# Text Mining


Universidad
de Valparaíso
CHILE


ESCUELA DE INGENIERÍA
INFORMÁTICA
UNIVERSIDAD DE VALPARAÍSO

# Unidad 2 - Representación de Texto

➔ Contenidos
- **Modelo espacio vectorial**
- Representación distribuida
- word embeddings

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial

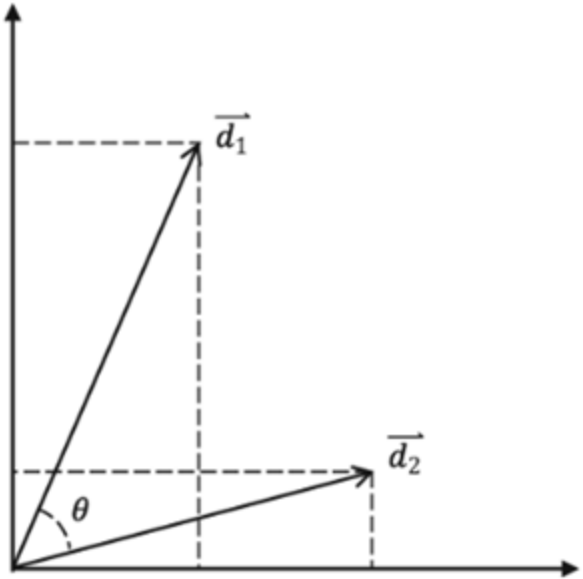
| | |
|---|---|
| Texto | Secuencia de caracteres con cierta granularidad: frases, sentencias, párrafos, o documento. |
| Término | Unidad más pequeña, que ya no se puede separar: caracter, palabras, ...<br>Texto corresponde a una colección de términos ($t_1$, $t_2$, ..., $t_n$) |
| Peso del término | Para un texto que contiene *n* términos, a cada término *t* se le asigna un peso *w*, que indica la importancia y relevancia en el texto.<br>Texto: ($t_1$:$w_1$ , $t_2$:$w_2$ , ..., $t_n$:$w_n$ ) |

Material basado en: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Vector Space Model (VSM)

- VSM considera, que un documento
  - Cada término es único
  - Los términos no tiene un orden
  - Considerar $(t_1, t_2, ..., t_n)$ como un sistema n-dimensional de **coordenadas** ortogonales
  - Texto puede ser representado como un **vector** n-dimensional $(w_1, w_2, ..., w_n)$
  - $d = (w_1, w_2, ..., w_n)$ – vector en VSM




¿Como armar el conjunto de términos?
¿Cómo calcular sus pesos?

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Construcción

- Primero (pre) procesamos el texto.
- Segundo: texto a secuencia de tokens
  - Palabras == Términos ➔ Coleccion == Vocabulario
  - Vocabulario
    - Texto
    - Léxico externo
  - Términos como bolsa de palabras (Bag-of-Words, BOW Model) – SVM
  - Con esto se fija el espacio vectorial.

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Construcción

- Tercero: Asignar los pesos
  - **Boolean Weight (BOOL)**

$$\text{BOOL}_i = \begin{cases} 1 \text{ , si } t_i \text{ aparece en el documento } d \\ 0 \text{ en otro caso} \end{cases}$$

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

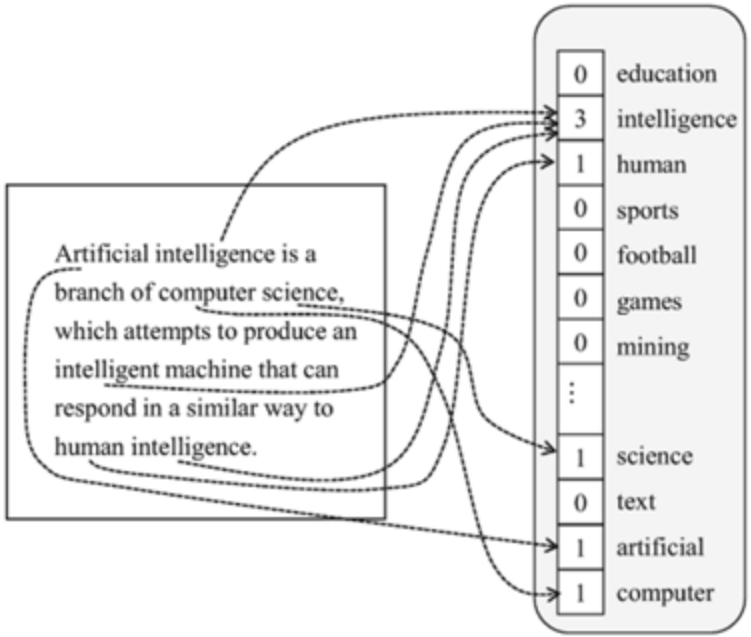
## Modelo Espacio Vectorial - Construcción

- Tercero: Asignar los pesos
  - **Term Frequency (TF)**
    - Frecuencia de un término en un documento
    - Idea: término con mayor frecuencia contiene más información que aquellos no frecuentes.

$$\mathrm{tf}_i = N(t_i, \boldsymbol{d})$$

Para palabras muy frecuente

$$f_i = \log(\mathrm{tf}_i + 1)$$




https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Construcción

- Tercero: Asignar los pesos
  - **Inverse Document Frequency (IDF)**
    - Indica la relevancia del término en el corpus.
    - DF: número de documentos que contienen el término específico en el corpus.
    - Cuando DF de un término es mayor, menor es la cantidad de información efectiva que tiene.

$$\mathrm{idf}_i = \log \frac{N}{\mathrm{df}_i}$$

$\mathrm{df}_i$: representa DF de $t_i$

N: Número total de documentos en el corpus.

Términos que aparecen pocas veces tendrán IDF alto, mientras que términos frecuentes tendrán IDF bajo.

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Construcción

- Tercero: Asignar los pesos
  - **Term Frequency - Inverted Document Frequency (TF-IDF)**

$$\text{tf\_idf}_i = \text{tf}_i \cdot \text{idf}_i$$

*"TF-IDF → características más discriminativas son aquellas que aparecen frecuentemente en el documento actual y raramente en otros documentos"*

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Normalización

- Textos tienen diferentes largos
- El largo del texto afecta en la representación
- ¿Qué pasa si duplicamos el texto?
- Normalizamos para reducir la influencia del largo del texto (text length normalization)

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Normalización

- Tipos de Normalización, dado $d = (w_1, w_2, ..., w_n)$
  - L1 norm normalization

$$d_1 = \frac{d}{\|d\|_1} = \frac{d}{\sum_i w_i}$$

Vectores normalizados están en el hiperplano $w_1 + w_2 + ... + w_n = 1$ del espacio vectorial.

  - L2 norm normalization

$$d_2 = \frac{d}{\|d\|_2} = \frac{d}{\sqrt{\sum_i w_i^2}}$$

Vectores normalizados están en la superficie esférica $w_1^2 + w_2^2 + ... + w_n^2 = 1$ del espacio vectorial.

  - Maximum word frequency normalization

$$d_{max} = \frac{d}{\|d\|_\infty} = \frac{d}{\max_i\{w_i\}}$$

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Ing. de características

- VMS considera que coordenadas en el espacio son ortogonal. Ejemplo que los términos son independientes entre sí.
- Problema:
  - “John is quicker than Mary”
  - “Mary is quicker than John”
- Surge *ingeniería de características*

Expresiones distintas que tendrán la misma representación del texto en VSM.

Proceso de definir de forma manual cuáles son las características que se utilizarán.

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Ing. de características

1) n-grams features

| | |
|---|---|
| | hoy nos juntaremos para estudiar cálculo multivariable |
| unigram | hoy, nos, juntaremos, para, estudiar, cálculo, multivariable |
| bigram | hoy nos, nos juntaremos, juntaremos para, para estudiar, estudiar cálculo, cálculo multivariable |
| trigram | hoy nos juntaremos, nos juntaremos para, juntaremos para estudiar, para estudiar cálculo, estudiar cálculo multivariable. |

https://link.springer.com/chapter/10.1007/978-981-16-0100-2_3

# Unidad 2 - Representación de Texto

**Modelo Espacio Vectorial - Ing. de características**

2) Syntactic features

- Analizar la frase basado en reglas gramaticales.
- Estructura del lenguaje, a través de la dependencia entre palabras
- Salida: árbol sintáctico

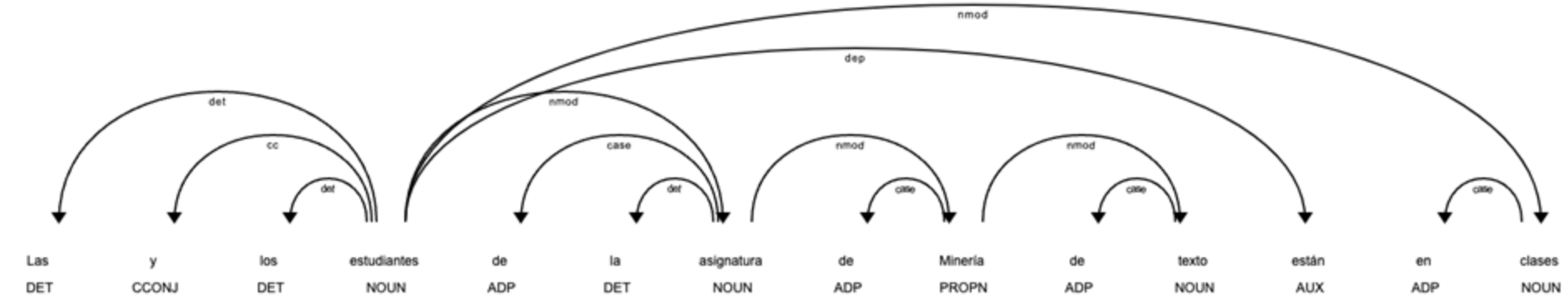
    - relación direccional, desde palabra dominante

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Ing. de características

2) Syntactic features

- Ejemplo: “Las y los estudiantes de la asignatura de Minería de Texto están en clases”



# Unidad 2 - Representación de Texto

**Modelo Espacio Vectorial - Ing. de características**

3) Lexicon Features

- Permite trabajar con polisemia y sinónimos
- Ejemplos: WordNet
- Tesaurio

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Otras Representaciones

- Concept Representation
  - Modelos tradicionales no logran captar las relaciones semánticas implícitas.
  - Otros métodos si lo permiten como:  Topic models
    - Latent Semantic Analysis (LSA)
    - Probabilistic Latent Semantic Analysis (PLSA)
    - Latent Dirichlet Allocation (LDA)

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Otras Representaciones

- Learning Representation
  - Deep learning: aprender vector densos de baja dimensionalidad.
  - Métodos de aprendizaje pueden captar  información semántica, evitando la ingeniería de características.
  - La representación del texto, varía según la tarea a resolver:
    - Análisis de sentimientos: atributos que representan emociones
    - Detección y seguimiento de temas: características de descripción de eventos.
  - No siempre existe una representación general del texto e ideal para todos los tipos de tareas. También se puede combinar con características de distintas tareas.

# Unidad 2 - Representación de Texto

## Modelo Espacio Vectorial - Otras Representaciones

- Learning Representation
  - Modelo muy utilizado: Bag-of-Word
    - Cada documento == vocabulario
    - Vectores one-hot (vectores booleanos)
    - Ejemplo: Primera palabra: [1,0, 0, 0, …., 0] , última palabra [0,0, 0, 0, …., 1] , si el vocabulario es de tamaño 30 mil, se tendrán 30 mil vectores.
    - Problemas
      - símbolos discretos, propenso a generar datos discretos
      - cada vector representa solo una palabra, por lo que dos palabras son independientes entre sí. El método no permite capturar similitudes entre palabras
    - Solución: representación distribuida, que ha sido tema estudiado y que ha tomado mucho interés por los resultados.