Representación del Texto

¿Enfoques Clásicos?

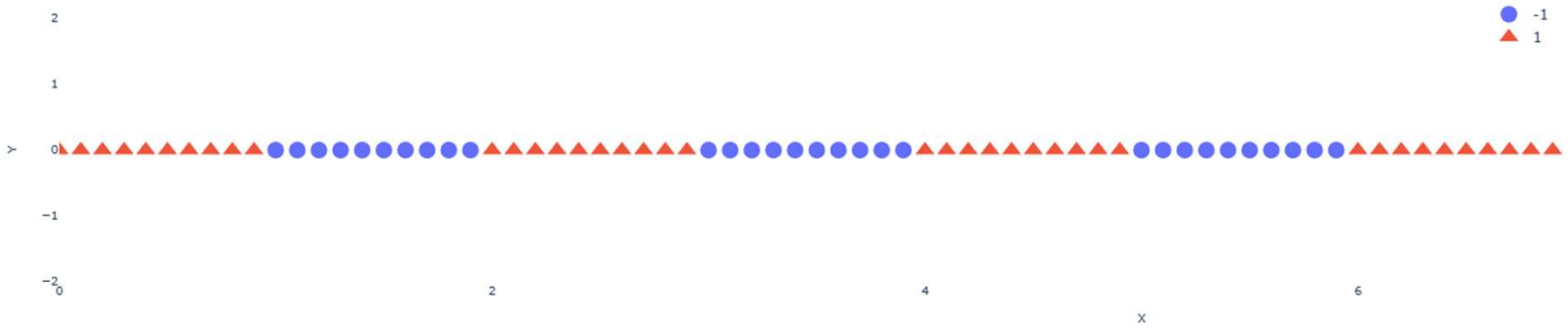
Universitat d'Alacant
Universidad de Alicante
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UJI UNIVERSITAT JAUME I
valgrAI
PLN

Contenidos

- Motivación
- Representación Mediante Texto Plano
- Representación Mediante Espacios Vectoriales

Introducción

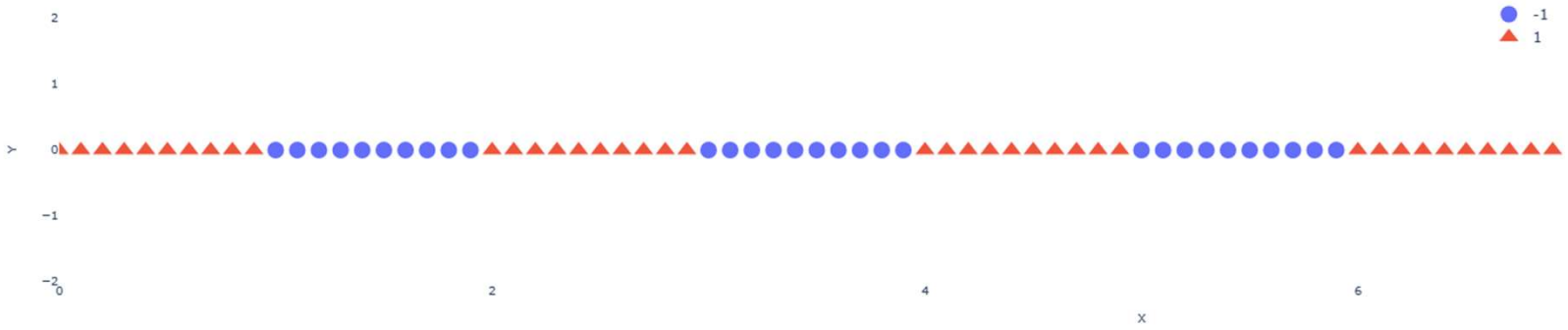
Representación del Texto. Motivación



¿será posible separar ambas clases trazando una única línea?

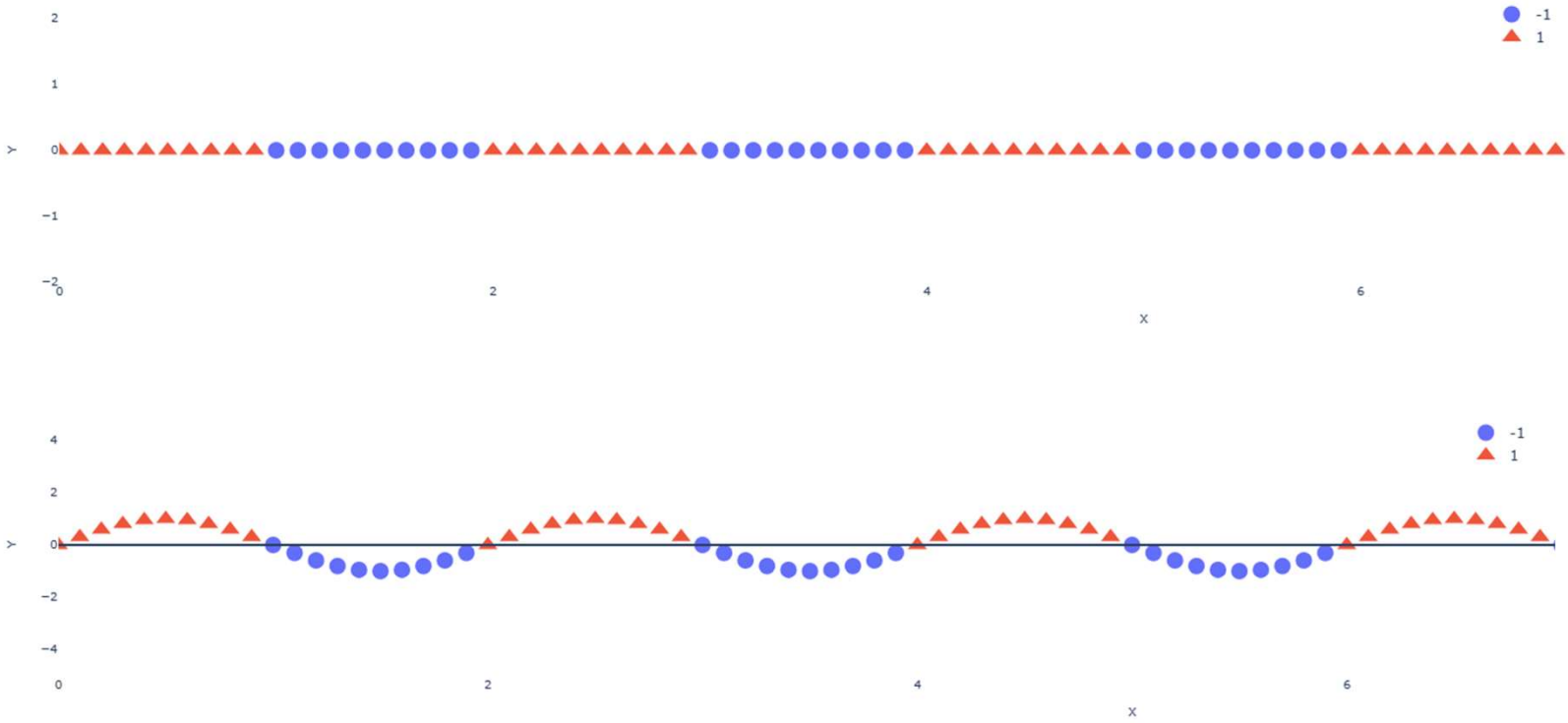
y si permitimos cierto margen de error, ¿podemos estimarlo?

Representación del Texto. Motivación



¿y si aplicamos alguna transformación a los datos sin cambiar su categoría, por ejemplo, $\sin(x)$?

Representación del Texto. Motivación



Representación del Texto

¿cómo representamos el texto al resolver un problema mediante aprendizaje automático?

Representación del Texto

Considerar

Características de los modelos

- Los modelos pueden limitar el tipo de datos que procesan.

Relevancia para el problema

- Deben capturar información relevante al problema.

Capacidad predictiva

- Deben permitir predecir la variable dependiente.

Coste de obtención

- No debe suponer una barrera. Ejemplo, requerir monto prohibitivo de recursos de cómputo o financieros.

Dimensionalidad

- Si es muy alta, partes del espacio pueden estar sub-representadas, influyendo negativamente la efectivad de los modelos

Representación Mediante Texto Plano

Texto Plano

- Representación más simple…pero quizás no la más adecuada para aprendizaje automático.

- Permite operaciones como búsquedas mediante expresiones regulares, o comparaciones con métricas como la distancia de Levenshtein

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)

LAGRANGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A

LAGRANGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P

LAPGRANGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L

LAPLGRANGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G

LAPLGRANGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G
- Borrar R

LAPLGRANGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G
- Borrar R
- Sustituir A x A

LAPLGRANGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G
- Borrar R
- Sustituir A x A
- Insertar C

LAPLGRACNGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G
- Borrar R
- Sustituir A x A
- Insertar C
- Borrar N

LAPLGRACNGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G
- Borrar R
- Sustituir A x A
- Insertar C
- Borrar N
- Borrar G

LAPLGRACNGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G
- Borrar R
- Sustituir A x A
- Insertar C
- Borrar N
- Borrar G
- Sustituir E x E

LAPLGRACNGE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- Insertar P
- Insertar L
- Borrar G
- Borrar R
- Sustituir A x A
- Insertar C
- Borrar N
- Borrar G
- Sustituir E x E

LAPLACE

Texto Plano. Medidas de Similitud

Representan la noción de “distancia” entre dos cadenas a partir de operaciones de edición

- Sustituciones (S)
- Inserciones (I)
- Borrado (B)
- Intercambio (W)

Ejemplo: LAGRANGE LAPLACE

- Sustituir L x L (no hacer nada)
- Sustituir A x A
- **Insertar P**
- **Insertar L**
- **Borrar G**
- **Borrar R**
- Sustituir A x A
- **Insertar C**
- **Borrar N**
- **Borrar G**
- Sustituir E x E

LAPLACE

Costos

S=0

I=1

B=1

Distancia = 7

Texto Plano. Medidas de Similitud

¿consideraciones?

Representación Mediante Espacios Vectoriales

Espacios Vectoriales

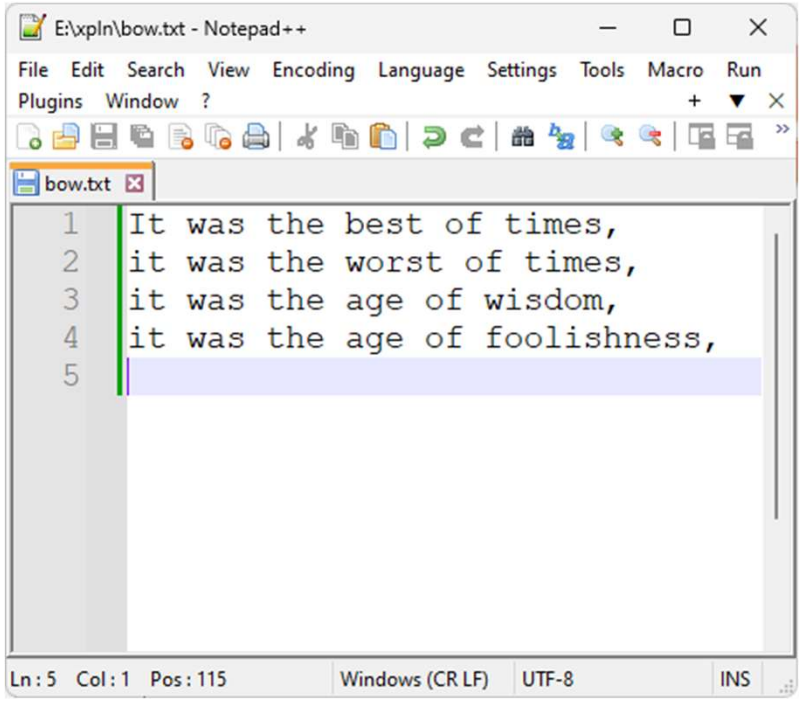
Bolsa de Palabras

Bolsa de Palabras (BoW)

- Simple de entender y utilizar.
- Describe la ocurrencia de palabras en el texto.
- No considera información sobre el orden o la estructura de las palabras.
- Extensible a n-gramas.

Bolsa de Palabras (BoW)

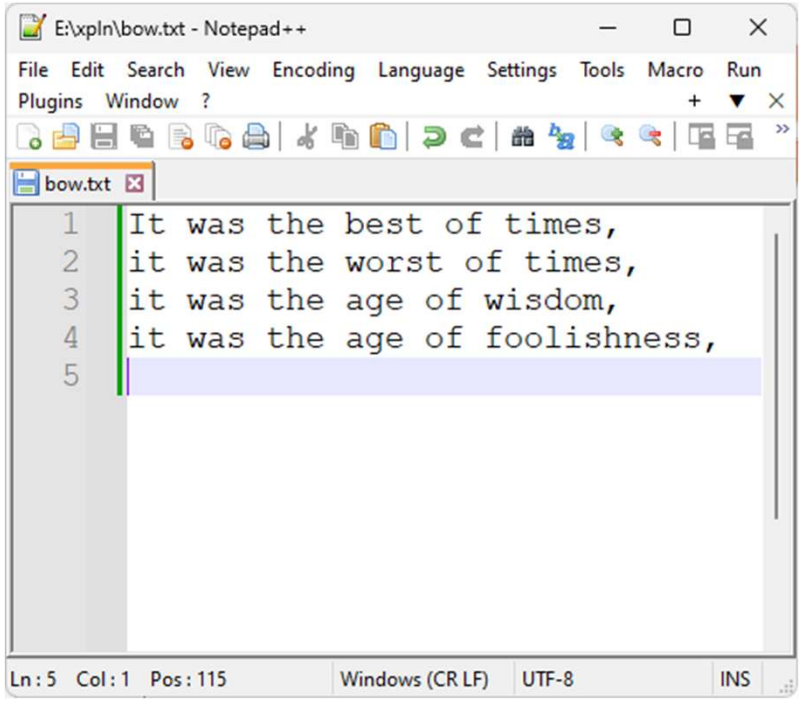
Consideraremos cada línea un documento.



Bolsa de Palabras (BoW)

Paso 1: Crear el vocabulario

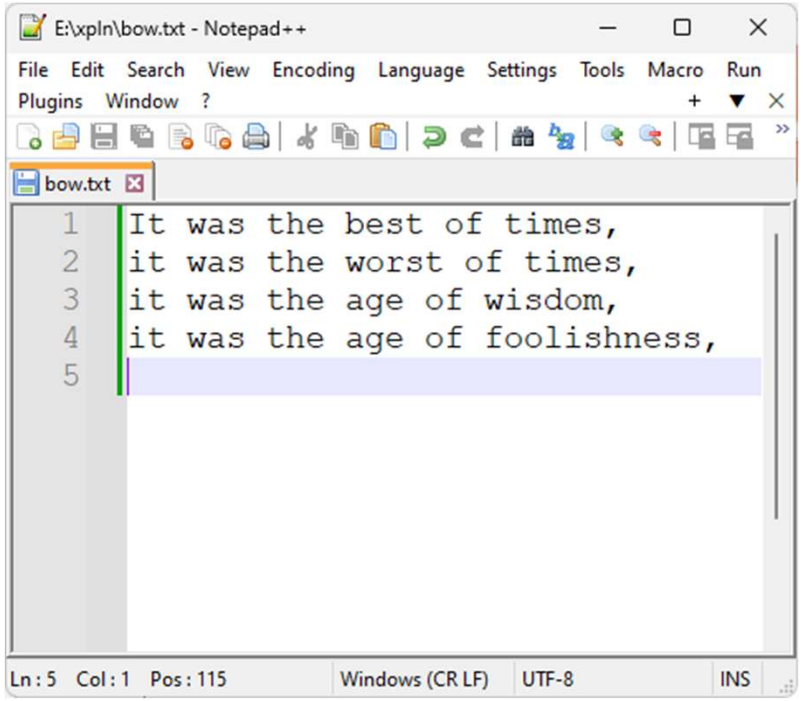
término
age
best
foolishness
it
off
the
times
was
wisdom
worst



Bolsa de Palabras (BoW)

Paso 2: Crear vectores para cada documento

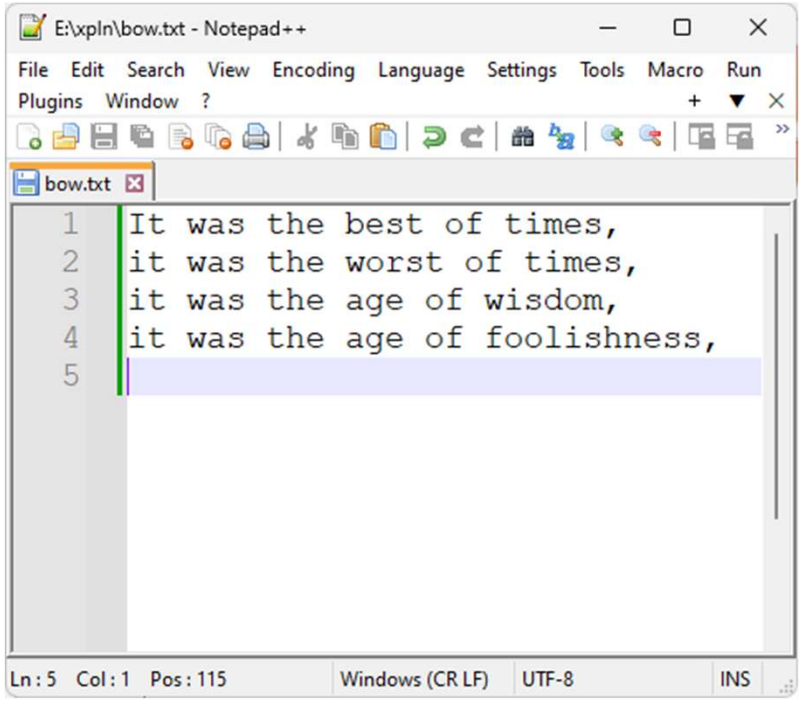
término	d1
age	0
best	1
foolishness	0
it	1
off	1
the	1
times	1
was	1
wisdom	0
worst	0



Bolsa de Palabras (BoW)

Paso 2: Crear vectores para cada documento

término	d1	d2	d3	d4
age	0	0	1	1
best	1	0	0	0
foolishness	0	0	0	1
it	1	1	1	1
off	1	1	1	1
the	1	1	1	1
times	1	1	0	0
was	1	1	1	1
wisdom	0	0	1	0
worst	0	1	0	0



Bolsa de Palabras (BoW)

¿consideraciones?

Term Frequency – Inverse Document Frequency

Term Frequency – Inverse Document Frequency (TF-IDF)

- Una de las representaciones más populares.

- Pondera cada palabra por:
  - Su frecuencia en el documento (TF)
  - El logaritmo del recíproco de su frecuencia en todo el corpus (IDF)

- Asume que la frecuencia de cada palabra provee información independiente a las otras palabras.

- Como BoW, no utiliza la similitud semántica entre las palabras.

Term Frequency – Inverse Document Frequency (TF-IDF)

Consideraremos cada línea un documento.

$$tf(t,d) = \frac{f(t,d)}{\sum_{t' \in d} f(t',d)}$$

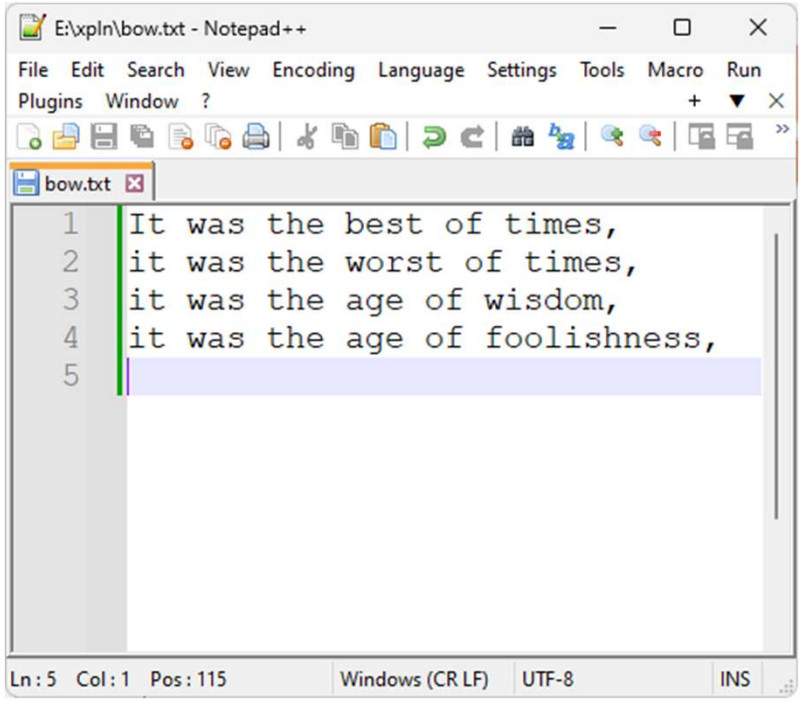
$$idf(t) = \log\left(\frac{(|D|)}{|\{d \in D : t \in d\}|}\right)$$

$$tf\_idf(t,D) = tf(t,D) * idf(t)$$

Donde

D es el corpus

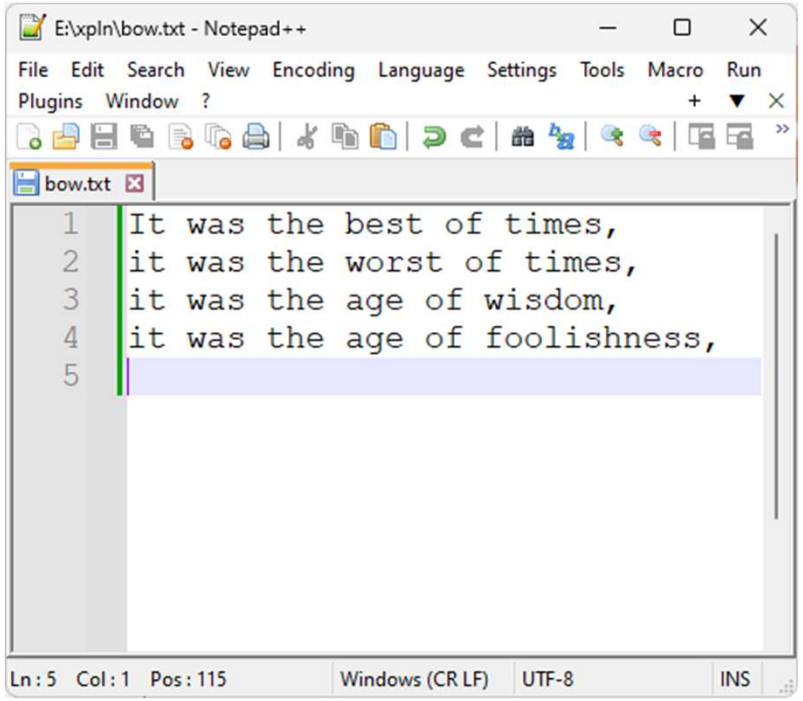
$f(d,t)$ es la frecuencia del término t en el documento d



Term Frequency – Inverse Document Frequency (TF-IDF)

Paso 1: Crear matriz frecuencia de términos (TF)

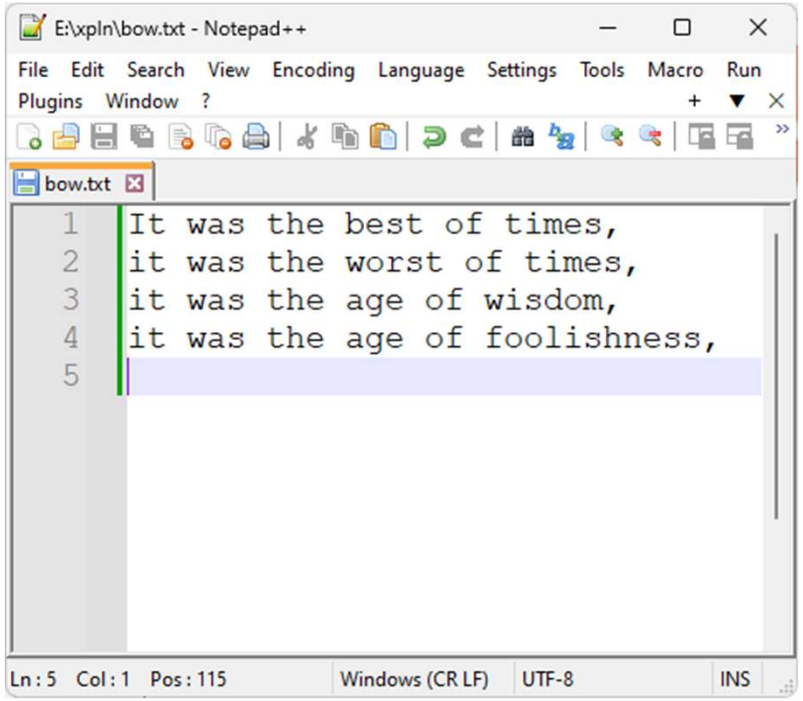
término	tf(t,d1)	tf(t,d2)	tf(t,d3)	tf(t,d4)
age	0	0	0.17	0.17
best	0.17	0	0	0
foolishness	0	0	0	0.17
it	0.17	0.17	0.17	0.17
off	0.17	0.17	0.17	0.17
the	0.17	0.17	0.17	0.17
times	0.17	0.17	0	0
was	0.17	0.17	0.17	0.17
wisdom	0	0	0.17	0
worst	0	0.17	0	0



Term Frequency – Inverse Document Frequency (TF-IDF)

Paso 2: Inversa Frecuencia Documento (IDF)

término	N	Dft	N/Dft	log(N/Dft)
age	4	2	2	0.30
best	4	1	4	0.60
foolishness	4	1	4	0.60
it	4	4	1	0
off	4	4	1	0
the	4	4	1	0
times	4	2	2	0.30
was	4	4	1	0
wisdom	4	1	4	0.60
worst	4	1	4	0.60



Term Frequency – Inverse Document Frequency (TF-IDF)

Paso 3: Calcular $tf\_idf(t, D) = tf(t, D) * idf(t)$

término	tf(t,d1)	tf(t,d2)	tf(t,d3)	tf(t,d4)
age	0	0	0.17	0.17
best	0.17	0	0	0
foolishness	0	0	0	0.17
it	0.17	0.17	0.17	0.17
off	0.17	0.17	0.17	0.17
the	0.17	0.17	0.17	0.17
times	0.17	0.17	0	0
was	0.17	0.17	0.17	0.17
wisdom	0	0	0.17	0
worst	0	0.17	0	0

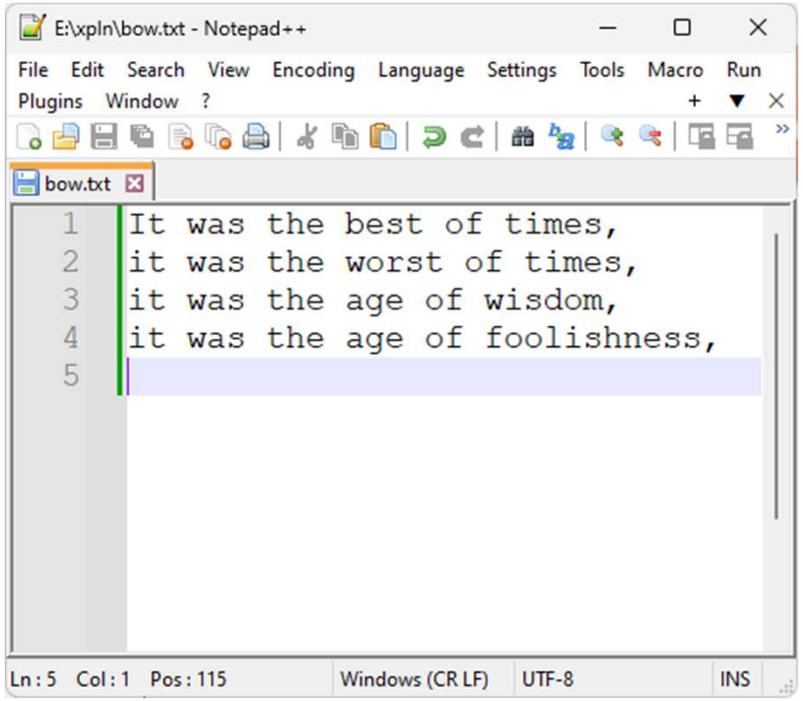
término	N	Dft	N/Dft	log(N/Dft)
age	4	2	2	0.30
best	4	1	4	0.60
foolishness	4	1	4	0.60
it	4	4	1	0
off	4	4	1	0
the	4	4	1	0
times	4	2	2	0.30
was	4	4	1	0
wisdom	4	1	4	0.60
worst	4	1	4	0.60

$$tf_{idf("age",D3)} = 0.17 * 0.3 = 0.05$$

Term Frequency – Inverse Document Frequency (TF-IDF)

Paso 4: Crear vectores para cada documento

término	d1	d2	d3	d4
age	0	0	0.05	0.05
best	0.1	0	0	0
foolishness	0	0	0	0.1
it	0	0	0	0
off	0	0	0	0
the	0	0	0	0
times	0.05	0.05	0	0
was	0	0	0	0
wisdom	0	0	0.1	0
worst	0	0.1	0	0



Term Frequency – Inverse Document Frequency (TF-IDF)

¿consideraciones?

Word Embeddings

Word Embeddings

- Son una representación “aprendida” donde palabras con similar significado tienen similar representación.

- Cada palabra se representa como un **vector denso** en $\mathbb{R}^n$ cuyas componentes se relacionan con otras palabras. En general los vectores tienen menor dimensión que representaciones tipo BoW.

- Basados en la idea de que, palabras con significado similar, aparecerán en contextos similares (Distributional Hypethesis).

- Generalizan mejor ante palabras no vistas o raras ya que la representación se basa en el contexto.

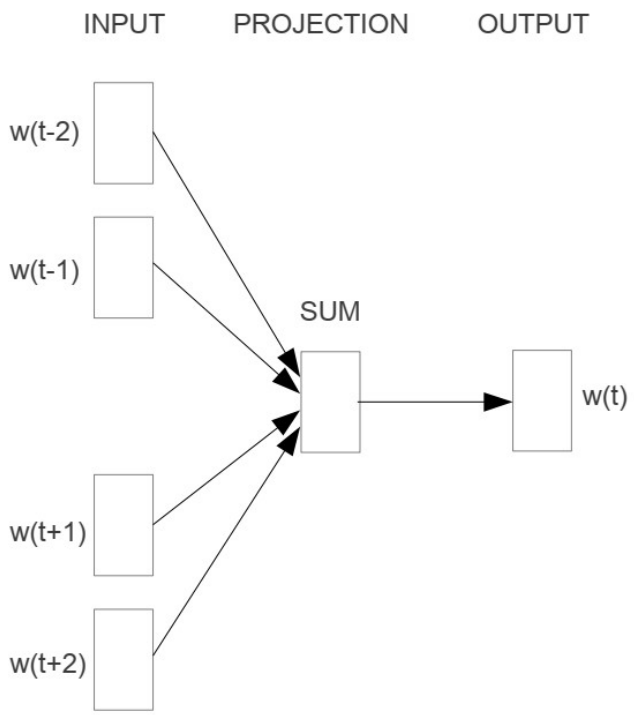
- Uno de los puntos de inflexión en aprendizaje profundo aplicado a NLP.

Word Embeddings. Word2Vec

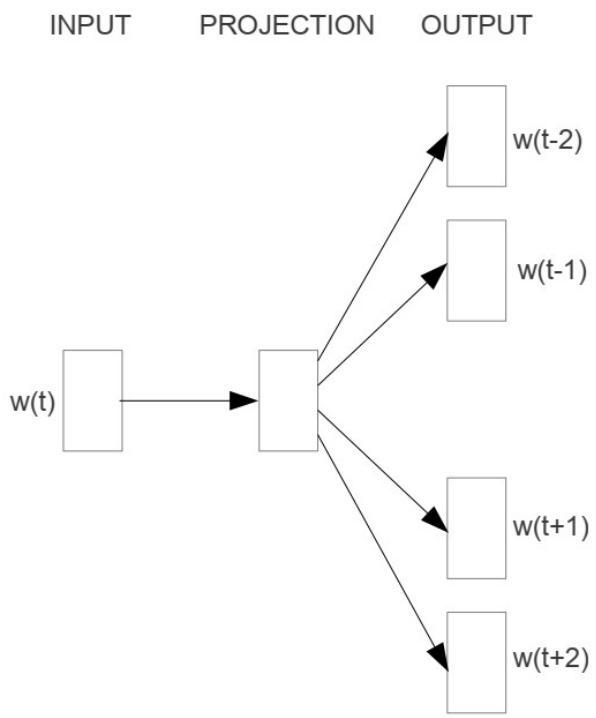
- Desarrollado por Google, 2013.

- Diferentes algoritmos de entrenamiento.
  - Continuous Bag-of-Words (CBOW): predice una a partir de su contexto.
  - Continous Skip-Gram Model: predice el contexto a partir de una palabra.

Word Embeddings. Word2Vec



CBOW



Skip-gram

Word Embeddings. GloVe

- Global Vectors for Word Representation

- No se basa sólo en información local, incorpora además información global (coocurrencias entre las palabras), siendo esa una diferencia distintiva respecto a Word2Vec

Word Embeddings.

¿consideraciones?
¿cómo comparar palabras … y documentos?

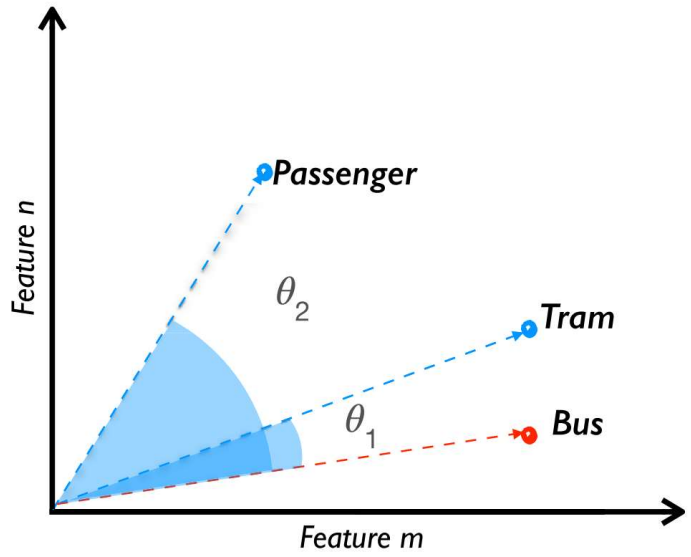
Comparando Palabras y Documentos: Cosine Similarity

$$\cos(\overrightarrow{t_1}, \overrightarrow{t_2}) = \frac{\overrightarrow{t_1}\overrightarrow{t_2}}{\|\overrightarrow{t_1}\|\|\overrightarrow{t_2}\|}$$

Donde $\|\vec{t}\|$ es la norma del $\vec{t}$ y $\overrightarrow{t_1}\overrightarrow{t_2}$ es el producto escalar.

$$\|\vec{t}\| = \sqrt[p]{|t_1|^p + |t_2|^p + \cdots + |t_n|^p}$$

$$\overrightarrow{t_1}\overrightarrow{t_2} = \vec{t}_1^1\vec{t}_1^2 + \vec{t}_2^1\vec{t}_2^2 + .. + .\vec{t}_n^1\vec{t}_n^2$$



Kalwar, S., Rossi, M., & Sadeghi, M. (2023). Automated creation of mappings between data specifications through linguistic and structural techniques. *IEEE Access*, *11*, 30324-30339.

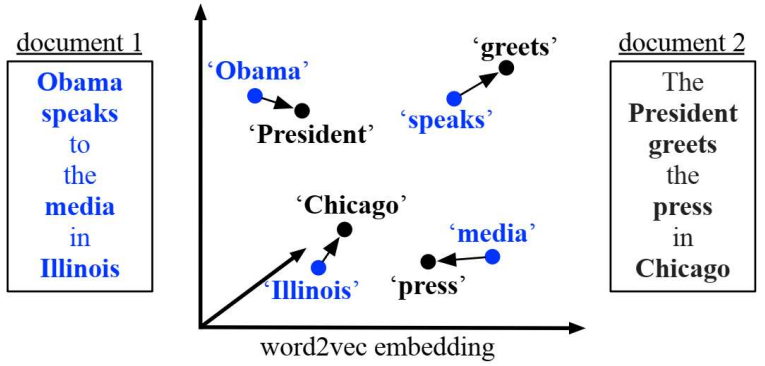
Comparando Documentos: Word Mover's Distance

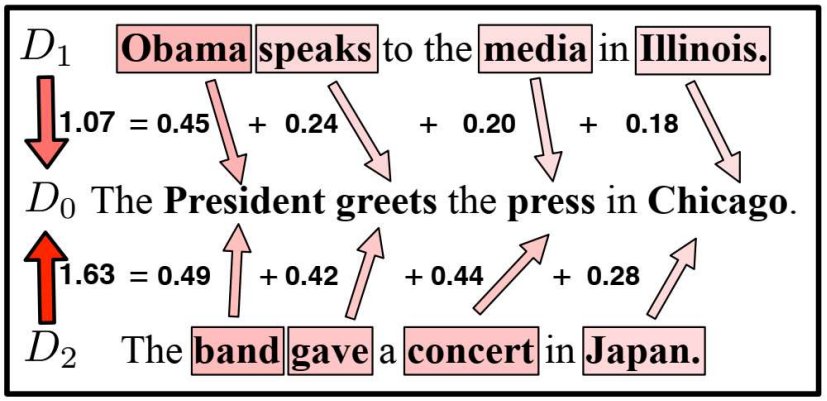
- Se basa en que, las distancias entre los vectores codifican información semántica.
- Los documentos se representan como un conjunto de vectores de palabras.
- La distancia entre dos documentos d1 y d2 es el mínimo de la distancia acumulada de todas palabras de d1 para llegar a d2 en el espacio vectorial.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

Comparando Documentos: Word Mover's Distance

- Se basa en que, las distancias entre los vectores codifican información semántica.
- Los documentos se representan como un conjunto de vectores de palabras.
- La distancia entre dos documentos d1 y d2 es el mínimo de la distancia acumulada de todas palabras de d1 para llegar a d2 en el espacio vectorial.





Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966). PMLR.

Word Embeddings.

¿consideraciones?
¿cómo comparar palabras … y documentos?

¿Document Embeddings?

Document Embeddings

...

Preparación de Datos

Fin