Bienvenidos!
Semana 4

Datos No Estructurados y Semiestructurados

Especialización en Economía, opción Ciencia de Datos

FCS - Udelar

Guillermo Lezama

La clase de hoy

- Embeddings!
- Casi que la frontera...

- Cuestionarios
- Tareas
- Videos
- Emails
- EVA

- ¿Preguntas?
- Clase de ayer

¿Qué son los embeddings?

- Representaciones numéricas de texto en forma de vectores.
- Capturan significado semántico (sentido/contexto), no solo frecuencia.
- Permiten medir la cercanía o distancia semántica entre textos.
- No fueron entrenados para detector polaridad sentimental

Pero… ¿de dónde salen?

- Modelos de lenguaje pre-entrenados con millones de textos.
- Redes neuronales profundas (Deep Learning) capturan significado contextual.
- Ejemplo usado en clase: "text-embedding-3-small" de OpenAI.
- Resultado: vector numérico (por ejemplo, de dimensión 1536).

¿Para qué se usan?

- Búsqueda (los resultados son rankeados por similitud)
- Clusterizar (agrupar)
- Recomendaciones
- Detección de anomalidades
- Medir diversidad
- Clasificación en grupos

¿Cómo los obtenemos?

- Hay versiones gratis
- Hay versiones pagas (por ejemplo: OpenAI) (1 centavo de dolar cada 1 millon de tokens)

Dimensionalidad

- Son vectores con 1536 entradas.
- Puede tener sentido reducir las dimensiones
  - Consume muchos recursos, memoria, etc.
  - Ruido

Ingeniería de Prompts

- Diseño cuidadoso y estratégico de instrucciones (*prompts*) para modelos de lenguaje como GPT
- Facilita transformar texto libre en información estructurada y utilizable.
- Mejora precisión en tareas de clasificación, resumen, extracción de entidades y análisis de sentimientos.

Para próxima clase

- Esta semana: un cuestionario cada **2 clases**
- Tarea del notebook
- Presentación del lunes