Text Mining


Universidad
de Valparaíso
CHILE


ESCUELA DE INGENIERÍA
INFORMÁTICA
UNIVERSIDAD DE VALPARAÍSO

Unidad 1 - Fundamentos de Text Mining

➔ Contenidos

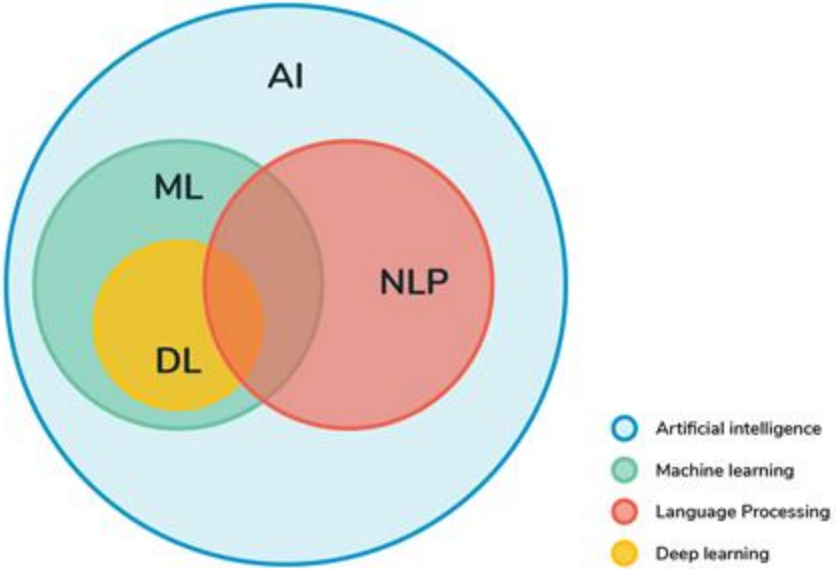
- **Conceptos básicos**
- Leyes del texto
- Procesamiento de texto
- Problemas en Text Mining

¿Qué entendemos por Text Mining?

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

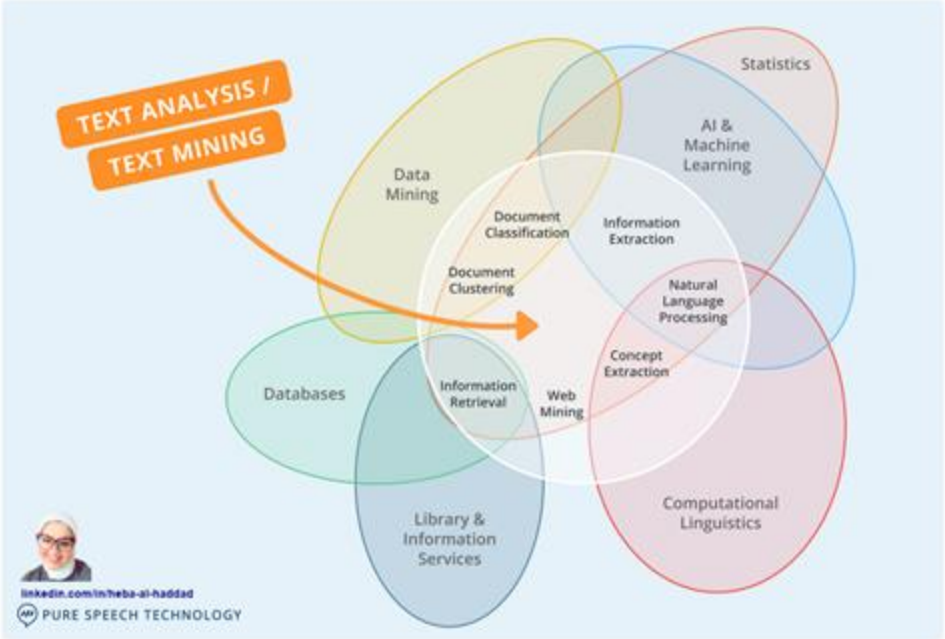
https://www.linkedin.com/pulse/10-common-nlp-terms-explained-text-mining-mohamadreza-mohtat

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos



https://www.sentisum.com/library/nlp-and-text-mining

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos



Fuente: Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.
https://www.linkedin.com/pulse/text-mining-natural-language-processing-business-heba-al-haddad

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Definición

“Extraer información útil desde las fuentes de datos, mediante la identificación y exploración de patrones interesantes”

Proceso conocimiento-intensivo

Colección de documentos

Herramientas de análisis

Tiempo

[Libro: *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*]
[imagen: https://www.flickr.com/]

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Colección de documentos

Documentos

Estáticos

Dinámicos

[Imagen: Designed by macrovector / Freepik"

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Colección de documentos

Estáticos

Dinámicos

Documentos

No estructurado

Semiestructurado

Estructurado

[Imagen: Designed by macrovector / Freepik"

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

No estructurado

Semiestructurado

Estructurado

- ➔ Perspectiva lingüística: estructura semántica y sintáctica (implícita)
- ➔ Elementos tipográficos: signos de puntuación, mayúsculas, números y caracteres especiales mezclado con elementos de diseño como: espacios en blanco, saltos de línea subrayados, asteriscos, tablas, formato de página, etc. ➔ permiten identificar subcomponentes como párrafos, títulos, fechas de publicación, nombres de autores, ...
- ➔ Dimensión estructural: secuencia de palabras
- ➔ Metadatos HTML WYSIWYG (What You See Is What You Get)

¿Qué debemos considerar?

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

No estructurado

Semiestructurado

Estructurado

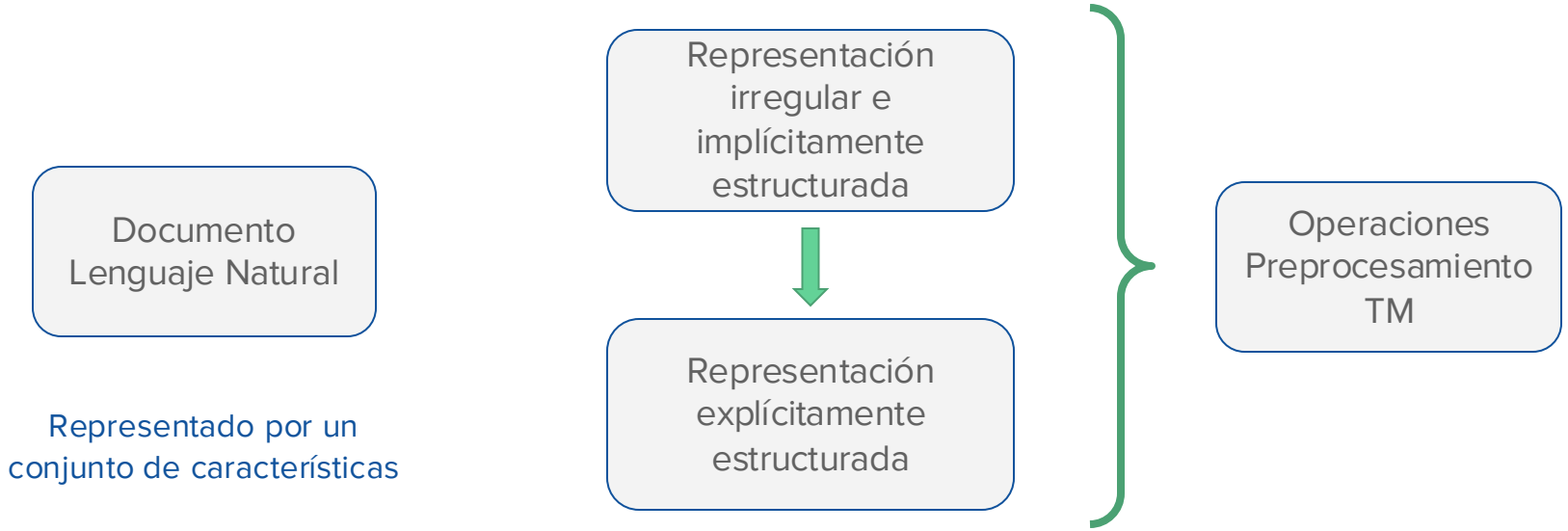
- ➔ Sistemas de Text Mining aplican los algoritmos de descubrimiento de conocimiento en colecciones de documentos preprocesados.
- ➔ Operaciones de preprocesamiento incluye diferentes tipos de técnicas seleccionadas y adaptadas de recuperación de información, la extracción de información y de investigación en lingüística computacional.

¿Para qué?

- ➔ Transforman contenido en bruto, no estructurado y en formato original en un formato estructurado, donde se aplican las operaciones de descubrimiento de conocimiento.

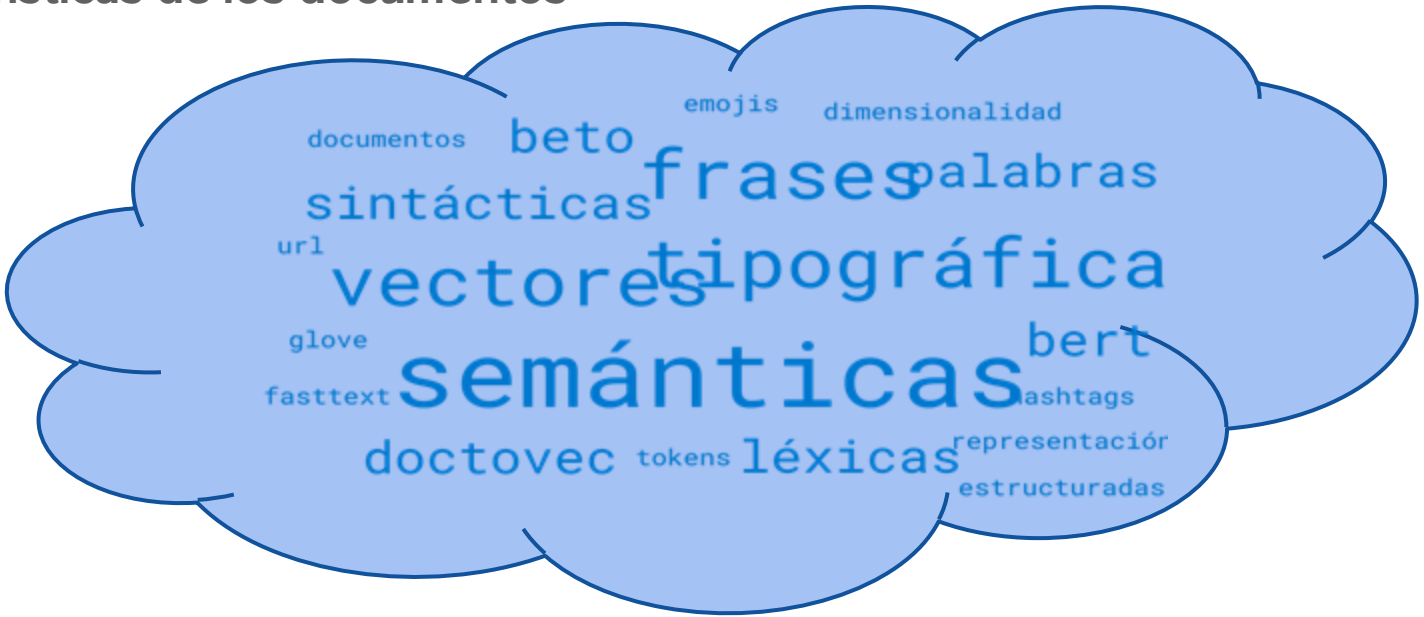
Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos



Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos



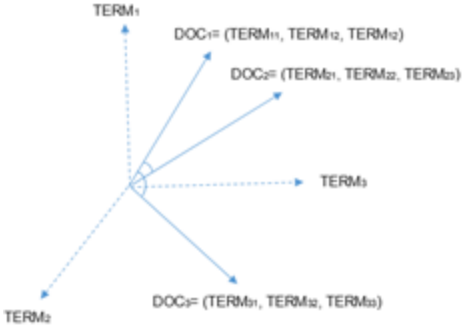
Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos

Feature dimensionality

Feature sparsity

¿Ejemplos?



Dimensión: términos
Documento: vector

Ref imagen: Framework for retrieving relevant contents related to fashion from online social network data

	Antonio y Cleopatra	Julio Cesar	La Tempestad	Hamlet	Otelo	Macbeth
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0
...						

Ejemplo libro: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos

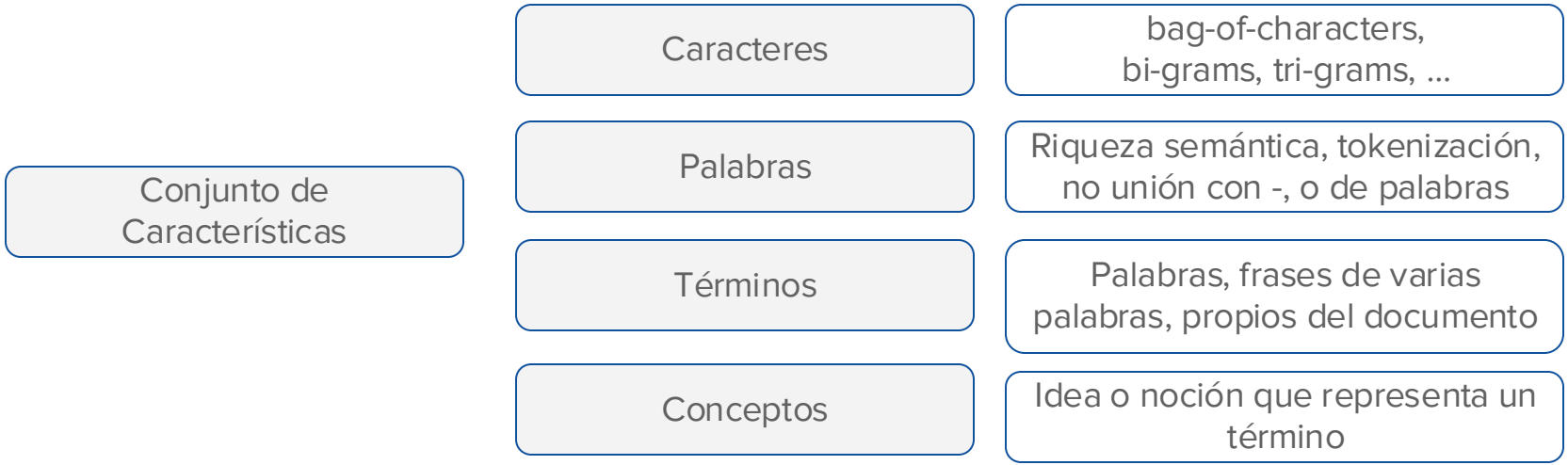
- Algoritmos de TM operan en base a características de los documentos
- Objetivos
  - Equilibrio entre volumen y nivel semántico de las características.

    ¿Para qué? Permite representar con precisión significado de un documento.

  - Identificar características más eficientes desde procesamiento computacional y práctica para descubrir patrones.

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos



Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos

- Caracteres
- Palabras
- Términos
- Conceptos
- Tokens

Ejemplo

- ➔ Palabra: libro
- ➔ Token: libro, libros (Tokenización, lematización)
- ➔ Término: Libro digital
- ➔ Concepto: “publicación que puede ser leída en dispositivo móvil o similar.” (idea abstracta)

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos

Normalización	Permite estandarizar el texto, por ejemplo eliminar tildes, dejar todo en minúscula, etc.
Tokenización	Permite dividir el texto de un documento en unidades más pequeñas, que son los tokens.
Lematización	Permite reducir las palabras de un texto a su forma base o canónica, considerando el contexto.
Stemming	Permite llegar a la raíz de una palabra, para esto es necesario cortar la palabra, sin considerar el contexto.

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Características de los documentos

Normalización	¡Hola, Mundo! ➔ hola mundo
Tokenización	¡Hola, Mundo! ➔ [“¡”, “H”, “o”, “l”, “a”, “,”, “M”, “u”, “n” ,“d” ,“o” ,“!”]
Lematización	corriendo ➔ correr
Stemming	corriendo, corrió, corres ➔ corr (alg. de Porter)

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

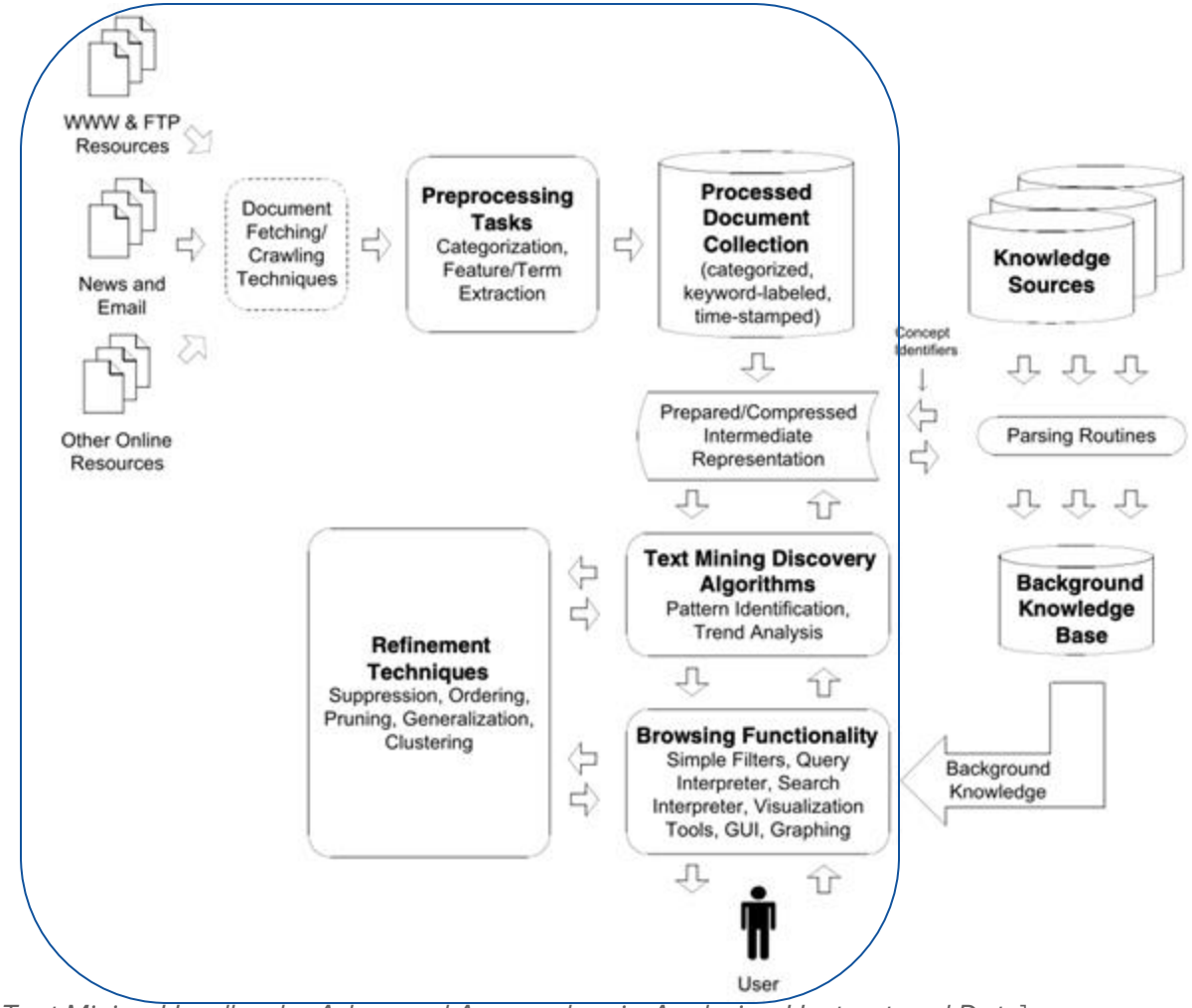
Considerar

- Dominios y antecedentes previos
- Búsqueda de patrones y tendencias
- Visualización

Unidad 1 - Fundamentos de Text Mining
Conceptos Básicos

Arquitectura sistemas de Text Mining (alto nivel)

[imagen: *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*]



[imagen: *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data*]