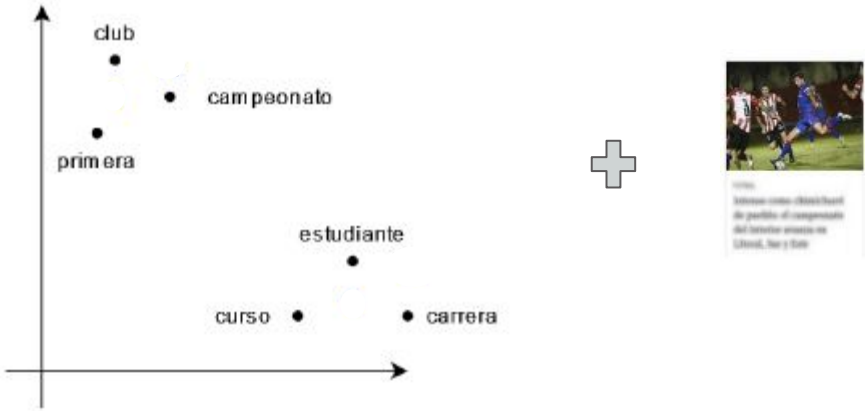
- Una matriz de embeddings con dimensión **d** x **V**
- Una matriz de tópicos
- Una red neuronal de predicción de tópicos, con entrada de tamaño **V** y salida de tamaño **n**

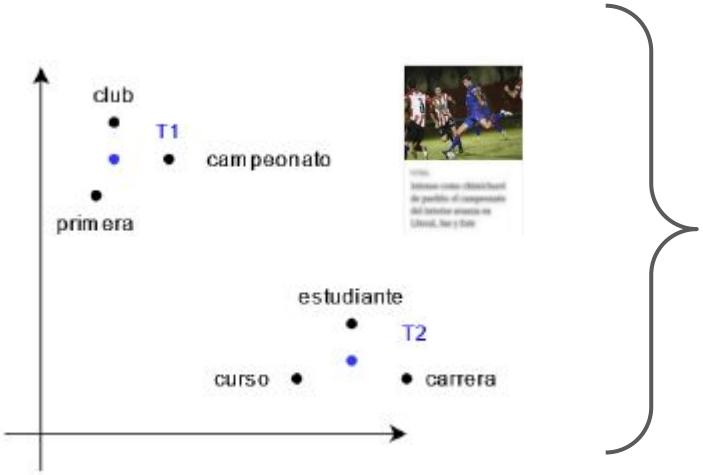


- tópicos inferidos con su % de probabilidad
- distribución de las palabras

ETM

Ejemplo





T1: 90%
T2: 10%

club: 32%
campeonato: 26%
primera: 29%
carrera: 2%
estudiante: 1%
curso: 1%
...