

■

TFG

Análisis de Clustering y Clusterización de Documentos

□

Autor: Aarón Casado Monge

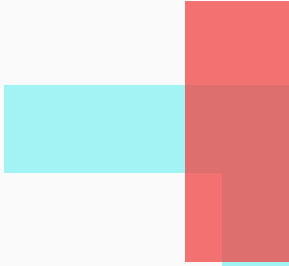
Tutor: Juan José Cuadrado Gallego

Grado en Ingeniería Informática con Mención en Ciencias de la Computación

Introducción: Era de los datos

Información sobre la situación actual y contexto

- Se generan más de 25.000 petabytes de datos diarios (2,5 trillones de bytes).
- Esta explosión de datos se ha denominado **Big Data**.
- Con consecuencias positivas como la accesibilidad y disponibilidad de la información.
- Pero también consecuencias negativas y problemas:
 - Dificultad para el almacenamiento y organización de los datos.
 - Desinformación.
 - Complicación para mostrar resultados.
 - Datos no estructurados que entorpecen su procesamiento.



“Ciencia que usa Estadística, Inteligencia Artificial, Programación y Bases de Datos para posibilitar la extracción de conocimiento a partir de datos”



Data Science (Ciencia de los Datos)





Ramas de la Ciencia de los Datos

Data Warehousing

Organización y agrupación de datos

Data Mining

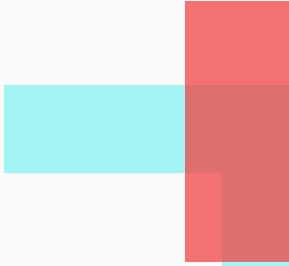
Análisis de datos

Visualization



Presentación de los resultados





“Proceso de descubrimiento de patrones interesantes y extracción de conocimiento a partir de grandes volúmenes de datos”



Data Mining (Minería de Datos)



Clasificación

- Actividad primitiva del ser humano con la que obtener información de los datos.
- Permite entender o asimilar nuevos objetos o fundamentos comparándolos con otros objetos o fenómenos en función de la similitud o disparidad que exista entre sus principales características.
- Dificultad para clasificar datos no estructurados que no dispongan de etiqueta o valor conocido que sirva como criterio de clasificación.
- ¿Solución?

CLUSTERING



01.

Definición

Concepto teórico y utilidades.

02.

Técnicas

Diferentes métodos y algoritmos existentes.

03.

Aplicaciones

Usos reales y énfasis en documentos.

04.

Clusterización

Ejemplo práctico enfocado sobre COVID-19.





Clustering

Concepto teórico y utilidades.



01.



Origen y definición

- Clustering, Cluster Analysis, Agrupamiento, Análisis de grupos o Clusterización es un método de Data Mining basado en Aprendizaje Automático, una rama de Inteligencia Artificial.
- El Aprendizaje Automático pretende resolver problemas basándose en resultados previos.
- Pero existe un método de aprendizaje que no dispone de experiencias previas sobre las que aprender

Aprendizaje No Supervisado

- Su objetivo principal es discernir patrones y relaciones entre los datos para poder agruparlos y trabajar sobre los resultados de la clasificación resultante.
- Clustering es un método de Aprendizaje No Supervisado.



■

“Proceso de organizar datos en grupos diferentes en base a la similitud o disparidad entre los mismos, definiendo en el propio proceso de clasificación los valores que delimitan cada grupo”

□


Clusterización





Utilidades

Clasificación de datos sin etiqueta o valor que permita clasificarlos.



- Detección de datos anómalos.
- Obtención de etiquetas y criterios de clasificación.
 - Pre-procesamiento de datos para Data Mining.
 - Compresión de información.

¿Resultado?

Obtención de nuevo conocimiento.



Campos en los que se usa

Ingeniería

Aprendizaje automático, IA, reconocimiento de imágenes...

Ciencias sociales

Sociología, psicología, educación...

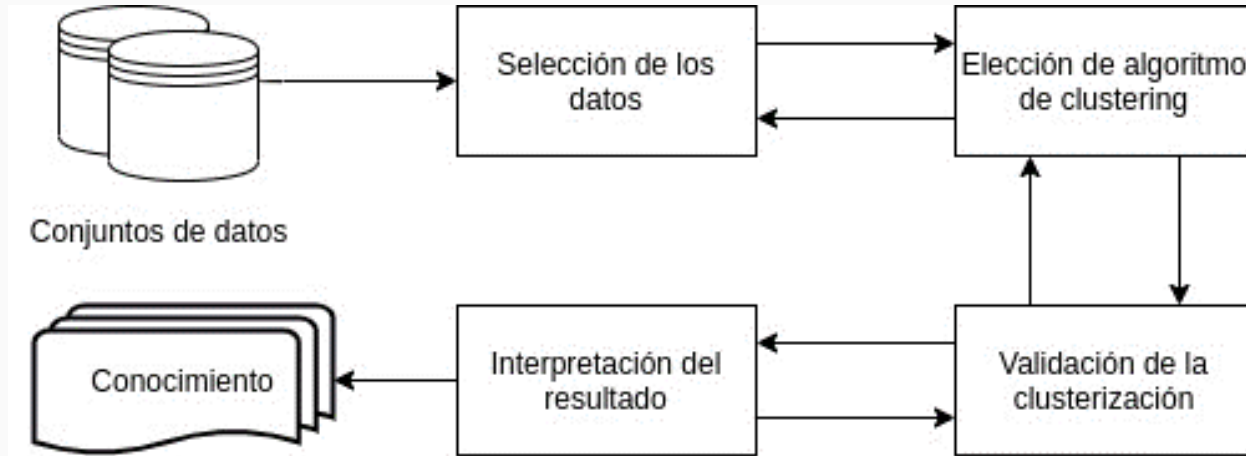
Ciencias de la salud

Genética, biología, microbiología, psiquiatría...

Economía

Marketing, negocios, detección de fraude...

Procedimiento de clustering





Técnicas de clustering

Métodos y algoritmos.



02.



Requisitos

- Escalabilidad.
- Habilidad para lidiar con diferentes tipos de datos.
- Descubrimiento de clusters con formas arbitrarias.
- No requerir información al usuario.
- Capacidad para trabajar con ruido en los datos.
- Clusterización incremental e insensibilidad al orden de entrada.
- Capacidad para clusterizar datos con múltiples atributos.
- Clustering basado en restricciones.
- Interpretabilidad y usabilidad.

■ Criterios de comparación

Criterio de división

Métodos de partición o jerárquicos.

Cálculo de similitud

Uso de distancia, densidad, contigüidad.

Separación de clusters

Métodos deterministas o “fuzzy”.

Espacio de clusterización

Espacio total o subconjuntos del espacio.

Principales métodos de clusterización

Métodos	Características generales	
Basados en particiones	<ul style="list-style-type: none">• Encuentran clusters de forma esférica mutuamente exclusivos• Utilizan distancia/proximidad• Usan la media o medoid para representar el centro del cluster• Efectivos en conjuntos de datos pequeños y medianos	<ul style="list-style-type: none">• Basados en modelos probabilísticos• Basados en la teoría de grafos• Basados en redes neuronales• Clusterización relajada
Jérárquicos	<ul style="list-style-type: none">• Clasifican en múltiples niveles• No pueden deshacer agrupaciones o divisiones erróneas• Pueden incorporar otras técnicas como microclustering y tener en cuenta vínculos entre los objetos	
Basados en densidad	<ul style="list-style-type: none">• Pueden encontrar clusters con formas arbitrarias• Los clusters son regiones con gran densidad de objetos separados por zonas con poca densidad• La densidad queda definida por un mínimo de objetos cercanos dentro del vecindario• Sirve para detectar datos anómalos	
Basados en rejillas	<ul style="list-style-type: none">• Utiliza una estructura de rejilla o cuadrícula• La velocidad de procesamiento es alta, pues no influye el número de objetos	

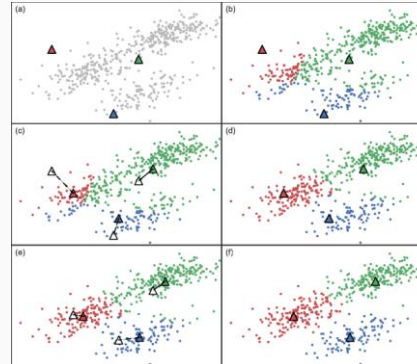
01. Basados en particiones

- Simples.
- Solicitan el número de clusters a formar.
- Optimizan el criterio de similitud usando el cálculo de la proximidad.
- Principalmente usan Distancia Euclídea.
- Utilizan el centroide para representar al cluster.
- Calculan la varianza como forma de comprobar la calidad de cada cluster.

Algoritmos más utilizados:

- **K-means.**
- **Fuzzy k-means.**
- **K-medoids.**
 - **PAM** (Partition Around Medoids).
 - **CLARA** (Clustering LARge Applications).

“Dado un conjunto de datos D con n objetos y siendo k el número de clusters a formar, un algoritmo basado en particiones organizará los objetos en k divisiones siendo $k \leq n$, donde cada una de ellas representa un cluster”



02. Jerárquicos

“Los métodos jerárquicos clasifican los datos en diferentes niveles utilizando un enfoque jerárquico”.

Aglomerativos

Estrategia de “abajo a arriba”,
une clusters individuales hasta
llegar a uno solo llamado raíz.

Divisivos

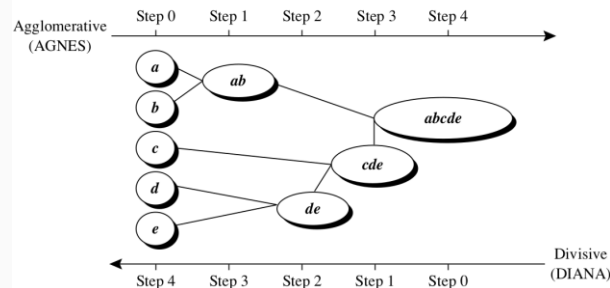
Estrategia de “arriba a abajo”,
parte de un solo cluster, la raíz, y
los va dividiendo hasta que cada
cluster es un único objeto.

02. Jerárquicos

- Dificultad a la hora de dividir o juntar clusters.
- Difícilmente escalable hacia grandes volúmenes de datos.
- Usan diferentes medidas de distancia.

Algoritmos más utilizados:

- **AGNES** (AGlomerative NESTing).
- **DIANA** (DIvisive ANALysis).
- **BIRCH** (Balanced Iterative Reducing and Clustering using Hierarchies).
- **Chameleon**.
- Métodos jerárquicos probabilísticos.



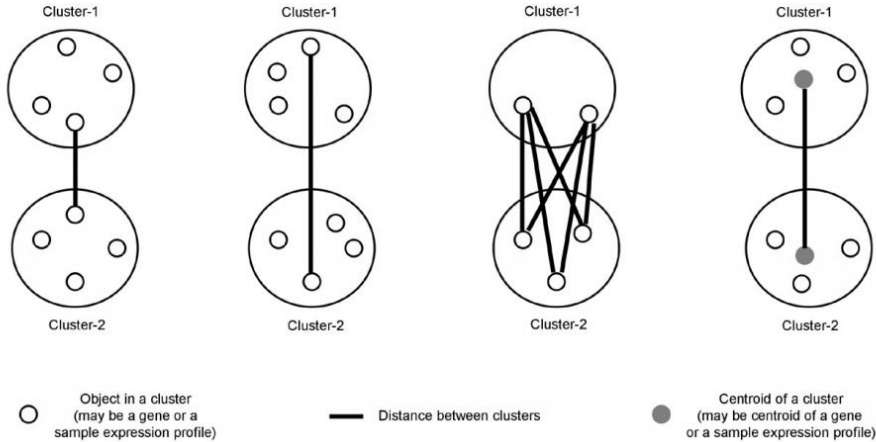
02. Medidas de distancia

Distancia mínima

Distancia máxima

Distancia media

Single linkage clustering Complete linkage clustering Average linkage clustering Centroid linkage clustering



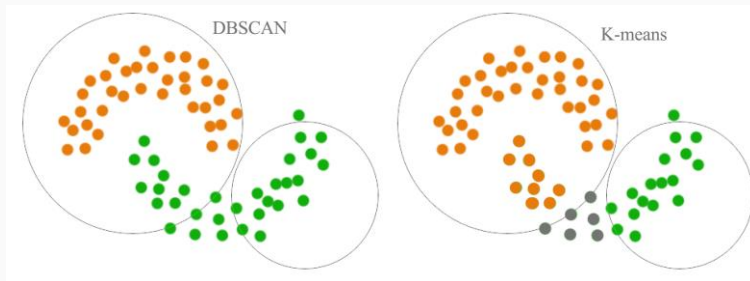
03. Basados en densidad

- Destacan encontrando clusters con formas arbitrarias.
- Lidian bien con ruido en los datos.

Algoritmos más utilizados:

- **DBSCAN** (Density-Based Spatial Clustering of Application with Noise).
- **OPTICS** (Ordering Points to Identify the Clustering Structure).
- **DENCLUE** (DENSity based CLUstEring).

“Los clusters se consideran regiones densas de objetos separadas por otras regiones de baja densidad”.



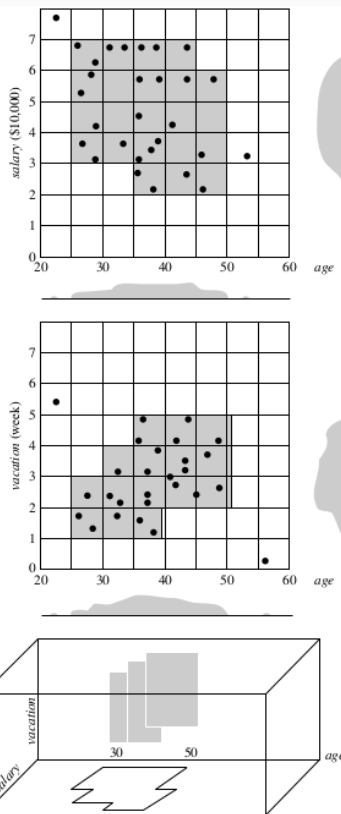
04. Basados en rejilla

- No dividen los datos, sino el propio espacio.
- Gran escalabilidad.
- Aumenta el rendimiento de procesamiento.

“Separan el espacio de datos en celdas equitativas independientemente de la distribución de los puntos”.

Algoritmos más utilizados:

- **STING** (Statistical Information Grid).
- **CLIQUE** (Clustering In QUEst).





Aplicaciones

Usos reales, validación y énfasis en
documentos



03.



Tipos de datos

	Particiones	Jerárquicos	Densidad	Rejilla	Fuzzy	Probabilísticos	Grafos	Redes neuronales
Categoricos								
Texto								
Multimedia								
Streaming								
Temporales								
Discretos								
Biológicos								
Red								
Difusos								

Evaluación del clustering

¿cómo sabemos que el resultado de una clusterización es preciso y correcto?

01

Evaluar la tendencia de la clasificación comprobando que existe una estructura no aleatoria en la distribución de los datos.

02

Determinar el número de clusters óptimo dado el algoritmo y la muestra de datos.

Calidad de la clusterización

Existen métodos que permiten evaluar el resultado del proceso de clustering.

Intrínsecos

Comprueban si los clusters acogen bien los datos examinando cuán compactos son los clusters y lo bien que están separados

Extrínsecos

Comprueban si los clusters se aproximan a una realidad existente utilizando información externa.

■ Aplicaciones reales

- Paso intermedio para otros métodos de data mining.
- Método de filtrado colaborativo.
- Segmentación de clientes.
- Resumen de datos.
- Detección de tendencias dinámicas.
- Análisis de datos multimedia.
- Análisis de redes sociales.
- Análisis de datos biológicos.


Ejemplos

- Clasificar tipos de consumidores con preferencias similares.
- Aunar bajo un mismo subconjunto muchas formas diferentes de escribir el mismo carácter.
- Agrupar resultados similares de una consulta en internet.
- Detección de fraudes mediante la identificación de datos anómalos.

Clusterización de documentos

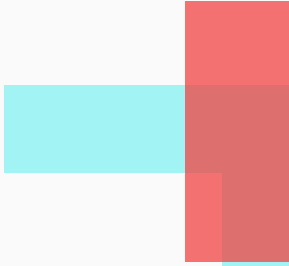


Clusterización de textos

- El 80% de la información se encuentra en textos.
 - Noticias, revistas, blogs, páginas web, etc.
 - Humanamente imposible procesar los datos
- 

¿Solución?

TEXT MINING



“Disciplina enfocada en la utilización de ordenadores para el descubrimiento de nueva información previamente desconocida mediante la automatización de la extracción de información”.



Text mining (Minería de textos)



Procesamiento del Lenguaje Natural

- Los documentos están formados por frases y palabras, una fuente de datos no estructurada y no numérica.
- Surge el **NLP** (Procesamiento del Lenguaje Natural) que permite:
 - Extraer información de documentos.
 - Controlar e identificar la temática de textos.
 - Crear resúmenes a partir del contenido.
 - Clasificar documentos.
 - Vincular conceptos similares.
 - Visualizar la información contenida en documentos.
 - Responder preguntas comunes.
 - Buscar y recuperar información.
 - Representar el conocimiento.
 - Analizar el sentimiento de un texto.



Preprocesamiento de los datos

Antes de aplicar algoritmos de clusterización a los documentos, estos se deben transformar en datos numéricos y manejables.



Eliminación de palabras vacías



Normalización

Stemming o lematización.



Tokenización



Reducción de la dimensionalidad

La cantidad de palabras relevantes en un texto debe ser minimizada buscando una representación de menor dimensionalidad que preserve toda la información posible sobre el documento original.

Selección de características

Escoge un subconjunto de los atributos originales de acuerdo a cierto criterio y seleccionando las palabras más relevantes.

Transformación de características

Proyecta el conjunto de datos original en un espacio dimensional más pequeño generando representaciones funcionales de los atributos originales.



Reducción de la dimensionalidad

La cantidad de palabras relevantes en un texto debe ser minimizada buscando una representación de menor dimensionalidad que preserve toda la información posible sobre el documento original.

Selección de características

- Frecuencia de Documento (TF)
- Frecuencia de Término – Frecuencia Inversa de Documento (TF-IDF)
- Fuerza de Término (FS)
- Ranking basado en Entropía (EN)
- Contribución del Término (TC)

Transformación de características

- Indexación Semántica Latente (LSI)
- Proyección Aleatoria (RP)
- Análisis de Componentes Independientes (ICA)

Algoritmos para clusterización de textos

1

K-means

7

SOM
(Self-Organizing Map)

2

PAM

8

EM
(Expectation-
maximization)

3

CLARA

4

DBSCAN

5

DENCLUE

6


Cualquier algoritmo
jerárquico





COVID-19

Ejemplo práctico de clusterización de documentos.



04.



Objetivos de la parte práctica

- Clasificar los más de 200.000 artículos sobre COVID-19 disponibles en Kaggle.
- Identificar las etiquetas que dividen cada grupo.
- Facilitar la búsqueda de información dentro del conjunto de documentos.
- Ver un ejemplo real de Procesamiento Natural del Lenguaje.

Material a utilizar

- Material facilitado por Kaggle.
- Lenguaje de programación R.
- Entorno de programación Rstudio.
- Entorno web JupyterNotebook.

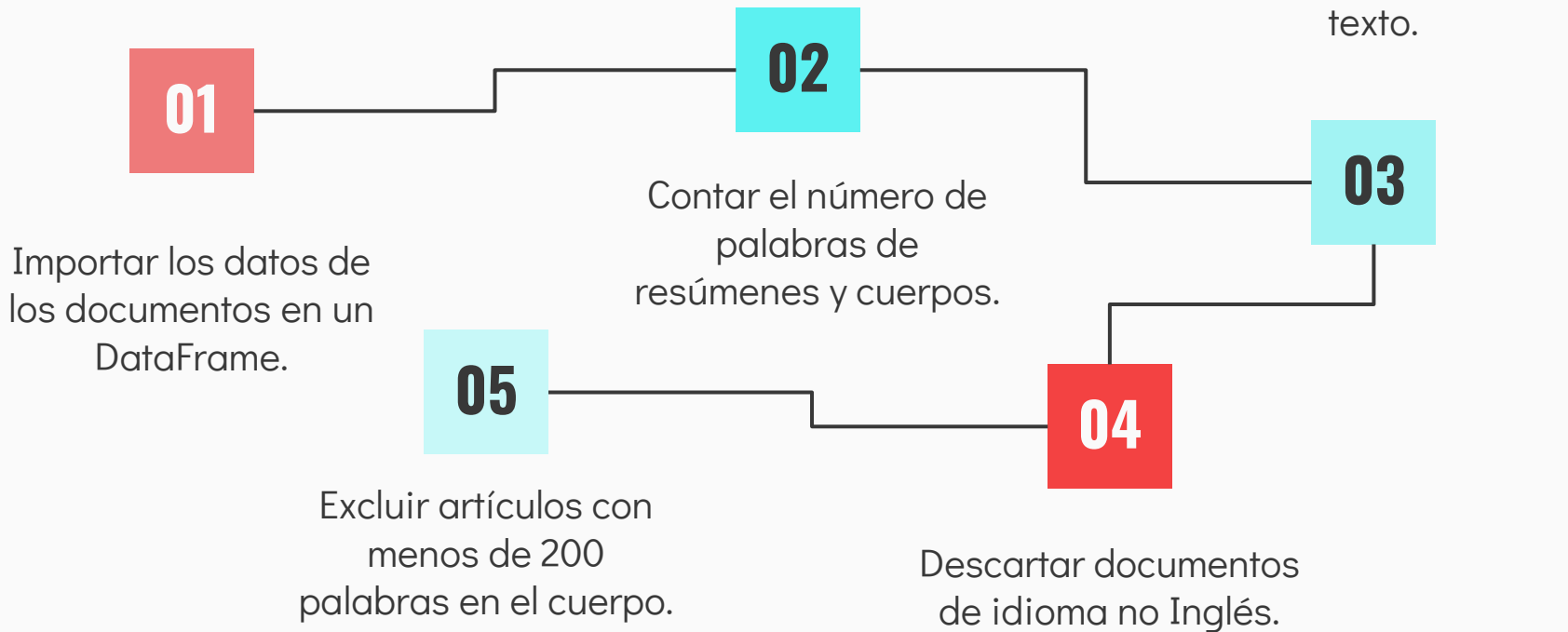
Pasos a seguir:

1. Carga y exploración de los datos
2. Preparación de los datos con NLP
3. Reducción de dimensionalidad
4. Aplicación de algoritmos de clusterización
5. Validación del resultado

Carga y exploración de los datos

- La base de datos cuenta con un archivo “metadata” con información básica de los documentos.
- 253.454 artículos con más de 19 atributos (título, DOI, resumen, autores, fecha de publicación, etc)
- La información individual de cada documento y el cuerpo de texto se encuentra en archivos JSON.
- El número final de artículos individuales incluidos en el directorio desciende a 106.137
- Extraemos la información tanto de los JSON como del archivo “metadata” para importar los datos a un DataFrame para obtener los siguientes atributos: **título**, **autores**, **revista de publicación**, **resumen** y **cuerpo de texto** además del **DOI** que sirve como identificador único.
- 10.157 artículos no cumplen con los requisitos de contenido mínimo y la cifra final de documentos que hemos cargado desciende a 95.980.

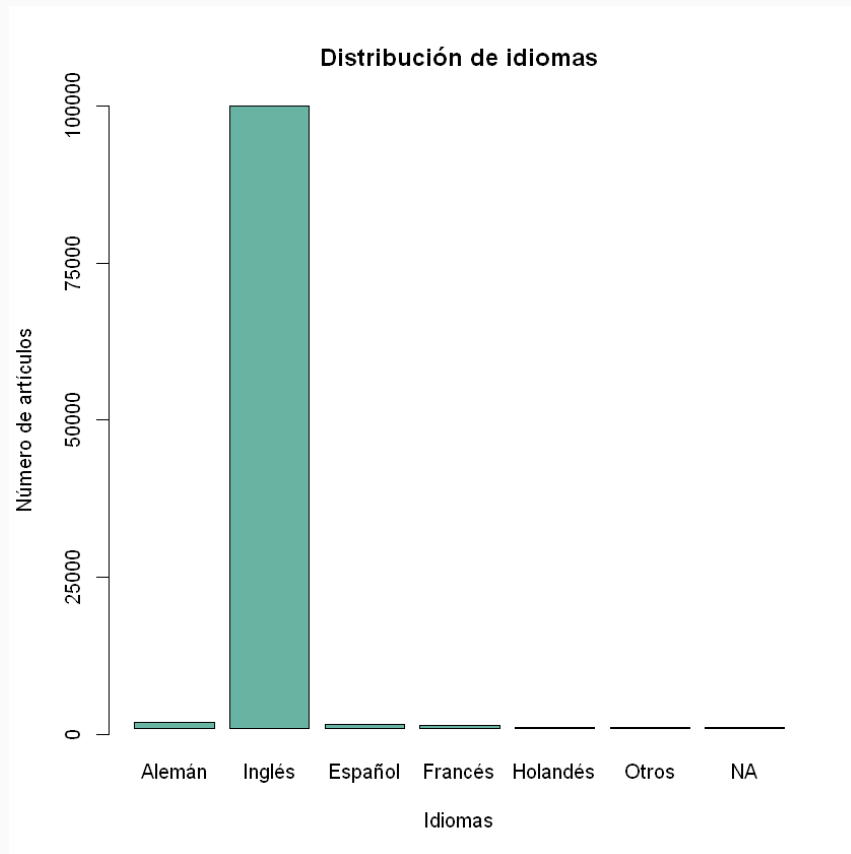
Carga y exploración de los datos



Resultado de EDA

253.454
documentos iniciales.

92.213
documentos finales.

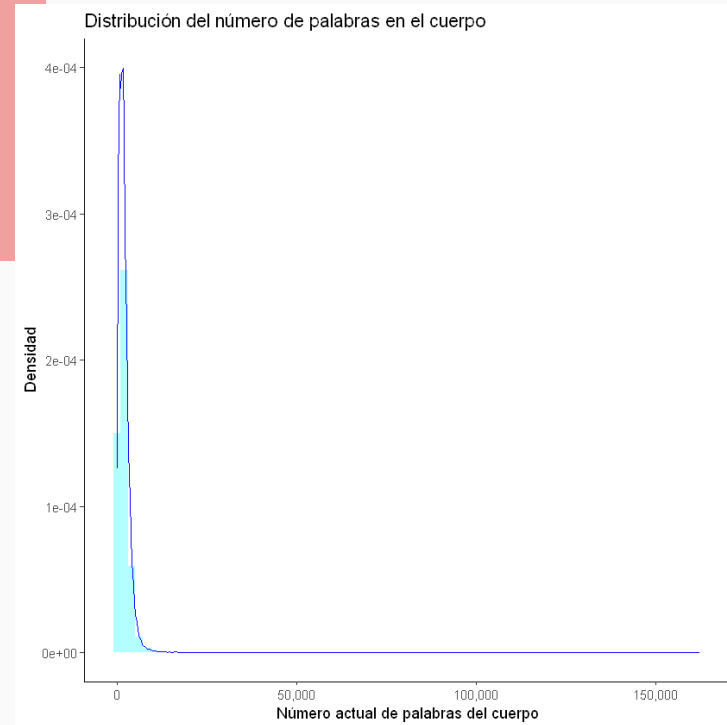
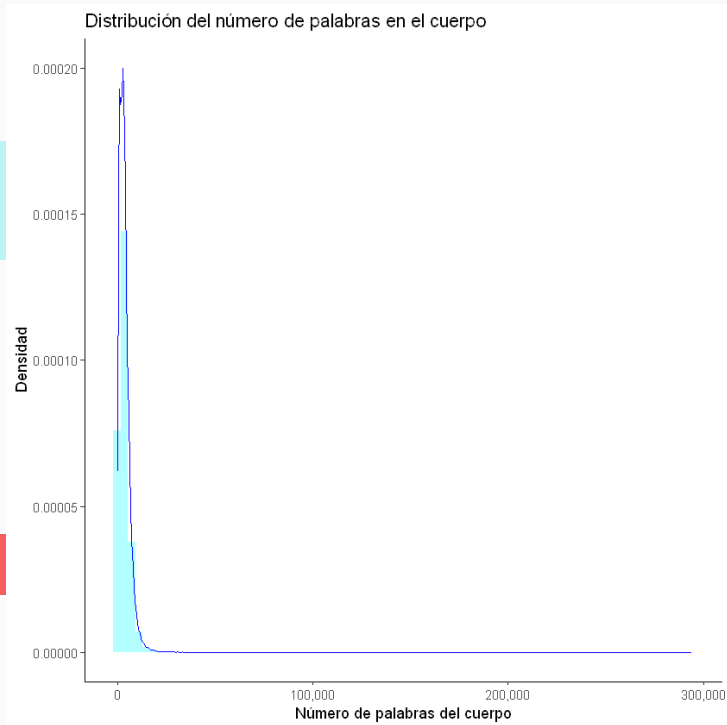


Eliminación de palabras vacías

- Utilizamos el paquete **TAU** y **STOPWORDS**, que contiene una variedad de palabras vacías.
- Añadimos palabras que son comunes en este tipo de documentos al diccionario de palabras vacías.
- Transformamos todas las palabras a minúscula.
- También eliminamos los signos de puntuación.
- Volvemos a contar las palabras finales.

	Inicial	Final	Eliminadas
Resumen	16134914	8828313	7306601
Cuerpo	371628162	192247004	179381158
Total	387763076	201075317	186687759

Eliminación de palabras vacías



Tokenización

- Se ha usado el paquete TEXT2VEC.

```
# Tokenizamos tanto abstracto como texto completo
covid_df['abstract_text'] <-
  ↪ apply(covid_df['abstract_text'], 2, function(s)
  ↪ word_tokenizer(s, xptr = TRUE, pos_keep = character('-')))
covid_df['body_text'] <- apply(covid_df['body_text'], 2,
  ↪ function(s) word_tokenizer (s, xptr = TRUE))
```

1. 'feline' 2. 'infectious' 3. 'peritonitis' 4. 'virus' 5. 'fipv' 6. 'positive' 7. 'cells'
8. 'pyogranulomas' 9. 'exudates' 10. 'cats'

Lematización

- Se utiliza el paquete **TEXTSTEM**.
- Aprovechamos la posibilidad de paralelizar esta operación aprovechando las funciones de R: *lapply*.

	Artículo 1	Artículo 2	Artículo 3	Artículo 4	Artículo 5
Iniciales	2120	2024	977	3331	2968
Finales	1089	1135	474	1838	1700
Unicas	417	496	260	486	631

Reducción de dimensionalidad

01.

Seleccionamos las 256 palabras más importantes.

03.

Transformamos DTM a TF-IDF.

02.

Generamos la matriz DTM.

04.

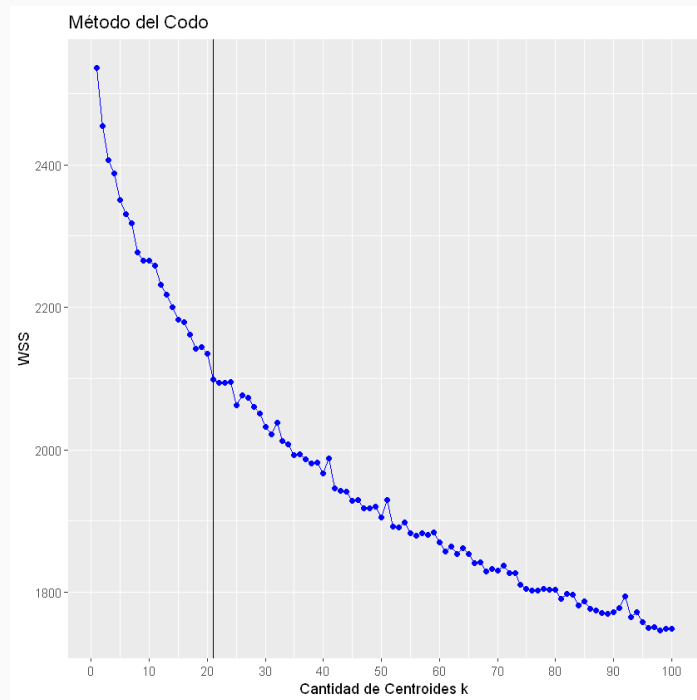
Convertimos el resultado a matriz no dispersa.

Clusterización con K-means

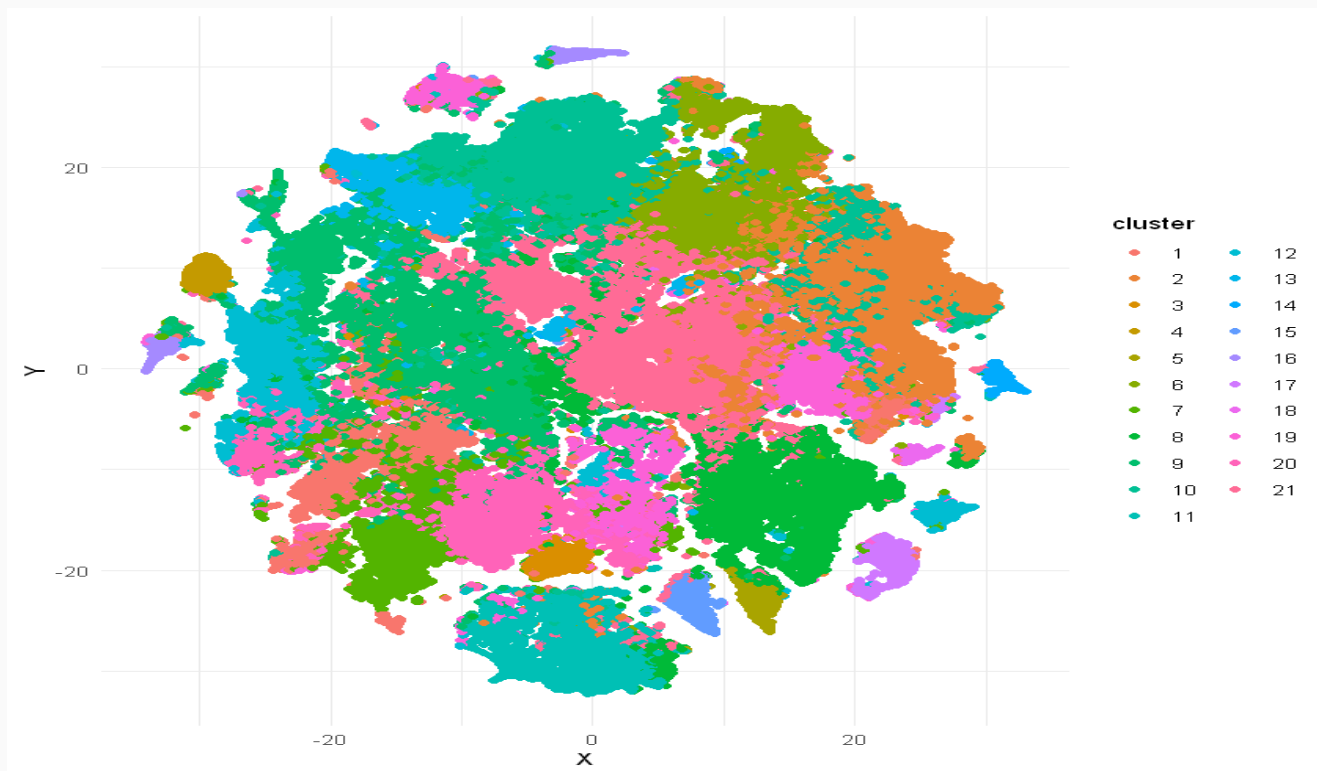
Aplicamos el Elbow Method.

Seleccionamos 21 clusters como punto óptimo.

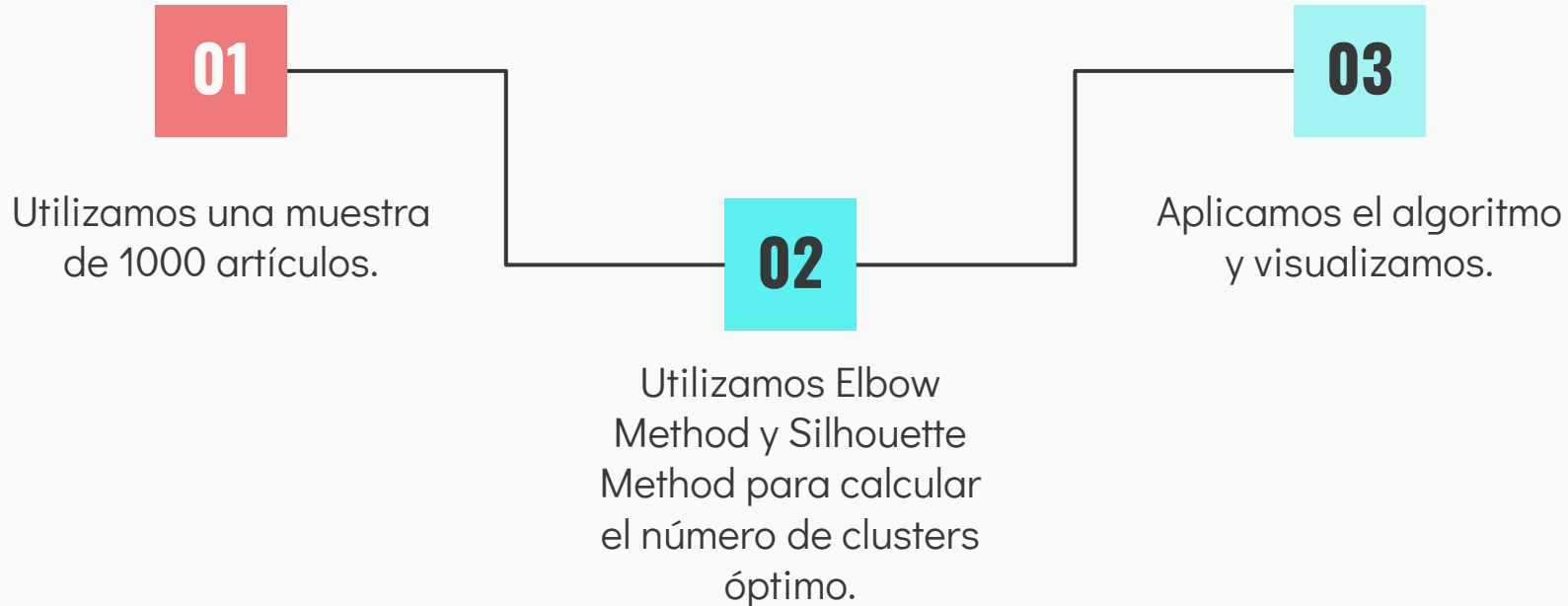
Aplicamos el algoritmo.



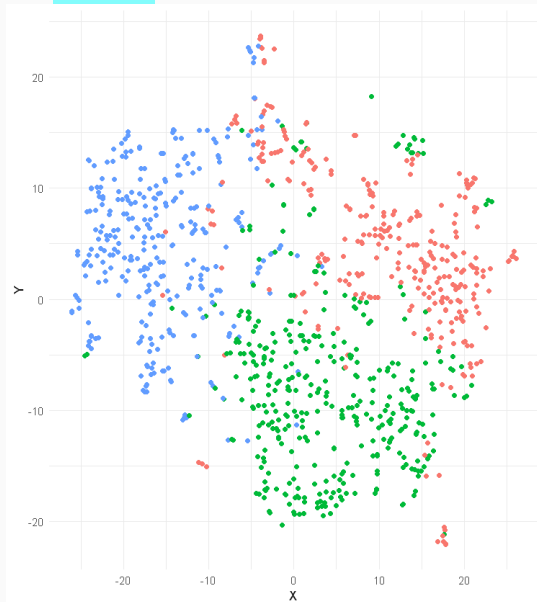
Visualización



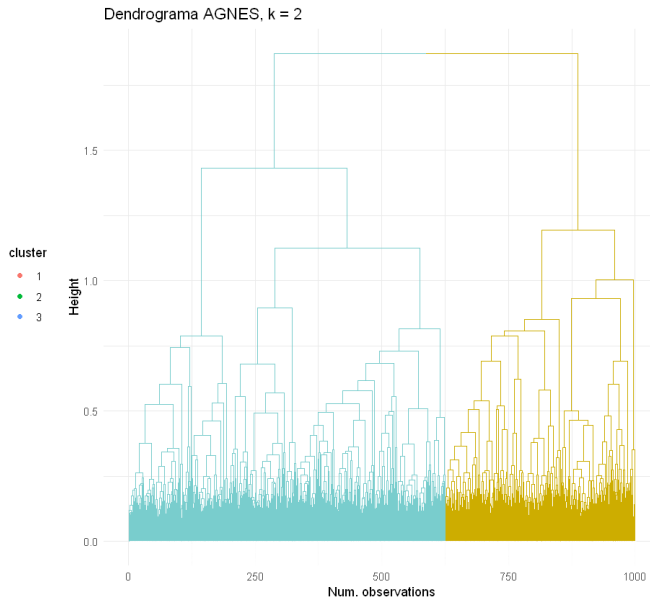
Clusterización con otros métodos



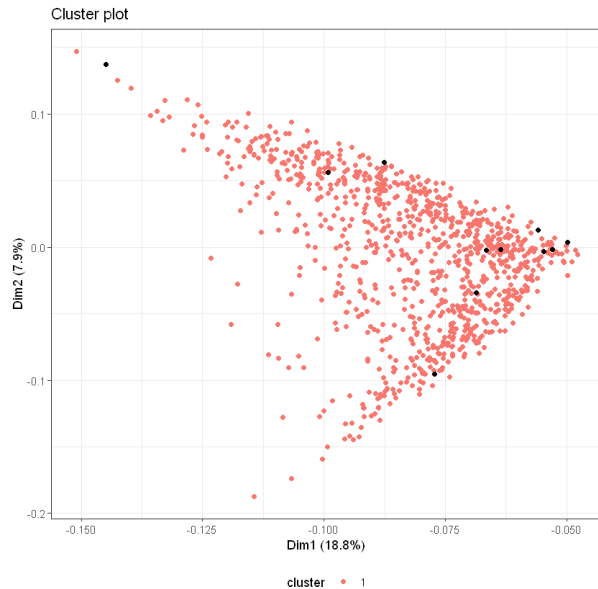
Otros métodos



PAM



AGNES



DBSCAN



- 





Fin

¿Preguntas?