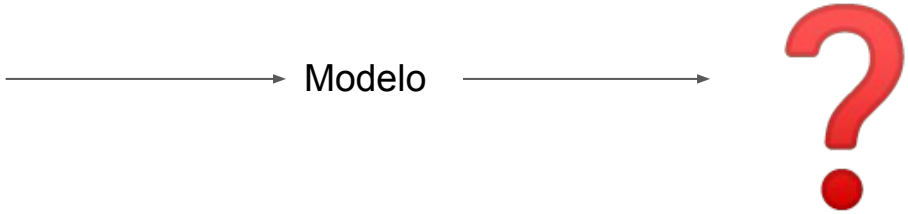
¿Qué es Natural Language Processing (NLP/PNL)?

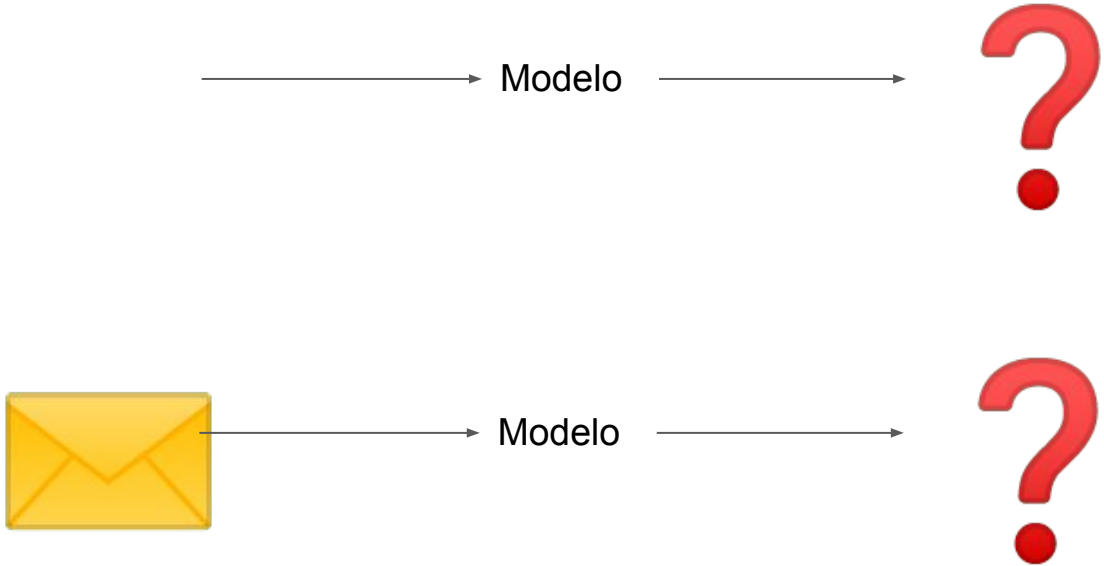
Clasificación de texto



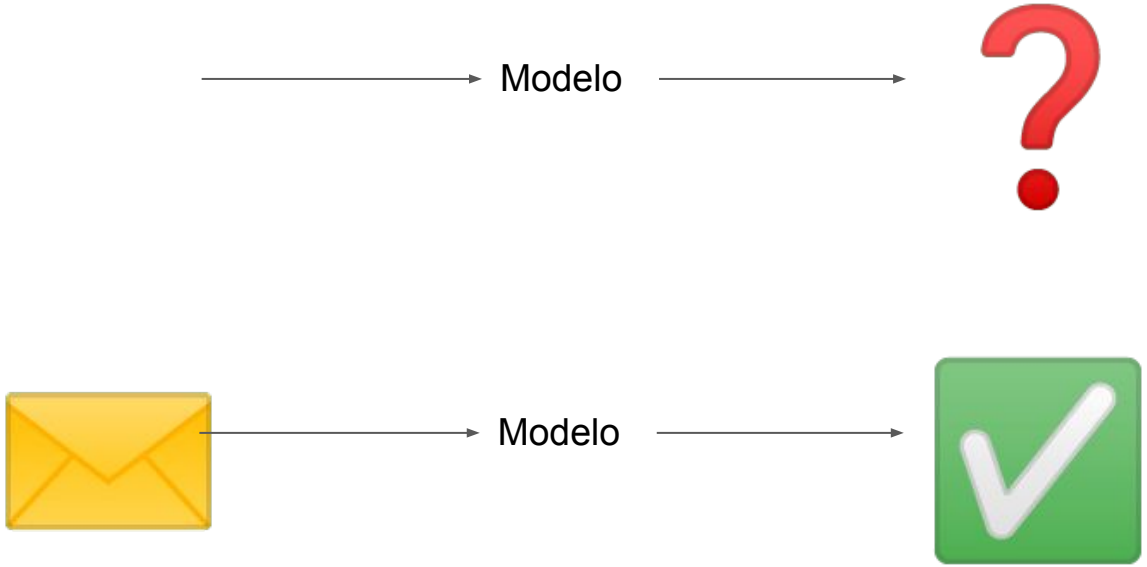
Clasificación de texto

Clasificación de texto

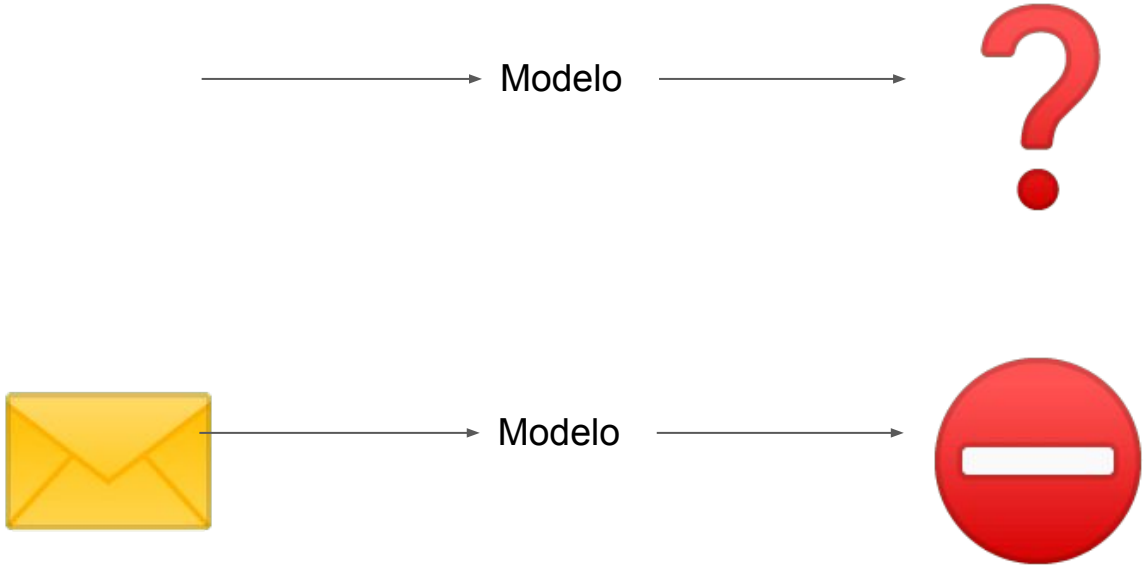
Clasificación de texto



Clasificación de texto

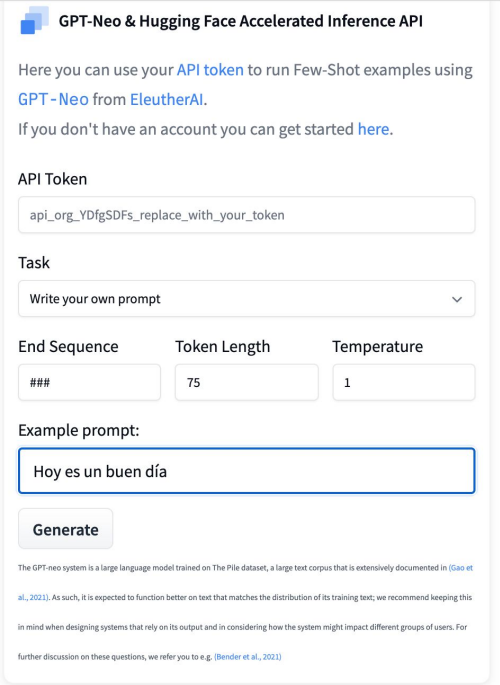


Clasificación de texto



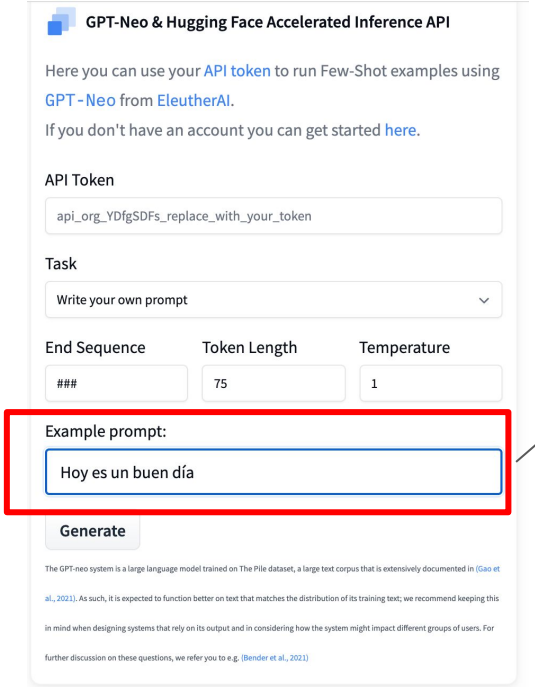
hf.co/blog/few-shot-learning-gpt-neo-and-inference-api

Generación de Texto



hf.co/blog/few-shot-learning-gpt-neo-and-inference-api

Generación de Texto



Example prompt:

Hoy es un buen día en San Juan, la capital del estado de Puerto Rico, donde las calles se han convertido en una de las tres grandes avenidas del mundo. El sábado, cuatro mil habitantes abandonan la ciudad, para irse a la isla de Navidad, a la cual el president

Y otros dominios

wav2vec: audio to text

Y otros dominios

wav2vec: audio to text

DALLE: text to image

https://openai.com/blog/dall-e/

Word Embeddings

Los word embeddings son representaciones del texto

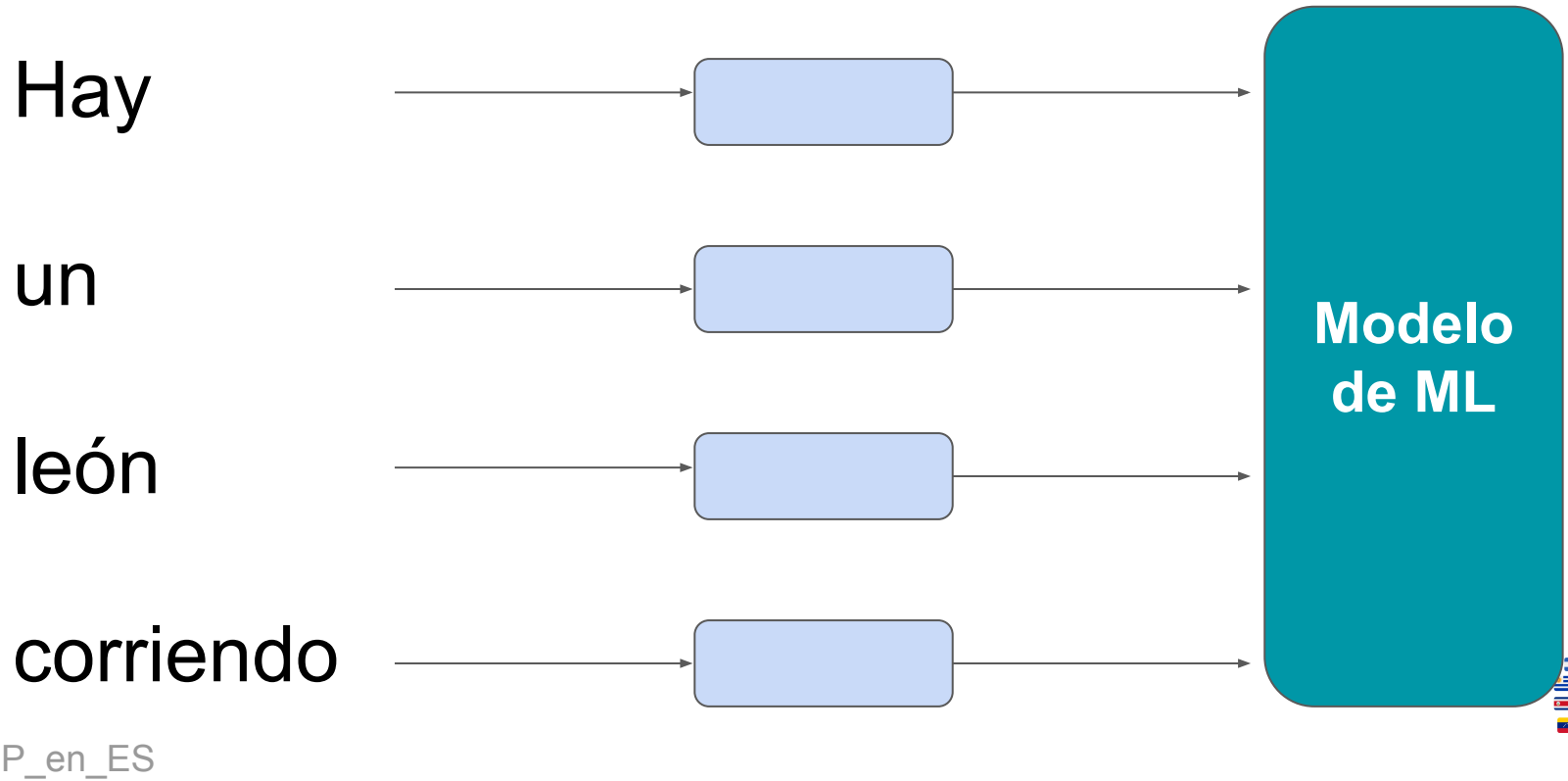
Hay

un

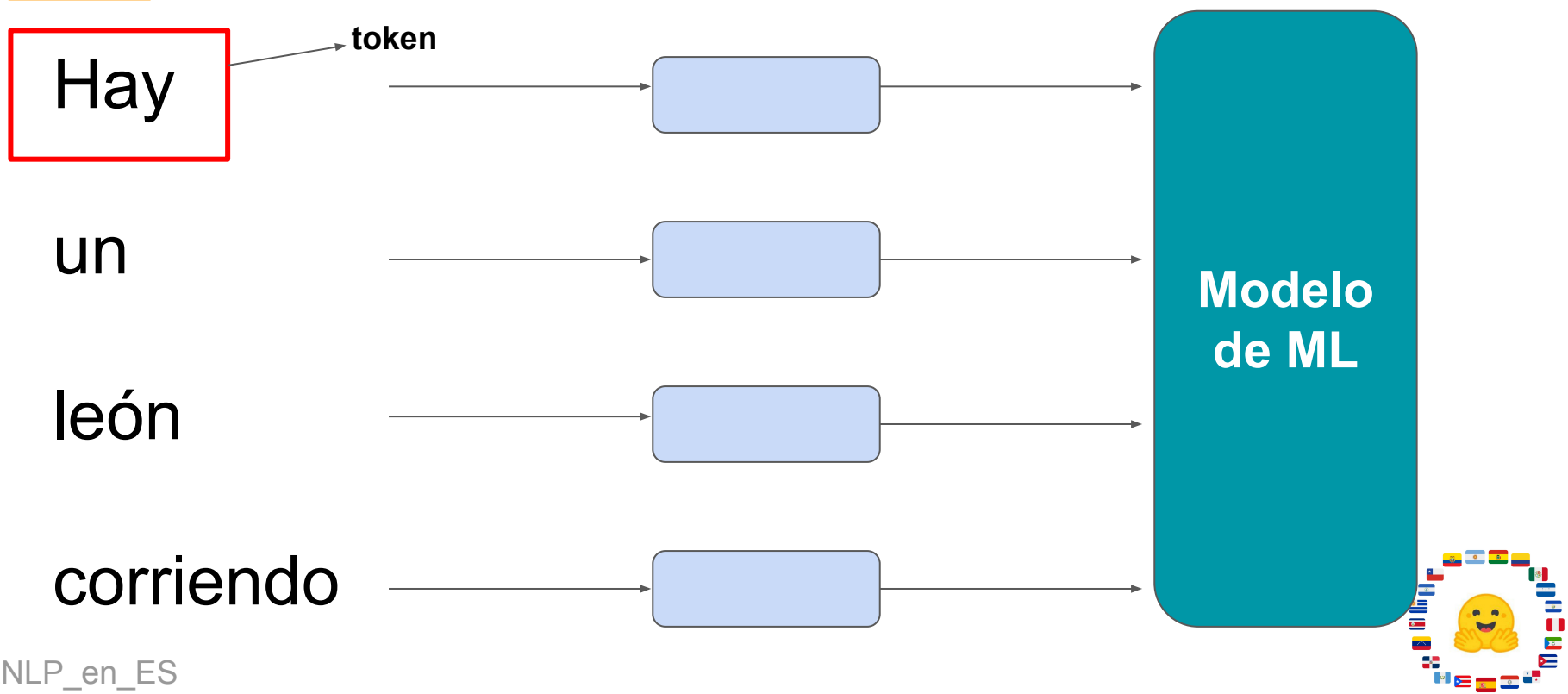
león

corriendo

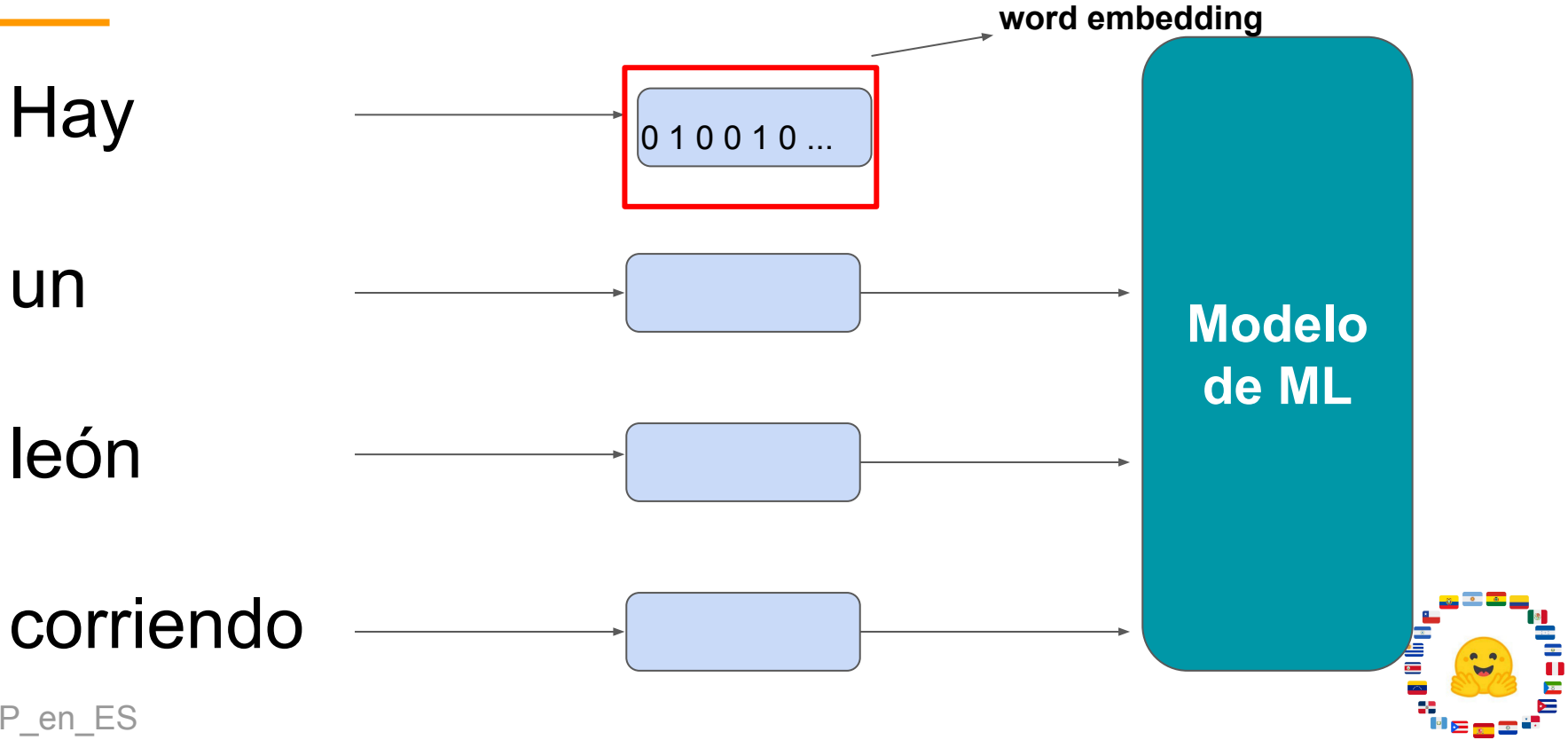
Los word embeddings son representaciones del texto



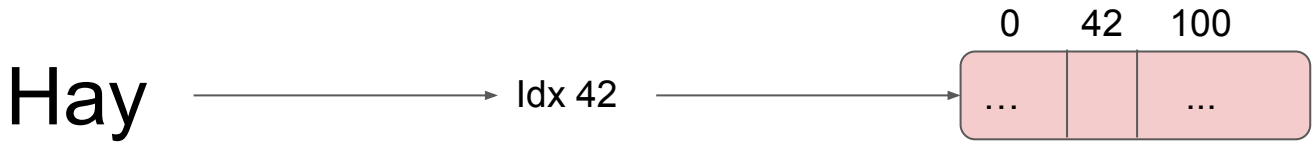
Los word embeddings son representaciones del texto



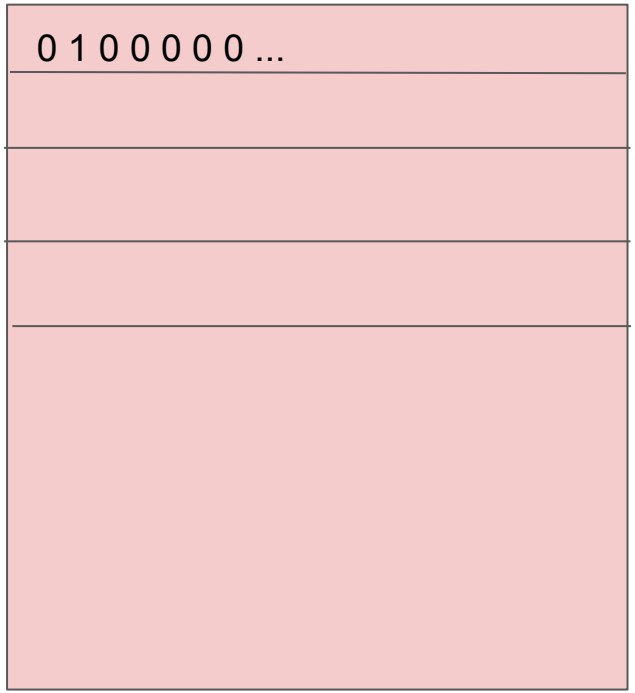
Los word embeddings son representaciones del texto



Se tiene una tabla que mapea cada palabra del vocabulario a la posición de su word embedding.



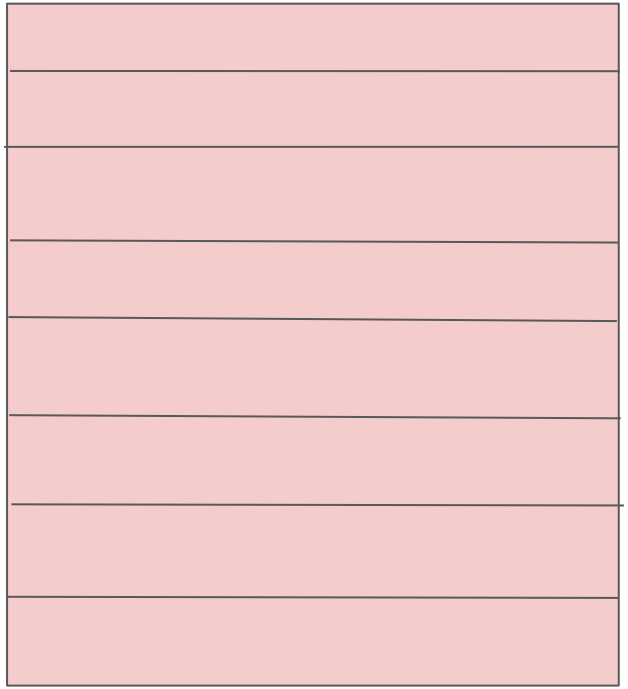
10000 word embeddings



representación de una palabra

256 números por palabra

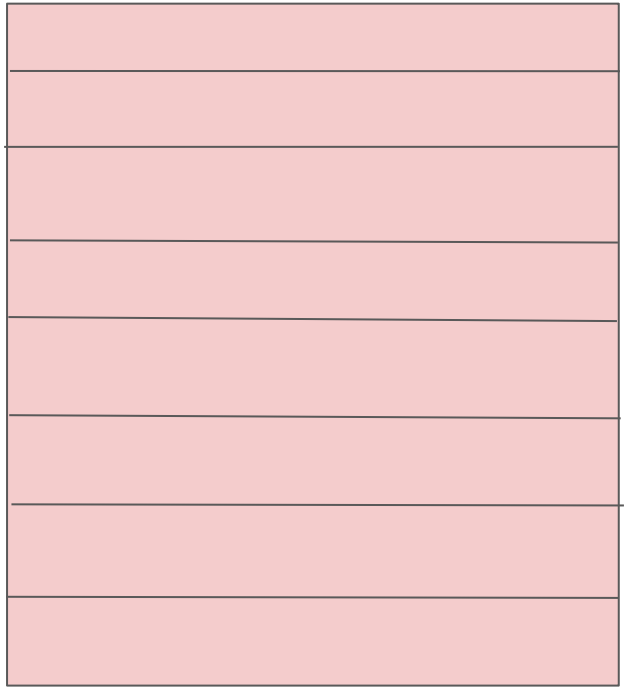
Incendio



El caracter especial UNK se suele utilizar para palabras fuera del vocabulario (out-of-vocabulary)

Incendio

UNK



One-hot Encoding

Vocabulario: Hola, Sol, Playa

One-hot Encoding

Sol [0, 1, 0, 0]
Playa [0, 0, 1, 0]
Hola [1, 0,0,0]
Feliz [0, 0, 0, 1]

Vocabulario: Hola, Sol, Playa

One-hot Encoding: Este es un ejemplo en el que el vocabulario sólo tiene las palabras de la oración, pero el vocabulario suele ser mucho más grande.

Ejemplo: Hoy es un buen día

```
[
  [0, 0, 0, 0,..., 1, 0],
  [1, 0, 0, ,....,0, 0, 0],
  [0,..., 1, 0, 0, 0, 0],
  [0, 0, 1, 0, …,0, 0],
  [0, 0, 0, 1, 0, …,0]
]
```

One-hot Encoding: Este es un ejemplo en el que el vocabulario sólo tiene las palabras de la oración, pero el vocabulario suele ser mucho más grande.

Ejemplo: Hoy es un buen día

```
[
  [0, 0, 0, 0, 1, 0],
  [1, 0, 0, 0, 0, 0],
  [0, 1, 0, 0, 0, 0],
  [0, 0, 1, 0, 0, 0],
  [0, 0, 0, 1, 0, 0]
]
```

¿Saben que es una **pachamanca**?

¿Saben que es una **pachamanca**?

El plato de **pachamanca** en la mesa.
Es demasiada **pachamanca** para mí
La **pachamanca** es típica en los Andes.

¿Saben que es una **pachamanca**?

El plato de **pachamanca** en la mesa.
Es demasiada **pachamanca** para mí
La **pachamanca** es típica en los Andes.

¿Saben que es una **pachamanca**?

Contexto

El plato de **pachamanca** en la mesa.
Es demasiada **pachamanca** para mí
La **pachamanca** es típica en los Andes.

¿Cómo supieron?

- Aunque nunca habían visto esa palabra antes, sí habían visto palabras en contextos similares.

El plato de **arroz** en la mesa.
Es demasiada **hamburguesa** para mí
La **quinoa** es típica en los Andes.

¿Cómo ponemos información de estos contextos en nuestras representaciones?

- Co-ocurrencias usando una ventana
- Información Mutua Puntual (PPMI)
- Y muchos métodos más

Word2vec

… Hoy es un gran día ...

Word2vec

Word2vec

… Hoy creo es un gran día para …

Ventana de 5

Variantes de Word2vec

- Skip-Gram: predice el contexto dada una palabra central

- Continuous Bag of Words (CBOW): predice la palabra central sumando los vectores de contexto

¿Cómo sabemos si el word embedding es bueno?

- Dadas las métricas al entrenarlo.

- Utilizándolo para entrenar un modelo en una tarea (por ejemplo, clasificación)

Propiedades - estructura linear

- Relaciones semánticas y sintácticas son lineares en el espacio vectorial

V(rey) - V(hombre) + V(mujer) = V(reina)

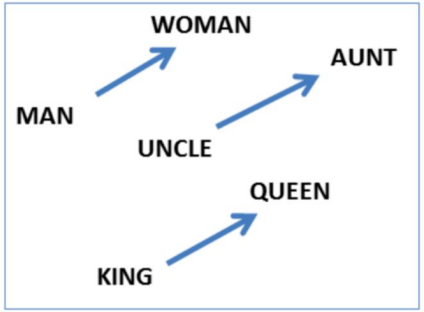
V(reyes) - v(rey) + V(reina) = V(reinas)

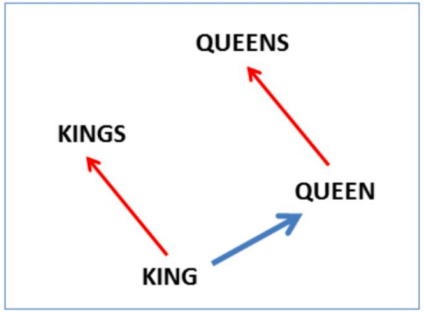
V(España) - V(Madrid) + V(Paris) = V(Francia)

V(fuego) - V(caliente) + V(frío) = V(hielo)

semantic: $v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen})$

syntactic: $v(\text{kings}) - v(\text{king}) + v(\text{queen}) \approx v(\text{queens})$





https://lena-voita.github.io/nlp_course/word_embeddings.html

Sesgos

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (2016), Bolukbasi et. al.

Enlaces útiles

@nlp-en-es/nlp-de-cero-a-cien

#nlp-de-cero-a-cien

@nlp_en_es

@company/nlp-en-es/