

Práctica: word2vec / skip-gram

En esta práctica trabajaremos con **modelos de representación vectorial de palabras** (word embeddings) mediante la técnica **Skip-gram** de *word2vec*. La idea central es que **el significado de una palabra puede inferirse por las palabras que la rodean**, como hemos visto en clase. Para ello, entrenaremos una red neuronal muy sencilla que, dado un **término central**, aprenda a **predecir palabras de su contexto** dentro de una ventana fija.

Partiremos de un corpus en texto plano (`dataset_word2vec.txt`) con frases cortas y vocabulario controlado. A partir de este fichero, deberás **tokenizar** las oraciones, **generar pares (centro, contexto)** con una ventana deslizante y **entrenar** un modelo Skip-gram.

Al finalizar, explora la calidad de los embeddings obtenidos mediante **consultas de similitud** (vecinos más cercanos) y **pruebas de analogías** (“parís : francia :: madrid : ?”), observando cómo los hiperparámetros del modelo influyen en los resultados.

Entregables

1. **Notebook** con el código (bien comentado).
2. Tabla/captura con **vecinos** de 5–10 palabras (p. ej., perro, gata, coche, parís, francia).
3. Resultados de **analogías** (p. ej., parís : francia :: madrid : ?).
4. Breve **comentario** sobre qué hiperparámetros influyen más y por qué.