UNAM iimas INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

Word embeddings

Materia: Procesamiento de lenguaje natural

Blanca Vázquez

Recordemos

Entrada → Modelo → Salida

¿Cómo entrar un modelo a partir de un conjunto de textos?

¿Cómo entrar un modelo a partir de un conjunto de textos?

Transformar los textos a una forma vectorial

Bolsa de palabras

Una bolsa de palabras (BOW, por las siglas en inglés de *Bag of Words*) es un modelo que representa texto como un conjunto no ordenado de palabras, ignorando la gramática y el orden de las palabras.

	cat	cute	dog	small
small dog	0	0	1	1
cute cat	1	1	0	0
cute dog	0	1	1	0

TF-IDF (Term Frequency — Inverse Document Frequency)

Es una técnica estadística que cuantifica la importancia de una palabra en un documento en función de la frecuencia con la que aparece en ese documento y en una colección determinada de documentos (corpus).

$$tfidf(t,d,D) = tf(t,d) \times idf(d,D)$$

TF-IDF ejemplo

Supongamos que contamos con 4 documentos:

- d1: *“The sky is blue.”*
- d2: "*The sun is bright today.*"
- d3: "*The sun in the sky is bright.*"
- d4: "*We can see the shining sun, the bright sun.*"

Paso 1: Remover las stop-words:

- d1: *“sky blue.”*
- d2: " *sun bright today.*"
- d3: "*sun sky bright.*"
- d4: *“can see shining sun bright sun.*"

TF-IDF ejemplo

- d1: *“sky blue.”*
- d2: " *sun bright today."*
- d3: "*sun sky bright."*
- d4: *“can see shining sun bright sun."*

Paso 2 calcular TF (term- frequency): construir la matriz palabra - documento.

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0

TF-IDF ejemplo

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0

Paso 2 (continúa): construir la matriz documento - palabra y normalizar las filas que sumen 1.

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}}$$

donde $f_{t,d}$ es el número de ocurrencias de *t* en *d*.

TF-IDF ejemplo

- d1: *“sky blue.”*
- d2: " *sun bright today."*
- d3: "*sun sky bright."*
- d4: *“can see shining sun bright sun."*

Paso 3 calcular IDF (Inverse document- frequency): Encuentra el número de documentos en los que aparece cada palabra.

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0
n_t	1	3	1	1	1	2	3	1

TF-IDF ejemplo

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0
n_t	1	3	1	1	1	2	3	1

Paso 3 (continúa): calcula usando la fórmula $idf(t, D) = \log(\frac{N}{n_t})$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

$\log_{10} \frac{4}{1} = 0.602$

$\log_{10} \frac{4}{3} = 0.125$

TF-IDF ejemplo

Paso 4 calcular TF-IDF: multiplicar los puntajes obtenidos por separado

tf (t,d)

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

idf (t,D)

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

$$tfidf(t,d,D) = tf(t,d) \times idf(d,D)$$

TF-IDF ejemplo

Paso 4 calcular TF-IDF: multiplicar los puntajes obtenidos por separado

tf (t,d)

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

idf (t,D)

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

	blue	bright	can	see	shining	sky	sun	today
1	0.301	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	**0.201**
3	0	0.0417	0	0	0	0.1	0.0417	0
4	0	0.0209	**0.1**	**0.1**	**0.1**	0	0.0417	0

TF-IDF ejemplo

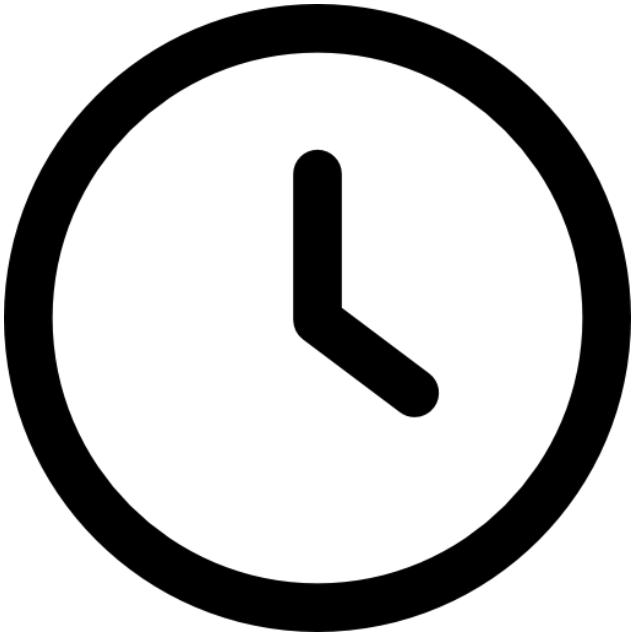
Paso 4 calcular TF-IDF: multiplicar los puntajes obtenidos por separado

	blue	bright	can	see	shining	sky	sun	today
1	**0.301**	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	**0.201**
3	0	0.0417	0	0	0	0.1	0.0417	0
4	0	0.0209	**0.1**	**0.1**	**0.1**	0	0.0417	0

TF-IDF se usa para ponderar la importancia de las palabras dentro de los documentos

¿Qué ventajas y desventajas encuentras en BOW y TF-IDF?

Ejercicio: TF-IDF



Word embeddings

Es un vector denso de valores de punto flotante los cuales son parámetros entrenables (pesos aprendidos por el modelo durante el entrenamiento).

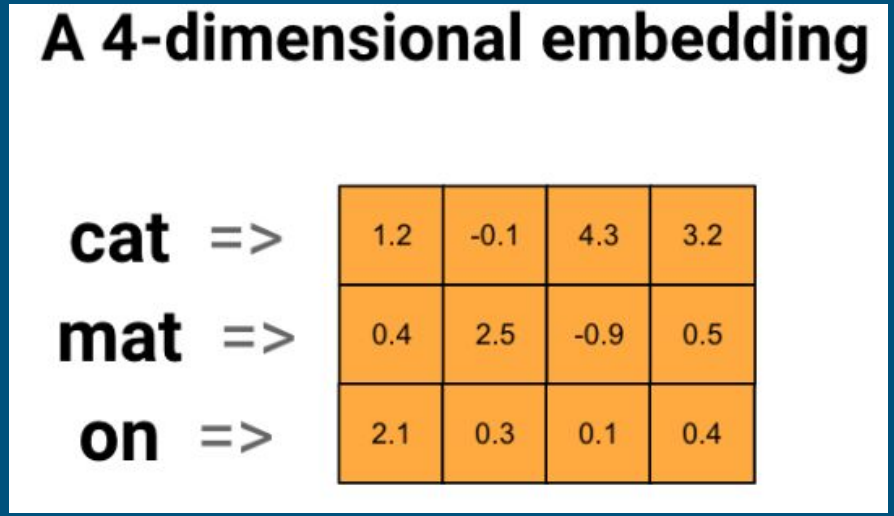


Imagen tomada de https://www.tensorflow.org/text/guide/word_embeddings

Word embeddings

La idea detrás de los embeddings es que a cada palabra se le asigna un vector en un **espacio multidimensional**.

La posición de estos vectores en el espacio refleja la proximidad semántica entre las palabras. Si dos palabras tienen significados similares, sus vectores estarán próximos. Si sus significados son opuestos o no están relacionados, estarán distantes en el espacio vectorial.

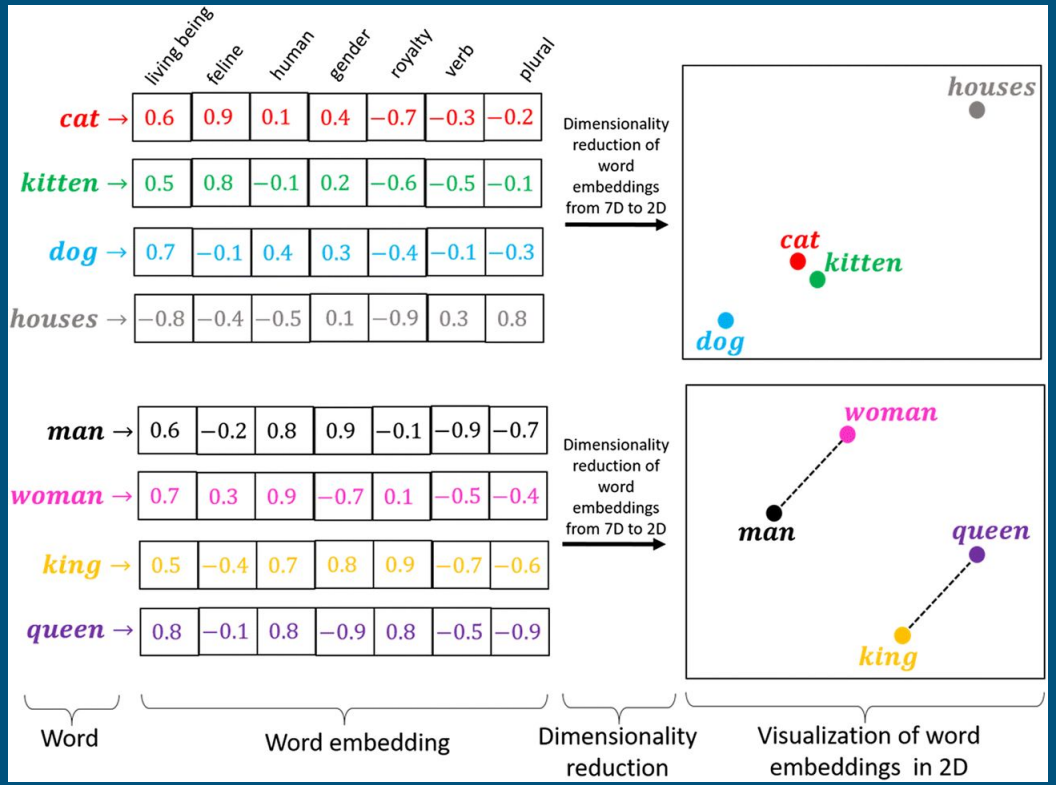


Imagen tomada de Hariom Gautam, 2020.

Time to code