Aplicaciones específicas

Taller de Procesamiento de Señales

Agenda

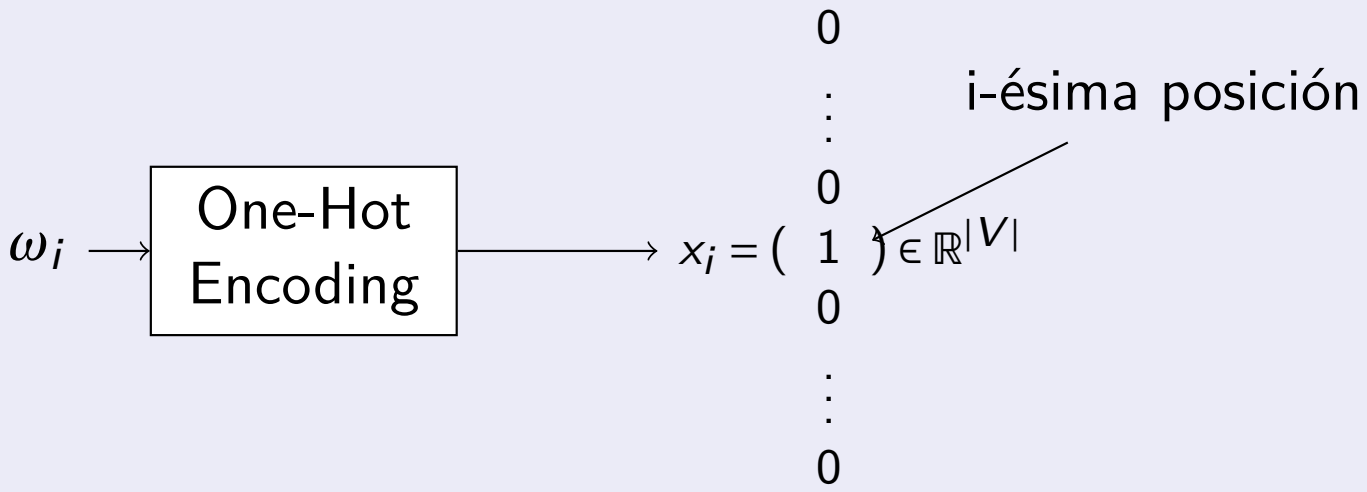
1. Modelo de Lenguaje

2. Sistemas de Recomendación

¿Como convertir un texto en un vector?

Word2vec

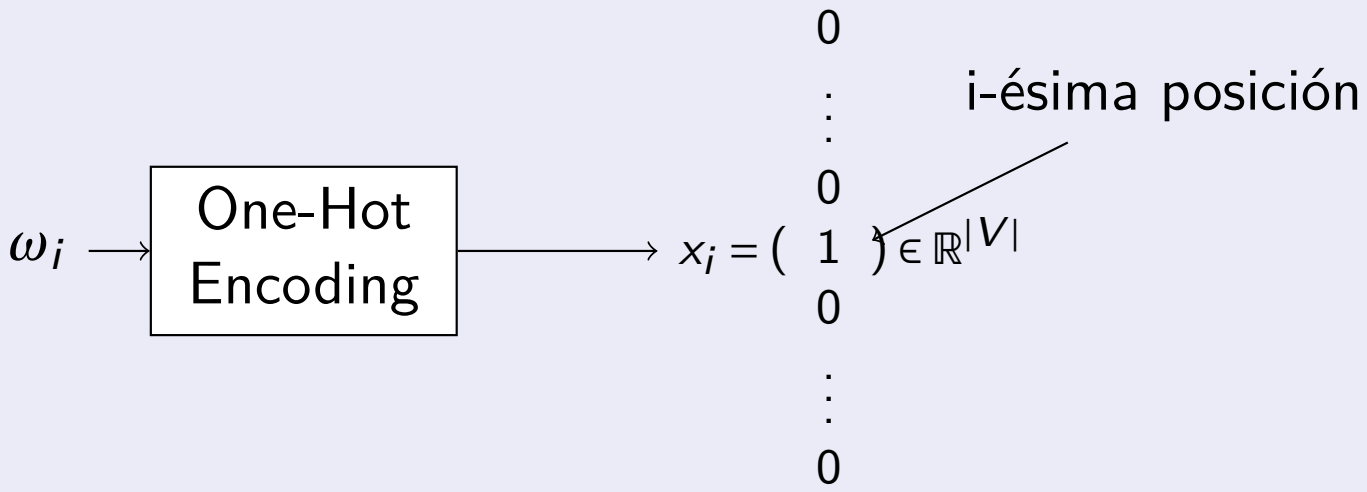
El método más sencillo para convertir una palabra en un vector es el *One-hot Encoding*. Dado un vocabulario $V = \{\omega_1, \cdots, \omega_{|V|}\}$, se puede convertir cada palabra en un vector *one-hot*.



¿Como convertir un texto en un vector?

Word2vec

El método más sencillo para convertir una palabra en un vector es el *One-hot Encoding*. Dado un vocabulario $V = \{\omega_1, \cdots, \omega_{|V|}\}$, se puede convertir cada palabra en un vector *one-hot*.



Bolsa de palabras

La vectorización de un documento consiste en definir una función $f(x_1, \cdots, x_n)$. El método más simple es la *bolsa de palabras* $f(x_1, \cdots, x_n) = x_1 + \cdots + x_n$, donde cada coeficiente representa la cantidad de veces que apareció cada palabra del vocabulario.

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.

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.

Vectorización

Transformación tf-idf se puede utilizar para vectorizar: cada documento d se puede expresar como un vector cuya dimensión es el largo del vocabulario y se define como $v_d = [\text{tf-idf}(0,d), \cdots, \text{tf-idf}(|V|,d)]^T$.

tf-idf ejemplo

- “El corazón tiene razones que la razón no conoce”

  Blaise Pascal

- “La razón debe conocer las razones del corazón y cualquier otra razón”

  Leonora Carrington

- “La anhelante espera apuñala el corazón”

  Mika Waltari

- “¿Qué espera tu corazón cuando niega lo que espera con su desesperación?”

  José Bergamín

tf-idf ejemplo: Stop words

- "~~El~~ corazón tiene razones ~~que~~ ~~la~~ razón ~~no~~ conoce"

  Blaise Pascal

- "~~La~~ razón debe conocer ~~las~~ razones ~~del~~ corazón ~~y~~ cualquier otra razón"

  Leonora Carrington

- "~~La~~ anhelante espera apuñala ~~el~~ corazón"

  Mika Waltari

- "¿~~Qué~~ espera ~~tu~~ corazón cuando niega ~~lo~~ ~~que~~ espera ~~con~~ ~~su~~ desesperación?"

  José Bergamín

tf-idf ejemplo: Matriz de Conteo (TF puro)

Frase 1: "El corazón tiene razones que la razón no conoce"
Frase 2: "La razón debe conocer las razones del corazón y cualquier otra razón"
Frase 3: "La anhelante espera apuñala el corazón"
Frase 4: "¿Qué espera tu corazón cuando niega lo que espera con su desesperación?"

	anhelante	apuñala	conoce	conocer	corazón	cualquier	cuando	debe	desesperación	espera	niega	otra	qué	razones	razón	tiene
Frase 1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1
Frase 2	0	0	0	1	1	1	0	1	0	0	0	1	0	1	2	0
Frase 3	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0
Frase 4	0	0	0	0	1	0	1	0	1	2	1	0	1	0	0	0

tf-idf ejemplo: Matriz TF Normalizado: $\frac{\#(t \in d)}{\#(d)}$

Frase 1: "El corazón tiene razones que la razón no conoce"
Frase 2: "La razón debe conocer las razones del corazón y cualquier otra razón"
Frase 3: "La anhelante espera apuñala el corazón"
Frase 4: "¿Qué espera tu corazón cuando niega lo que espera con su desesperación?"

	anhelante	apuñala	conoce	conocer	corazón	cualquier	cuando	debe	desesperación	espera	niega	otra	qué	razones	razón	tiene
Frase 1	.00	.00	.20	.00	.20	.00	.00	.00	.00	.00	.00	.00	.00	.20	.20	.20
Frase 2	.00	.00	.00	.13	.13	.13	.00	.13	.00	.00	.00	.13	.00	.13	.25	.00
Frase 3	.25	.25	.00	.00	.25	.00	.00	.00	.00	.25	.00	.00	.00	.00	.00	.00
Frase 4	.00	.00	.00	.00	.14	.00	.14	.00	.14	.29	.14	.00	.14	.00	.00	.00

tf-idf ejemplo: Vector IDF: $1 - \log\left(\frac{\mathrm{df}(t)}{n}\right)$

Frase 1: "El corazón tiene razones que la razón no conoce"
Frase 2: "La razón debe conocer las razones del corazón y cualquier otra razón"
Frase 3: "La anhelante espera apuñala el corazón"
Frase 4: "¿Qué espera tu corazón cuando niega lo que espera con su desesperación?"

	anhelante	apuñala	conoce	conocer	corazón	cualquier	cuando	debe	desesperación	espera	niega	otra	qué	razones	razón	tiene
IDF	3	3	3	3	1	3	3	3	3	2	3	3	3	2	2	3

tf-idf ejemplo: Matriz TF-IDF:

$v_d = [\text{tf-idf}(0, d), \cdots, \text{tf-idf}(|V|, d)]^T$

Frase 1: "El corazón tiene razones que la razón no conoce"
Frase 2: "La razón debe conocer las razones del corazón y cualquier otra razón"
Frase 3: "La anhelante espera apuñala el corazón"
Frase 4: "¿Qué espera tu corazón cuando niega lo que espera con su desesperación?"

	anhelante	apuñala	conoce	conocer	corazón	cualquier	cuando	debe	desesperación	espera	niega	otra	qué	razones	razón	tiene
Frase 1	.00	.00	.60	.00	.20	.00	.00	.00	.00	.00	.00	.00	.00	.40	.40	.60
Frase 2	.00	.00	.00	.37	.12	.37	.00	.37	.00	.00	.00	.37	.00	.25	.50	.00
Frase 3	.75	.75	.00	.00	.25	.00	.00	.00	.00	.50	.00	.00	.00	.00	.00	.00
Frase 4	.00	.00	.00	.00	.14	.00	.43	.00	.43	.57	.43	.00	.43	.00	.00	.00

significado –da

3 Cosa significada (→ significar [1]) [por otra].
Esp: Concepto o pensamiento representado [por una palabra o grupo de palabras]. En lingüística se opone a significante y a sentido.

Diccionario del español actual

Carta abierta a la FIFA: científicos advierten por el riesgo del calor extremo en el Mundial 2026 14/05/2026 | Deportes

“Su vida pende de un hilo”: preocupación y carta de 113 premios Nobel por la salud de Narges Mohammadi, condenada por el régimen de Irán 12/05/2026 | Mundo

El fabricante de los electrodomésticos Peabody pidió la apertura de su concurso de acreedores

> *En una **carta** que envió a clientes y proveedores, informó que atraviesa “una etapa de reestructuración de pasivos”.* 03/03/2026 | Economía

Carta abierta a la FIFA: científicos advierten por el riesgo del calor extremo en el Mundial 2026
14/05/2026 | Deportes

“Su vida pende de un hilo”: preocupación y carta de 113 premios Nobel por la salud de Narges Mohammadi, condenada por el régimen de Irán
12/05/2026 | Mundo

El fabricante de los electrodomésticos Peabody pidió la apertura de su concurso de acreedores

> *En una **carta** que envió a clientes y proveedores, informó que atraviesa “una etapa de reestructuración de pasivos”.*
> 03/03/2026 | Economía

Tarot: ¿qué significan los cuatro palos de las cartas?
10/05/2024 | Astrología

Carta abierta a la FIFA: científicos advierten por el riesgo del calor extremo en el Mundial 2026 14/05/2026 | Deportes

"Su vida pende de un hilo": preocupación y carta de 113 premios Nobel por la salud de Narges Mohammadi, condenada por el régimen de Irán 12/05/2026 | Mundo

El fabricante de los electrodomésticos Peabody pidió la apertura de su concurso de acreedores

> *En una **carta** que envió a clientes y proveedores, informó que atraviesa "una etapa de reestructuración de pasivos".* 03/03/2026 | Economía

Tarot: ¿qué significan los cuatro palos de las cartas? 10/05/2024 | Astrología

Son holandeses, apostaron a la Argentina y abrieron en Recoleta un restaurante de lujo y un café siempre lleno de vecinos

> *[...] Presencia es un fine dining diferente que no ofrece menú de pasos, sino un servicio a la **carta** con base en cocina europea.* 07/05/2026 | Gourmet

Wordnet (1985 / 1998)

Base de datos léxica que agrupa palabras en inglés en conjuntos de sinónimos llamados synsets.

Un **Synset** representa un concepto único y contiene diversos campos y relaciones semánticas

Atributos Básicos

- **Definición y Ejemplos** del concepto.
- **Lemas:** Las palabras exactas que expresan este significado.
- **Antónimos:** Términos con significado directamente opuesto.

Relaciones Parte-Todo

- **Holónimos:** El "todo" al que pertenece el concepto. Puede ser de *Miembro*, *Parte* o *Sustancia*.
- **Merónimos:** Las "partes" que componen al concepto.

Jerarquía (Relación ES-UN)

- **Hiperónimos:** Conceptos más generales (ej. *vehículo* para *coche*).
- **Hipónimos:** Conceptos más específicos (ej. *coche* para *vehículo*).

Otras Relaciones

- **Implicaciones:** Acciones obligatorias que se desprenden de un verbo (ej. *comprar* obliga a *pagar*).
- **Similares a:** Agrupa adjetivos con significados cercanos o relacionados.

Open Multilingual Wordnet: Reporte para 'carta'

```
-- SYNSET 1:  card.n.01 (n) --

Definición:
   one of a set of small pieces of stiff paper marked in
   various ways and used for playing games or for telling
   fortunes
Ejemplos:
   ['he collected cards and traded them with the other
   boys']
Lemmas:
   ['carta']

Hiperónimos:
   ['paper.n.01']

Hipónimos:
   ['tarot_card.n.01', 'punched_card.n.01',
   'playing_card.n.01', 'trading_card.n.01']
```

Open Multilingual Wordnet: Reporte para 'carta'

```
-- SYNSET 4:  menu.n.01 (n) --

Definición:
   a list of dishes available at a restaurant

Ejemplos:
   ['the menu was in French']

Lemmas:
   ['carta', 'menú', 'menú_del_día']

Hiperónimos:
   ['bill.n.07']

Hipónimos:
   ['a_la_carte.n.01', "table_d'hote.n.01",
   'prix_fixe.n.01']
```

Open Multilingual Wordnet: Reporte para 'carta'

```
-- SYNSET 5:  letter.n.01 (n) --
Definición:
   a written message addressed to a person or organization
Ejemplos:
   ['mailed an indignant letter to the editor']
Lemmas:
   ['carta']
Hiperónimos:
   ['text.n.01', 'document.n.02']
Hipónimos:
   ['round_robin.n.02', 'dead_letter.n.02', 'business_letter.n.01',
   'encyclical.n.01', 'epistle.n.01', 'invitation.n.01', 'open_letter.n.01',
   'letter_of_intent.n.01', 'covering_letter.n.01', 'chain_letter.n.01',
   'airmail_letter.n.01', 'form_letter.n.01', 'crank_letter.n.01',
   'fan_letter.n.01', 'pastoral.n.02', 'personal_letter.n.01']
Holónimos:
   ['correspondence.n.01', 'mail.n.01']
Merónimos:
   ['postscript.n.01', 'line.n.05', 'address.n.06']
```

significado –da

"Conocerás una palabra por la compañía que mantiene".

John Rupert Firth 1957

Procesamiento del Lenguaje Natural

La palabra es un símbolo de un símbolo.
El lenguaje es un sistema de signos arbitrarios

Jorge Luis Borges

Procesamiento del Lenguaje Natural

La palabra es un símbolo de un $\underbrace{\text{símbolo}}_{t-1}$.

El $\underbrace{\text{lenguaje}}_{t-2}$ es un $\underbrace{\text{sistema}}_{t}$ de $\underbrace{\text{signos}}_{t+1}$ $\underbrace{\text{arbitrarios}}_{t+2}$

Procesamiento del Lenguaje Natural

La palabra es un símbolo de un $\underbrace{\text{símbolo}}_{t-1}$.

El $\underbrace{\text{lenguaje}}_{t-2}$ es un $\underbrace{\text{sistema}}_{t}$ de $\underbrace{\text{signos}}_{t+1}$ $\underbrace{\text{arbitrarios}}_{t+2}$

Definimos

- $\mathcal{V}$ El vocabulario del corpus (conjunto de palabras).
- $c \in \mathcal{V}$ La palabra Central.
- $w \in \mathcal{V}$ Una palabra del contexto (palabras que la rodean).

Problemas

- $p(c|w)$ La probabilidad de que la palabra central c aparezca dado el contexto w.
- $p(w|c)$ La probabilidad de que el contexto w aparezca dado la palabra central c.

Word2vec: Skip-gram (2013) - $p(w|c)$

Objetivo

Encontrar vectorizaciones U_w (del contexto) y V_c (de la palabra central) tal que el producto $U_w^T V_c$ de información de si w se encuentra en el contexto de c.

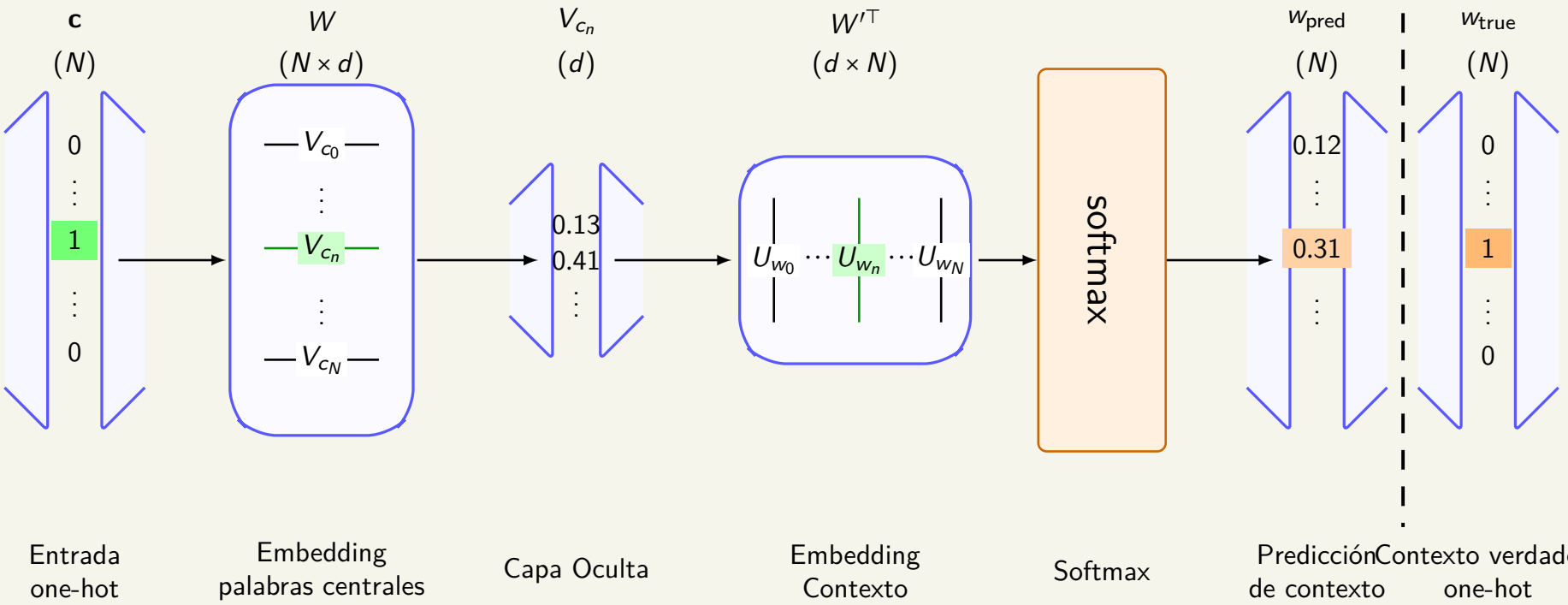
⚠ Permitimos codificaciones distintas para las palabras cuando aparecen como contexto o como palabra central.

Propuesta: softmax

$$p(w|c,\theta) = \frac{e^{U_w^T V_c}}{\sum_{w' \in \mathcal{V}} e^{U_{w'}^T V_c}}$$

Donde los parametros $\theta = \{W, W'\}$, con W, la matrix de codificación de c y W' la matrix de codificación de w.

Skip-gram - Pipeline



$N = |\mathcal{V}|$

Skip-gram - Optimización

Función costo

$$J(\theta) = -\sum_{(c,w)\in\mathscr{D}} \log p(w|c,\theta)$$

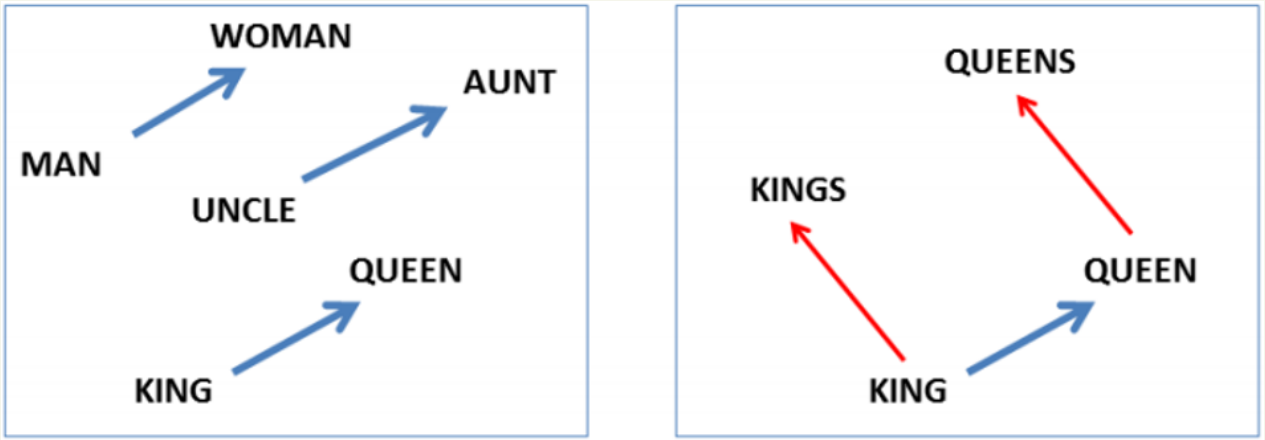
$\mathscr{D}$ es el dataset de pares (palabra central, contexto) extraídos del corpus.

Gradientes

$$\frac{\partial J(\theta)}{\partial V_c} = \sum_{(c,w)\in\mathscr{D}} \left([\sum_{w'\in\mathscr{V}} p(w'|c,\theta)U_{w'}] - U_w\right)$$

$$\frac{\partial J(\theta)}{\partial U_k} = \sum_{(c,w)\in\mathscr{D}} \left(p(k|c,\theta) - \mathbf{1}_{k=w}\right) V_c$$

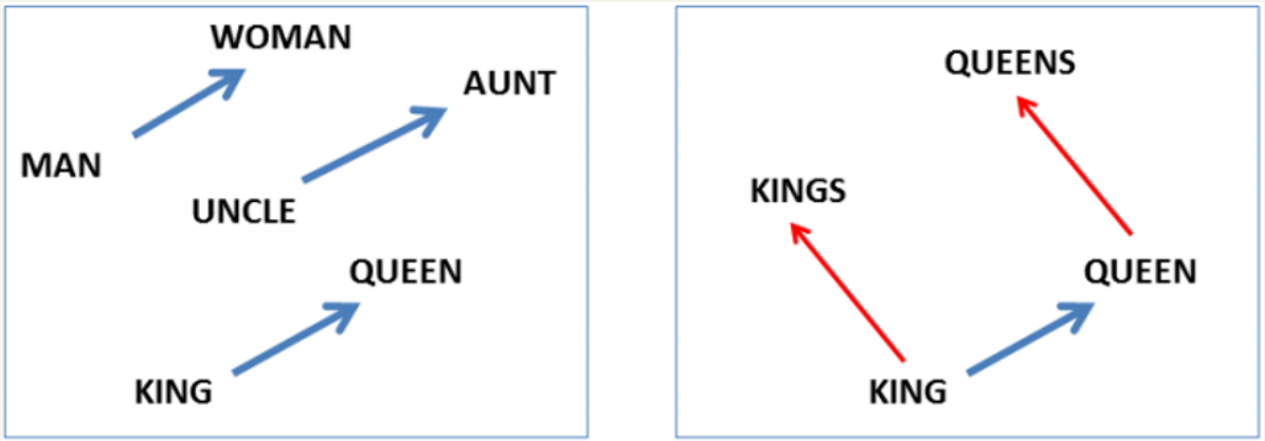
Word Vectors + PCA



(Mikolov et al., NAACL HLT, 2013)

vector(KINGS) - vector(KING) + vector(QUEEN) = vector(QUEENS)

Word Vectors + PCA



(Mikolov et al., NAACL HLT, 2013)

vector(KINGS) - vector(KING) + vector(QUEEN) = vector(QUEENS)

Similitud Coseno

El significado suele relacionarse con la dirección de los vectores, y por lo tanto, el ángulo entre vectores indica que tan similares son dos representaciones. La similitud coseno se define como el coseno del ángulo entre dos vectores $\mathbf{SC}(u, v) = \frac{u \cdot v}{\|u\| \|v\|}$.

Procesamiento del Lenguaje Natural

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)

Vectorizaciones más Sofisticadas

En la práctica suelen utilizarse representaciones pre-entrenadas.

- **GloVe** (2014): estadísticas globales de co-ocurrencia en el corpus.
- **FastText** (2016) (n-grams, capturan información morfológica)
- **BERT** (2018): Transformers. Represtaciones contextuales.
- **GTE** (2024): Transformer. Recuperación semántica..

Procesamiento del Lenguaje Natural

Normalizaciones de NLP

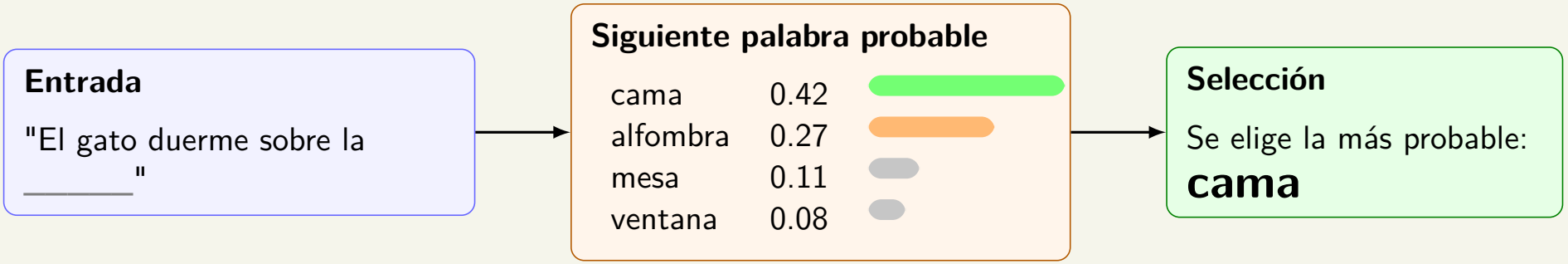
- Eliminar caracteres raros e inusuales
- Convertir todo a minúsculas
- Eliminar palabras no informativas (stop words)
- Descartar las palabras poco observadas
- Descartar las palabras más comunes
- Lemmatization (significado)
- Stemming (quedarse con la raíz)

Vectorizaciones más Sofisticadas

En la práctica suelen utilizarse representaciones pre-entrenadas.

- **GloVe** (2014): estadísticas globales de co-ocurrencia en el corpus.
- **FastText** (2016) (n-grams, capturan información morfológica)
- **BERT** (2018): Transformers. Represtaciones contextuales.
- **GTE** (2024): Transformer. Recuperación semántica..

Síntesis de texto



Síntesis de texto

Entrada

"El gato duerme sobre la _____"

Siguiente palabra probable

cama	0.42
alfombra	0.27
mesa	0.11
ventana	0.08

Selección

Se elige la más probable: **cama**

Nueva entrada

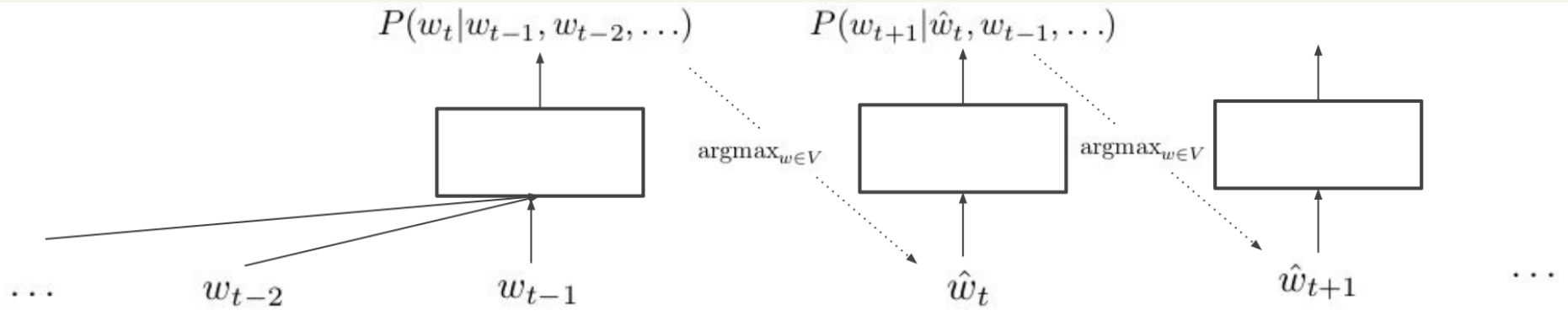
"El gato duerme sobre la cama _____"

Se vuelve a estimar

mientras	0.31
y	0.22
en	0.18
.	0.12

Selección

Se elige la más probable: **mientras**



Outline

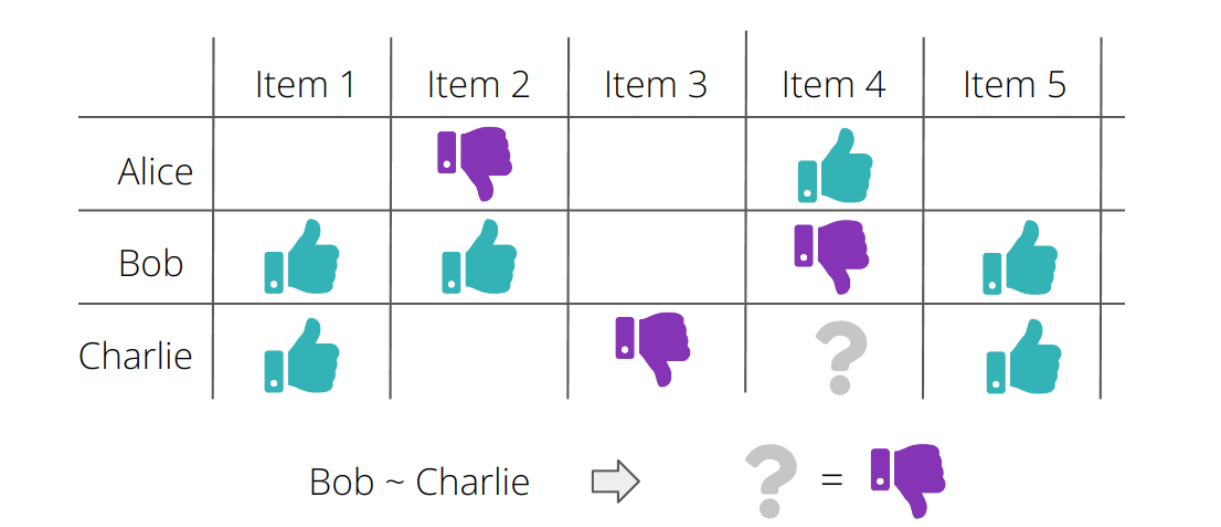
Sistemas de Recomendación

Algunas 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.
- *Manipulaciones*. Los algoritmos no solo recomiendan según los gustos del usuario, sino que priorizan algunos contenidos por sobre otros. Pero no todos detallan los criterios utilizados para ellos.

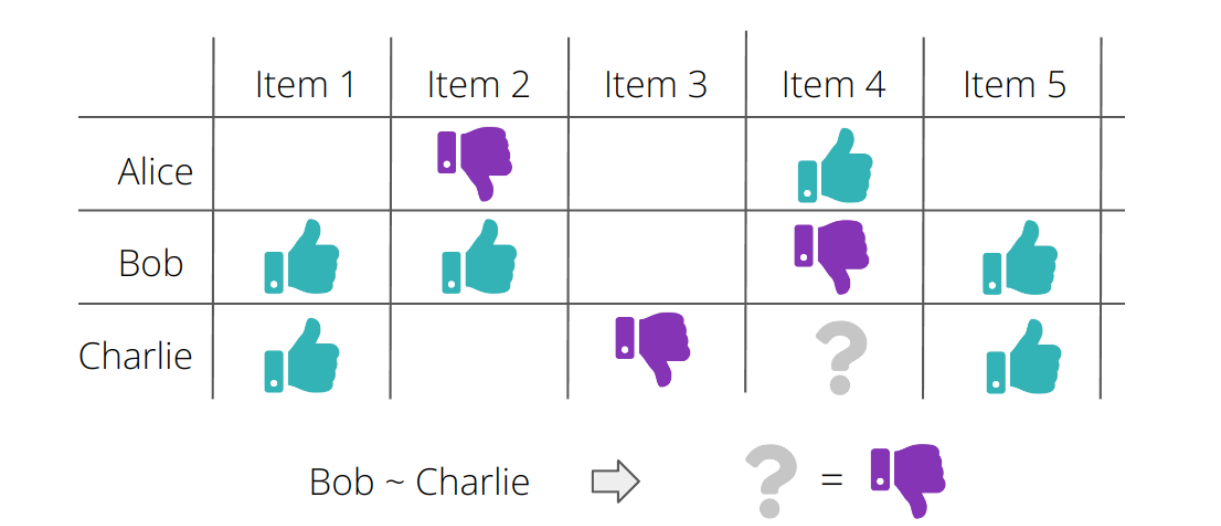
Filtro Colaborativo

Aprender por Colaboración



Filtro Colaborativo

Aprender por Colaboración



Entrenamiento

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

donde $\mathscr{R} = \{(i,j) : y_{i,j} \text{ tiene dato cargado}\}$, $y \in \mathbb{N}^{n_{\text{items}} \times n_{\text{users}}}$ 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; ν la dimensión del espacio latente y $\lambda \geq 0$ un hiperparámetro de regularización.

Filtro Colaborativo

Combinación convexa de factores durante la inferencia

Inferencia (Rating)

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

donde Ψ_i es la calificación promedio del item i-ésimo (dentro de los datos cargados) y $0 \leq p \leq 1$ es un hiperparámetro que indica cuanto peso le damos al aprendizaje y cuanto al valor medio.

Filtro Colaborativo

Combinación convexa de factores durante la inferencia

Inferencia (Rating)

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

donde Ψ_i es la calificación promedio del item i-ésimo (dentro de los datos cargados) y $0 \leq p \leq 1$ es un hiperparámetro que indica cuanto peso le damos al aprendizaje y cuanto al valor medio.

TECH / ELON MUSK

Yes, Elon Musk created a special system for showing you all his tweets first

/ After his Super Bowl tweet did worse numbers than President Biden's, Twitter's CEO ordered major changes to the algorithm.

theverge.com

by Zoë Schiffer and Casey Newton
Feb 14, 2023, 10:19 PM GMT-3

219 Comments (219 New)

Photo illustration by William Joel / The Verge, photo by Christian Marquardt / Getty Images