

Aplicaciones específicas

Taller de Procesamiento de Señales

Agenda

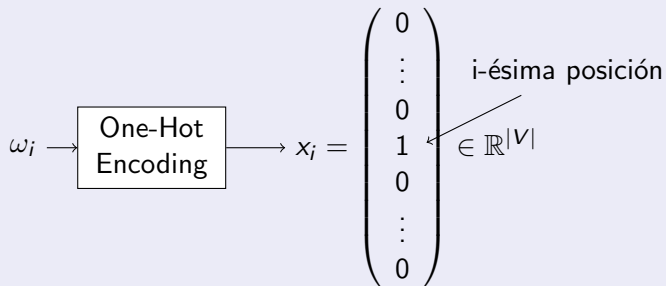
1 Lenguaje Natural

2 Sistemas de Recomendación

¿Como convertir un texto en un vector?

One-hot Encoding

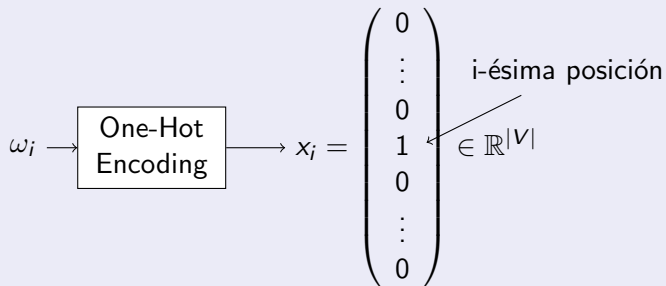
Dado un vocabulario $V = \{\omega_1, \dots, \omega_{|V|}\}$, se puede convertir cada palabra en un vector *one-hot*.



¿Como convertir un texto en un vector?

One-hot Encoding

Dado un vocabulario $V = \{\omega_1, \dots, \omega_{|V|}\}$, se puede convertir cada palabra en un vector *one-hot*.



Bolsa de palabras

Para vectorizar un documento $f(x_1, \dots, x_n)$, la manera más simple es *bolsa de palabras*: $f(x_1, \dots, x_n) = x_1 + \dots + x_n$.

Procesamiento del Lenguaje Natural

Vectorizaciones Sofisticadas

En la práctica suelen utilizarse representaciones pre-entrenadas (ej. FastText).

Procesamiento del Lenguaje Natural

Vectorizaciones Sofisticadas

En la práctica suelen utilizarse representaciones pre-entrenadas (ej. FastText).

Normalizaciones de NLP

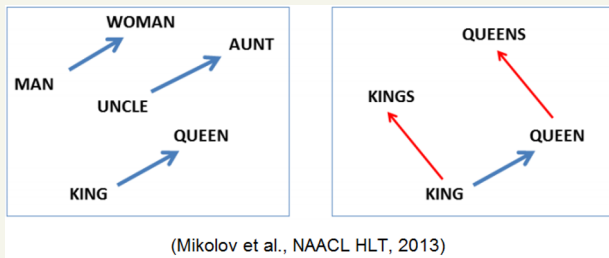
- Eliminar caracteres raros e inusuales
- Convertir todo a minúsculas
- Eliminar palabras no informativas (stop words)
- Descartar las palabras poco observadas
- Descartar las palabras más comunes
- Lemmatization (significado)
- Stemming (quedarse con la raíz)

Term Frequency - Inverse Document Frequency

Transformación tf-idf

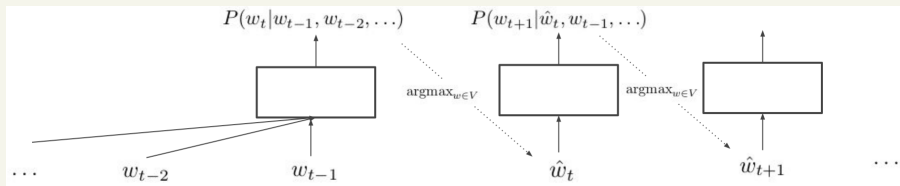
Medida numérica que expresa cuán relevante es una palabra para un documento dentro de un dataset. El tf-idf para un término t de un documento d perteneciente a una colección de n documentos es $\text{tf-idf}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t)$. El primer factor $\text{tf}(t, d) = \frac{\#(t \in d)}{\#(d)}$ es la cantidad de veces que aparece el término t en el documento d dividido la cantidad de términos que aparecen en el documento d . El segundo factor $\text{idf}(t) = 1 - \log\left(\frac{\text{df}(t)}{n}\right)$, donde $\text{df}(t)$ es la cantidad de documentos que poseen el término t en su interior.

Word Vectors + PCA



$$\text{vector}(\text{KINGS}) - \text{vector}(\text{KING}) + \text{vector}(\text{QUEEN}) = \text{vector}(\text{QUEENS})$$

Síntesis de texto

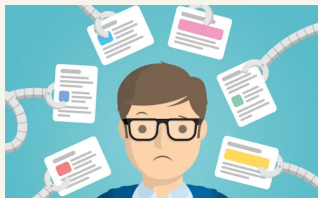


Outline

1 Lenguaje Natural

2 Sistemas de Recomendación

Sistemas de Recomendación




Problemáticas asociadas

- *Cámara de eco.* Los algoritmos de recomendación tienden a juntar a personas con ideología similar, creando un ciclo de realimentación donde todos escuchan lo que ya creen, no se expone a puntos de vista diferentes, fomenta la radicalización y el dogmatismo.
- *Filtro burbuja.* Los algoritmos filtran el contenido que no coincide con tus intereses o interacciones previas, creando una especie de burbuja en la que solo accedés a información que refuerza tus creencias.
- *Manipulaciones.* Muchos de estos algoritmos no publican su código, y por lo tanto no hay garantías que no se fomente algún tipo de contenido en particular.

Filtro Colaborativo

Aprender por Colaboración











	Item 1	Item 2	Item 3	Item 4	Item 5
Alice					
Bob					
Charlie					


Bob ~ Charlie



Filtro Colaborativo

Aprender por Colaboración

	Item 1	Item 2	Item 3	Item 4	Item 5
Alice					
Bob					
Charlie					

Bob ~ Charlie \Rightarrow ? = 

Entrenamiento

$$\min_{x, \theta} \frac{1}{2} \sum_{(i,j): y_{i,j} > 0} \left(\theta_j^T \cdot x_i - y_{i,j} \right)^2 + \frac{\lambda}{2} \left(\sum_{i=1}^{n_{\text{items}}} \|x_i\|^2 + \sum_{j=1}^{n_{\text{users}}} \|\theta_j\|^2 \right)$$

donde $y \in \mathbb{N}^{n_{\text{items}} \times n_{\text{users}}}$ contiene el dataset, $x \in \mathbb{R}^{n_{\text{items}} \times \nu}$ y $\theta \in \mathbb{R}^{n_{\text{users}} \times \nu}$ son los parámetros a entrenar; con ν la dimensión del espacio latente y $\lambda \geq 0$ un hiperparámetro de regularización.

Inferencia (Rating)

$$\hat{y}_{i,j} = p(\theta_j^T \cdot x_i) + (1 - p)\bar{y}_i$$

donde \bar{y}_i es la calificación promedio del item i -ésimo y $0 \leq p \leq 1$ es un hiperparámetro que indica cuanto peso le damos al aprendizaje y cuanto al valor medio.