

Día 08 · Tema 3 — Ejercicios: embeddings y distancia coseno

Cómo se usa esta hoja. Hay dos tipos de ejercicio, y sólo dos.

El ejercicio del día. Uno por hoja. Es el obligatorio y es el que entra en el examen, con el cálculo incluido. Basta con hacer ese para seguir el curso con solvencia.

Los ejercicios de ampliación. El resto. Están para enseñar de qué es capaz el método, no para exigir: **no se piden calculados en el examen**. De ellos sólo puede preguntarse la idea, a nivel conceptual.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

1. Similitud coseno paso a paso

Dados los vectores $\mathbf{u} = (1, 2, 3)$ y $\mathbf{v} = (4, 5, 6)$:

- Calcula el producto escalar $\mathbf{u} \cdot \mathbf{v}$.
- Calcula las normas $\|\mathbf{u}\|$ y $\|\mathbf{v}\|$.
- Calcula la similitud coseno $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$.
- Interpreta el resultado: ¿están estos vectores muy alineados o poco?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

2. Coseno frente a distancia euclídea

Considera tres vectores en $\mathbb{R}^3$: $\mathbf{a} = (1, 0, 0)$, $\mathbf{b} = (2, 0, 0)$, $\mathbf{c} = (0, 1, 0)$.

- Calcula $\cos(\mathbf{a}, \mathbf{b})$ y $\|\mathbf{a} - \mathbf{b}\|$.
- Calcula $\cos(\mathbf{a}, \mathbf{c})$ y $\|\mathbf{a} - \mathbf{c}\|$.
- ¿Qué métrica captura la *dirección* y cuál la *magnitud*? ¿En cuál de los dos pares la similitud coseno da un resultado más extremo que la distancia euclídea?

EL EJERCICIO DEL DÍA *obligatorio · entra en el examen*

3. Aritmética de Word2Vec

Se dan los siguientes embeddings de palabras en $\mathbb{R}^3$:

$$\varphi(\text{España}) = (0,8, 0,3, 0,6), \quad \varphi(\text{Madrid}) = (0,9, 0,5, 0,7), \quad \varphi(\text{Francia}) = (0,7, 0,2, 0,8).$$

- Usando la analogía $\varphi(\text{París}) \approx \varphi(\text{Madrid}) - \varphi(\text{España}) + \varphi(\text{Francia})$, calcula el vector predicho para París.
- Se proponen tres candidatos: $\mathbf{c}_1 = (0,8, 0,4, 0,9)$, $\mathbf{c}_2 = (0,5, 0,1, 0,3)$, $\mathbf{c}_3 = (0,9, 0,9, 0,9)$. Calcula la similitud coseno entre el vector predicho y cada candidato. ¿Cuál es el más cercano?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

4. Pérdida de negative sampling

En el modelo skip-gram con negative sampling, la función de pérdida para una palabra objetivo w_t con un contexto positivo w_c y una muestra negativa w_n es:

$$\mathcal{L} = \log \sigma(\varphi(w_c) \cdot \varphi(w_t)) + \log \sigma(-\varphi(w_n) \cdot \varphi(w_t)),$$

donde $\sigma(x) = 1/(1 + e^{-x})$ es la función sigmoide.

Datos: $\varphi(w_t) = (1, 0)$, $\varphi(w_c) = (0,5, 0,5)$, $\varphi(w_n) = (-0,3, 0,8)$.

Valores aproximados de la sigmoide: $\sigma(0,5) \approx 0,622$, $\sigma(-0,5) \approx 0,378$, $\sigma(0,3) \approx 0,574$, $\sigma(-0,3) \approx 0,426$.

- Calcula los productos escalares $\varphi(w_c) \cdot \varphi(w_t)$ y $\varphi(w_n) \cdot \varphi(w_t)$.
- Evalúa $\mathcal{L}$ usando los valores aproximados anteriores.

- (c) Interpreta el signo y la magnitud: ¿la pérdida indica un buen o mal estado del modelo?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

5. Ranking en búsqueda semántica

Un sistema de búsqueda semántica codifica una consulta como $\mathbf{q} = (1, 0, 1)$ y tiene cuatro documentos indexados:

$$\mathbf{d}_1 = (1, 1, 0), \quad \mathbf{d}_2 = (0, 0, 1), \quad \mathbf{d}_3 = (1, 0, 1), \quad \mathbf{d}_4 = (0, 5, 0, 5).$$

- Calcula $\cos(\mathbf{q}, \mathbf{d}_j)$ para $j = 1, 2, 3, 4$.
- Ordena los documentos de mayor a menor similitud.
- Si el sistema devuelve el *top-2*, ¿qué documentos se recuperan?
- ¿Habría cambiado el ranking si se usara distancia euclídea en lugar de similitud coseno? Justifica brevemente.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

6. Compromiso dimensionalidad–almacenamiento

Un corpus tiene un vocabulario de $|V| = 50\,000$ palabras. La matriz de embeddings tiene tamaño $|V| \times d$, donde cada parámetro se almacena en FP32 (4 bytes).

- Calcula el número de parámetros y el almacenamiento en MB para $d = 100$, $d = 300$ y $d = 768$.
- Completar la tabla siguiente:

	$d = 100$	$d = 300$	$d = 768$
Parámetros (millones)			
Almacenamiento (MB)			

- ¿Qué factores, más allá del almacenamiento, afectan a la elección de d en la práctica?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

7. Mean pooling y similitud de frases

Se tiene una frase de tres tokens con los siguientes embeddings:

$$\mathbf{e}_1 = (1, 2, 0), \quad \mathbf{e}_2 = (0, 1, 3), \quad \mathbf{e}_3 = (2, 0, 1).$$

- Calcula el embedding de la frase por *mean pooling*: $\bar{\mathbf{e}} = \frac{1}{3}(\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)$.
- Dado otro embedding de frase $\mathbf{s} = (1, 1, 1)$, calcula $\cos(\bar{\mathbf{e}}, \mathbf{s})$.
- ¿Qué información se pierde con mean pooling respecto a un modelo que genera embeddings contextuales (p. ej. BERT)?