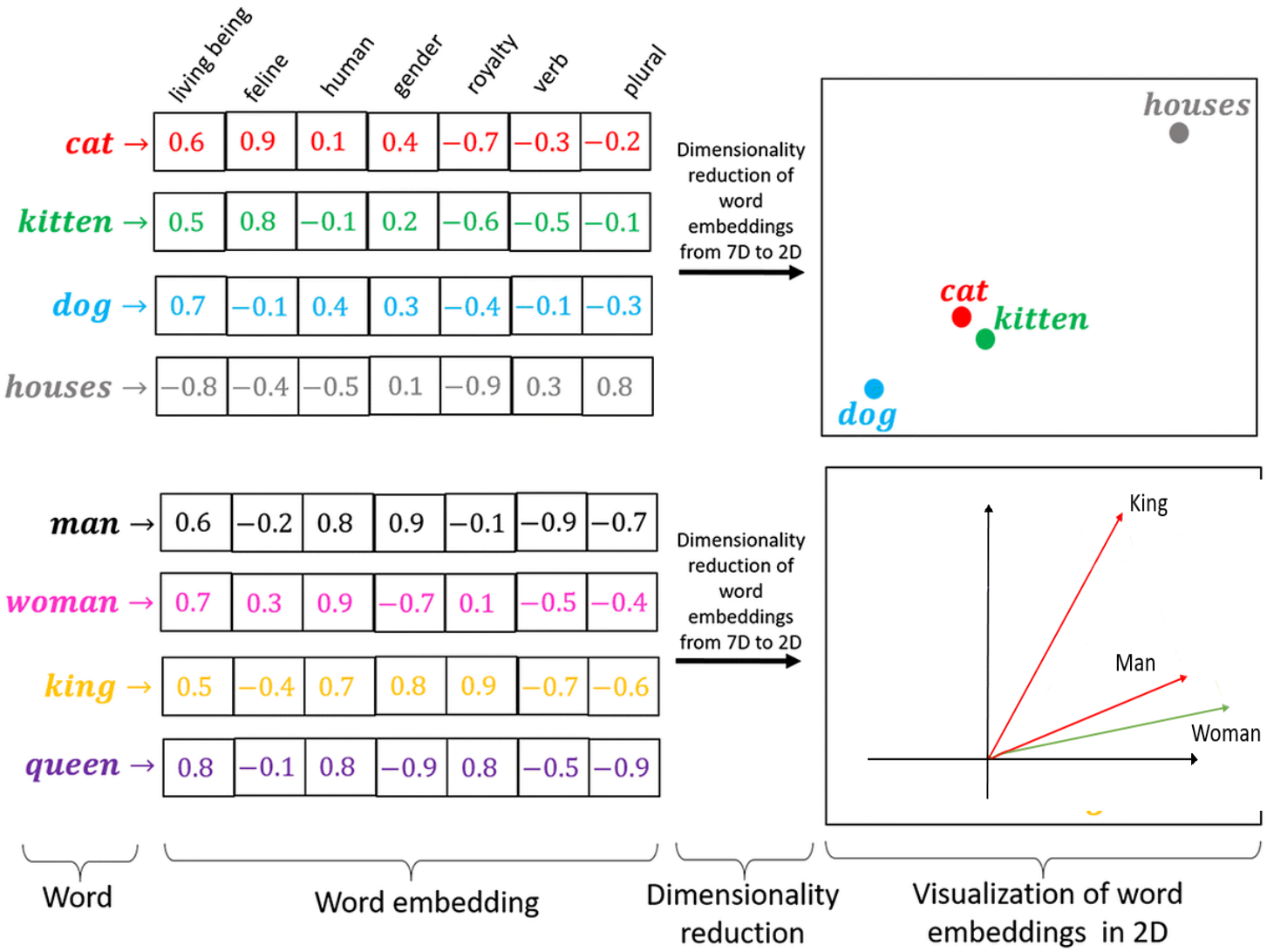
Word Embeddings

Santiago López

¿Qué son?

- Los word embeddings son representaciones vectoriales de palabras.
- Capturan relaciones semánticas entre palabras.
- Palabras similares tienen vectores cercanos en el espacio.



¿Por qué se utilizan?

- Las computadoras no entienden texto, solo números.
- Los embeddings convierten palabras en vectores numéricos.
- Permiten a los modelos de machine learning procesar y entender el lenguaje natural.

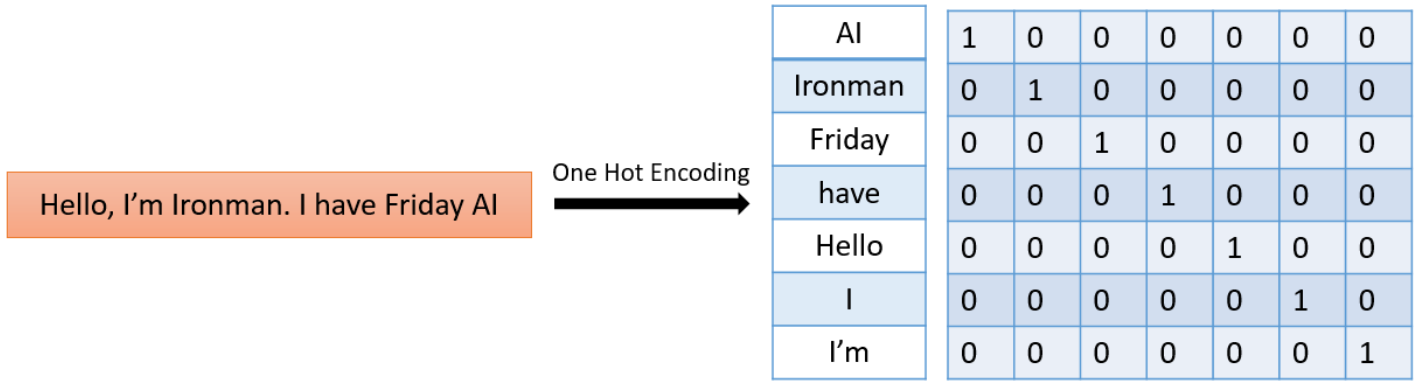
One Hot Encoding

- Representación tradicional: cada palabra es un vector disperso con un 1 en la posición correspondiente y 0 en las demás.
- Problemas:

-Alta dimensionalidad.

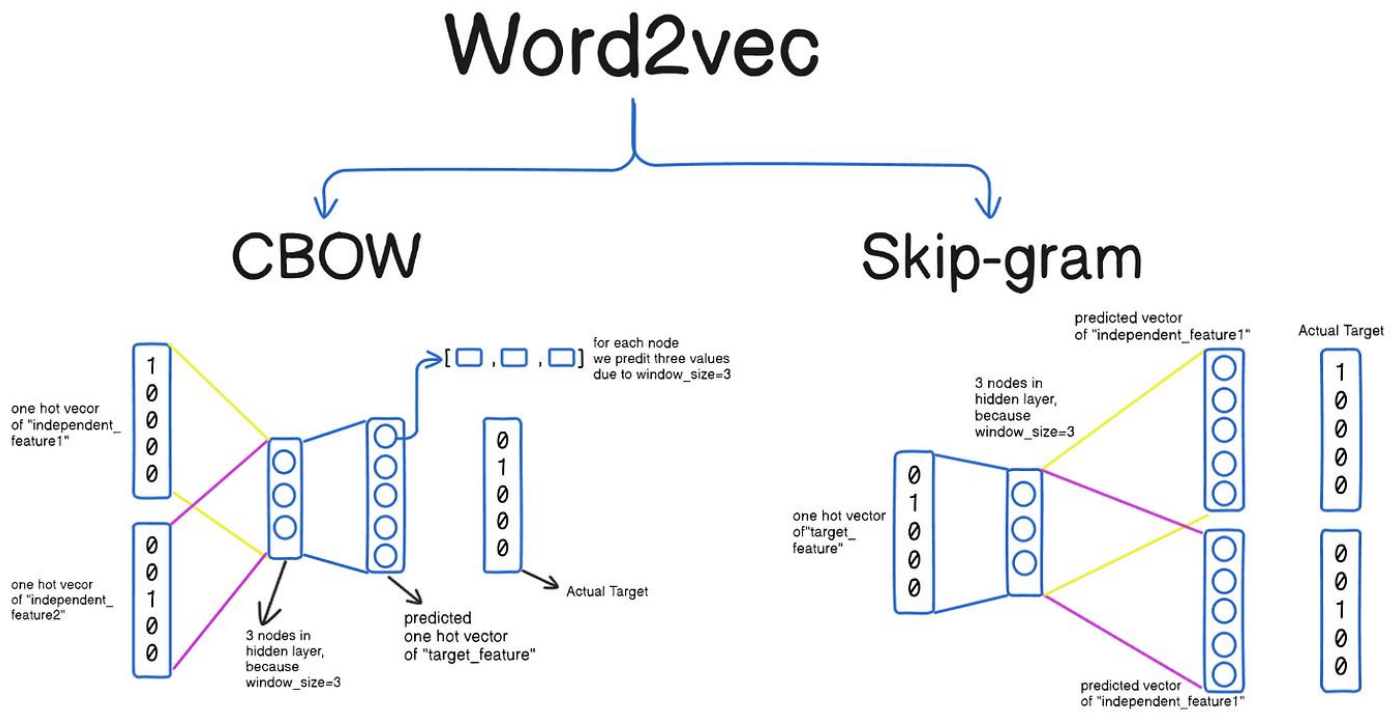
-No captura relaciones semánticas entre palabras.

-Ineficiente para textos grandes.



Word2Vec

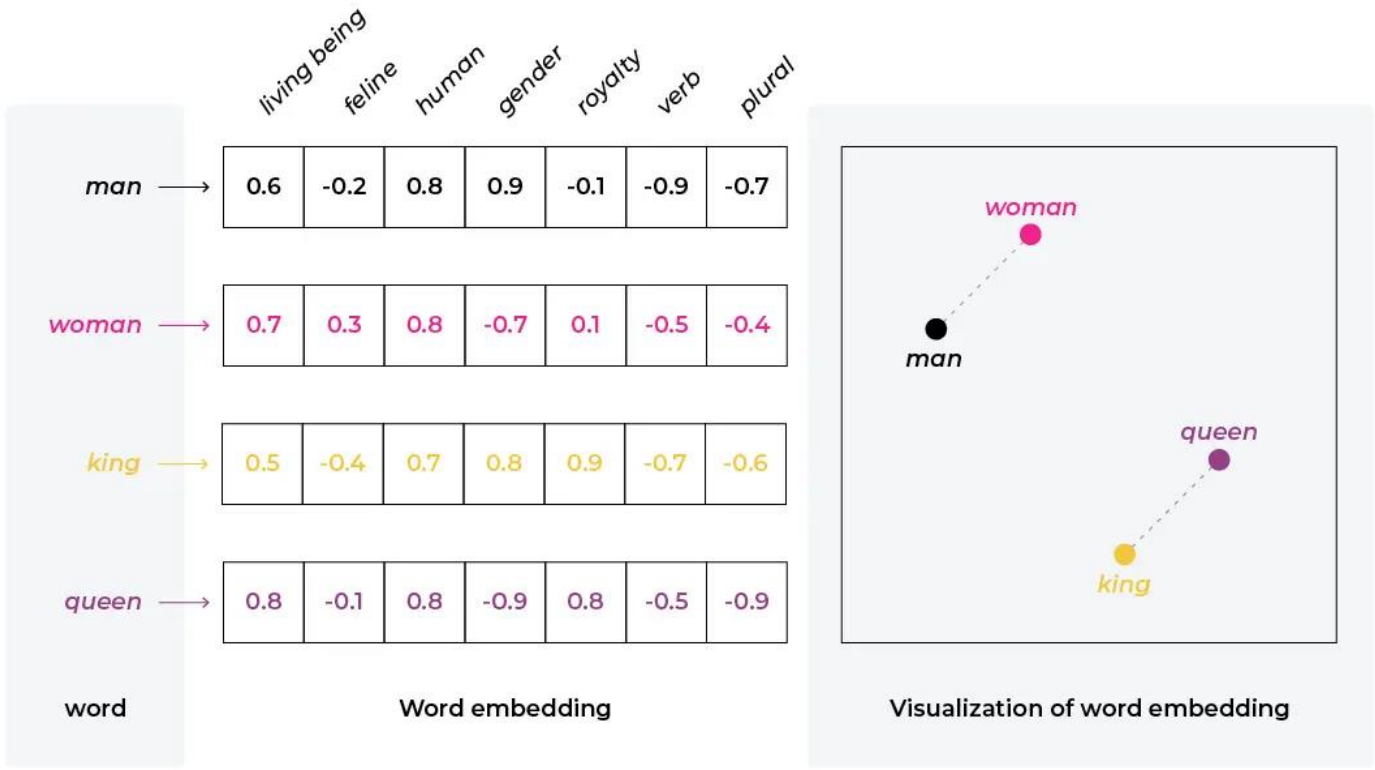
- Modelo que aprende embeddings usando redes neuronales.
- Dos métodos de entrenamiento: Skip-gram y CBOW (Continuous Bag of Words)
- Eficiente y captura relaciones semánticas.



Propiedades y beneficios

- Captura relaciones semánticas.
- Dimensionalidad:

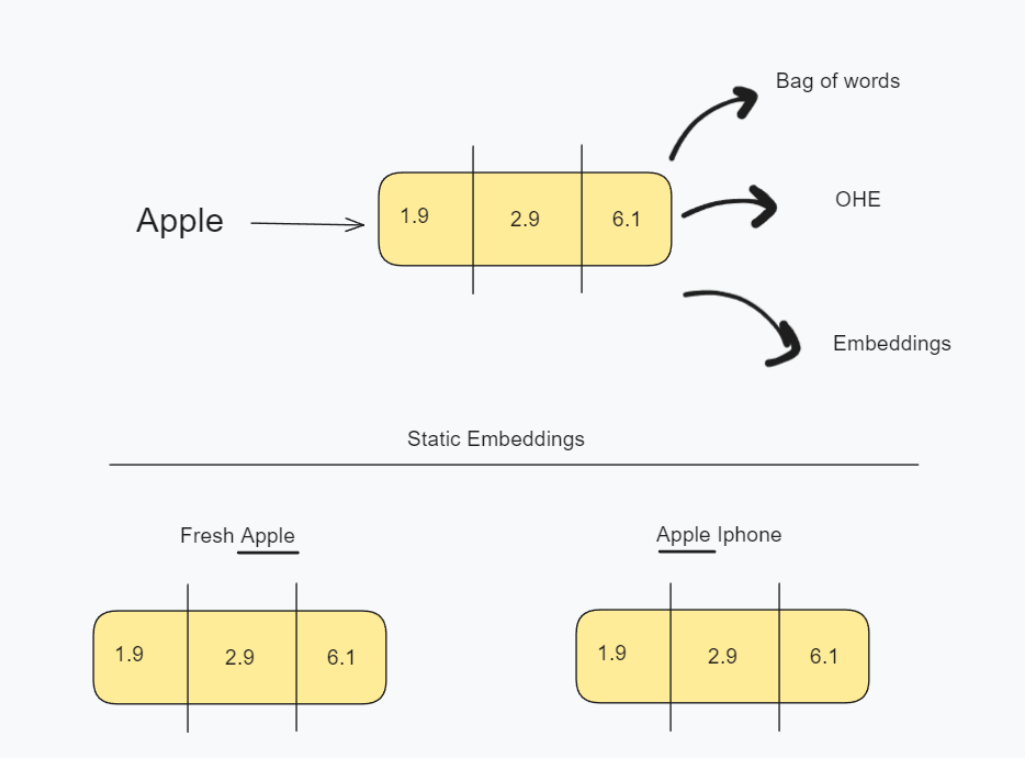
- Vectores típicamente de 50 a 400 dimensiones.

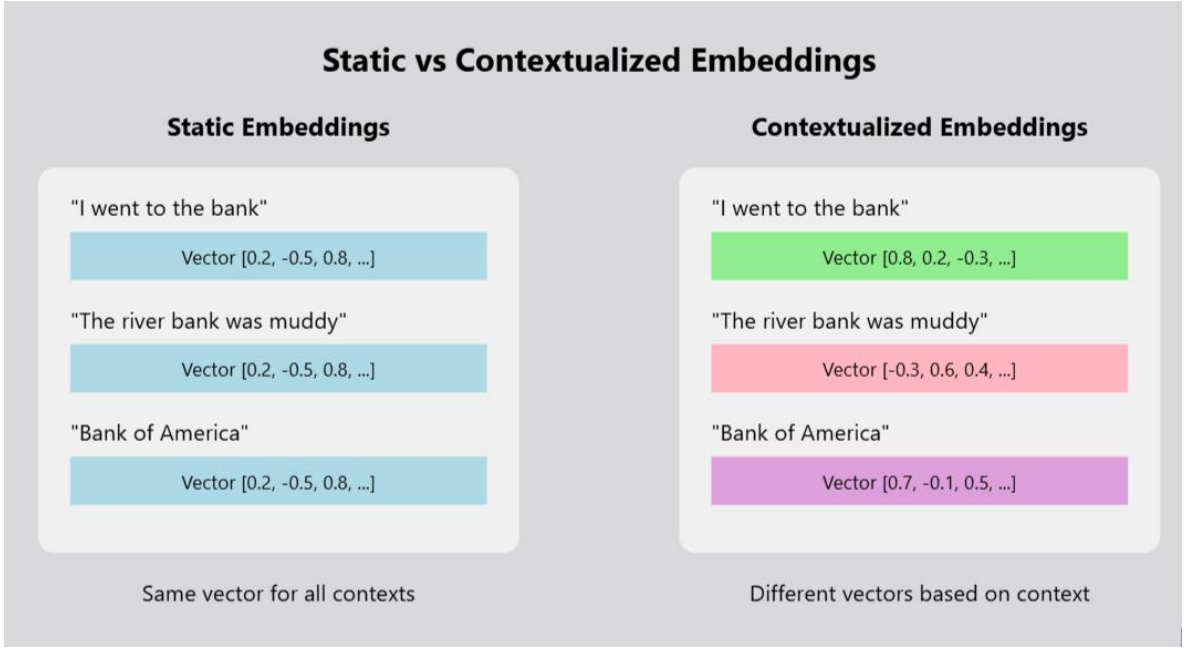


Embeddings con contexto

- Los embeddings dependen del contexto de la palabra (ej: Bank)
- Modelos como BERT y GPT generan representaciones dinámicas.
- La misma palabra tiene diferentes embeddings en diferentes oraciones.

Embeddings con contexto





Entrenamiento embeddings con contexto

Arquitectura basada en Transformers (autoatención).

Pre-entrenamiento:

MLM (Masked Language Modeling): Predecir palabras enmascaradas en una oración.

CLM (Causal Language Modeling): Predecir la siguiente palabra en una secuencia.

Fine-tuning:

Añadir una capa adicional para la tarea (e.g., clasificación).

Entrenar con datos etiquetados (e.g., análisis de sentimientos).

Algunos usos

- Sentiment Analysis
- Traducción automática: Representar palabras en diferentes idiomas.
- Búsqueda de información: Encontrar documentos relevantes.
- Reconocimiento de entidades: Identificar nombres, lugares, etc.