INTELIGENCIA ARTIFICIAL GENERATIVA

¿QUÉ ES EL PROCESAMIENTO DEL LENGUAJE NATURAL?

PLN es un subcampo de la Inteligencia Artificial que se ocupa de las interacciones entre los ordenadores y los lenguajes humanos (naturales).

- Se usa PNL para crear sistemas de IA que permitan:
  - Reconocimiento del habla,
  - Resumir documentos,
  - Traducción automática,
  - Detección de Spam,
  - Reconocimiento de Entidades Nombradas,
  - Respuesta a preguntas,
  - Autocompletar,
  - Escritura predictiva, etc.

Hoy en día, la mayoría de nuestros smartphones cuentan con un sistema de reconocimiento de voz. Estos smartphones utilizan PNL para entender el lenguaje natural y dar la respuesta. Además, la mayoría de la gente utiliza ordenadores portátiles cuyo sistema operativo tiene incorporado el reconocimiento de voz.

ENTENDIENDO EL PLN

- Como humanos, no es una tarea muy difícil realizar el procesamiento del lenguaje natural (PNL), pero aun así, no somos perfectos. A menudo malinterpretamos una cosa por otra y a menudo interpretamos las mismas frases o palabras de manera diferente.
- Por ejemplo, considera las siguientes frases y trata de entender su interpretación de muchas maneras diferentes:

Frase: Vi a un estudiante en una colina con un prismático

Interpretaciones

- Hay un estudiante en la colina, y lo observé con mi prismático.
- Hay un estudiante en la colina, y tiene un prismático.
- Estoy en una colina, y he visto a un estudiante con mi prismático.
- Estoy en una colina, y vi a un estudiante que tiene un prismático.
- Hay un estudiante en una colina, y le vi algo con mi prismático.

DIFERENCIA ENTRE EL PLN BASADO EN REGLAS Y PLN ESTADÍSTICO

Procesamiento del lenguaje natural basado en reglas

- Utiliza el razonamiento de sentido común para las tareas de procesamiento.
- Por ejemplo,
  - La temperatura de congelación puede provocar la muerte, o
  - El café caliente puede quemar la piel de las personas,
  - Algunas otras tareas de razonamiento de sentido común, etc.
- Sin embargo, este proceso puede llevar más tiempo y requiere un esfuerzo manual.

Procesamiento estadístico del lenguaje natural

- Este tipo de PLN utiliza grandes cantidades de datos y pretende extraer conclusiones de ellos. Para entrenar los modelos de PLN, utiliza algoritmos de aprendizaje automático. Una vez completado el proceso de entrenamiento en grandes cantidades de datos, el modelo entrenado tendrá resultados positivos con la deducción.

TERMINOLOGÍA USADA EN PLN

Documento

- Conjunto de datos con la información a procesar.

Corpus

- Colección de todos los documentos presentes en nuestro conjunto de datos.

Característica (Feature)

- Cada palabra única del corpus se considera una característica.

Ejemplo:

Frases:

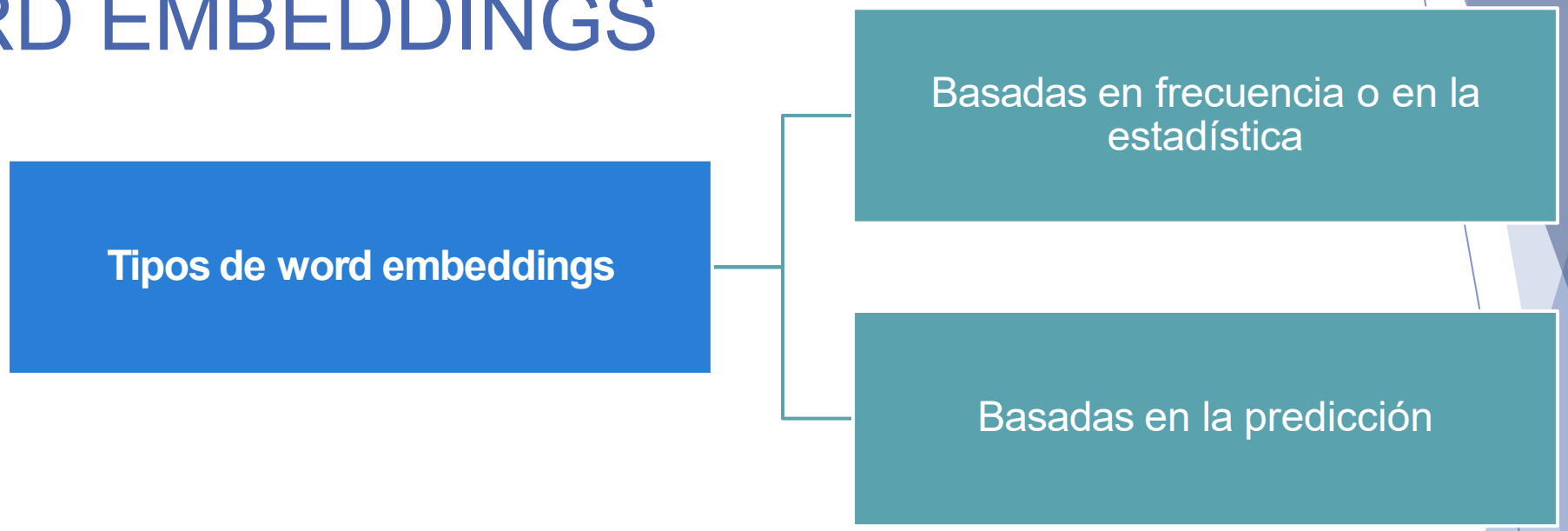
- El perro odia al gato. Le encanta salir a jugar
- Al gato le encanta jugar con una pelota.

Podemos construir un corpus a partir de los 2 documentos anteriores simplemente combinándolos.

- ✓ Corpus = “El perro odia al gato. Le encanta salir a jugar. Al gato le encanta jugar con una pelota.”
- ✓ Features = [‘El’, ‘perro’, ‘odia’, ‘al’, ‘gato’, ‘le’, ‘encanta’, ‘salir’, ‘jugar’] ←VECTOR DE CARACTERÍSTICAS

Nota: Eliminamos la ‘a’ por tener un sólo caracter

INCRUSTACIONES DE PALABRAS O WORD EMBEDDINGS



Para aplicar este tipo de incrustaciones podemos aplicar las siguientes técnicas:

1. One-Hot Encoding (OHE)
2. Count Vectorizer
3. Bag of Words (BOW) N-gramas
4. Term Frequency-Inverse
5. Document Frecuency (Vectorización TF-IDF)

INCRUSTACIONES DE PALABRAS O WORD EMBEDDINGS

1. One Hot Encoding (OHE)

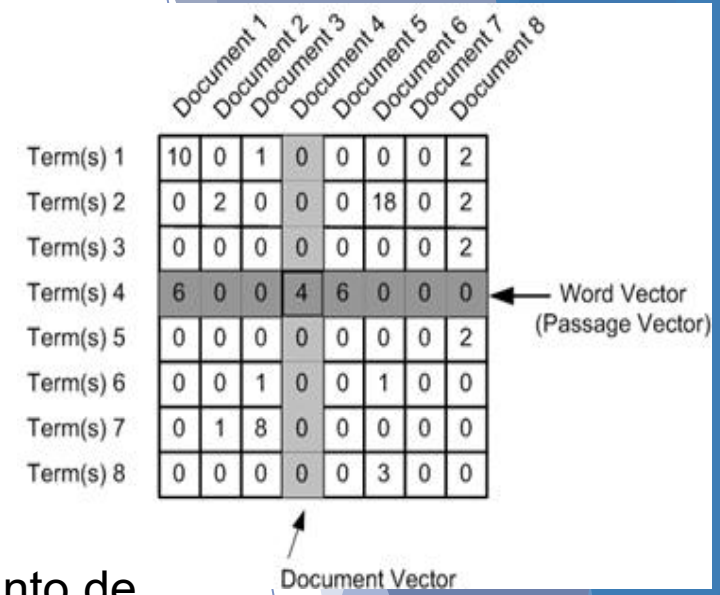
- En esta técnica, representamos cada palabra única en el vocabulario estableciendo un token único con valor 1 y el resto 0 en otras posiciones del vector.
- En palabras sencillas, la representación vectorial de un vector codificado OHE se representa en forma de 1 y 0, donde 1 representa la posición en la que existe la palabra y 0 en todas las demás.

Frase: I am teaching NLP in Python

Diccionario: ['I', 'am', 'teaching',' NLP',' in', 'Python']

Vector OHE para NLP: [0,0,0,1,0,0]

INCRUSTACIONES DE PALABRAS O WORD EMBEDDINGS



2. Count Vectorizer

- Es una de las formas más sencillas de hacer vectorización de textos

Crea una matriz de términos del documento, que es un conjunto de variables ficticias que indican si una determinada palabra aparece en el documento.

Registramos la frecuencia con la que aparecen los términos, y las columnas están dedicadas a cada palabra del corpus.

Se crea una matriz de términos del documento en la que las celdas individuales denotan la frecuencia de esa palabra en un documento concreto, lo que también se conoce como **frecuencia de términos**, y las columnas están dedicadas a cada palabra del corpus.

INCRUSTACIONES DE PALABRAS O WORD EMBEDDINGS

3. Bag of Words (BOW)

- Esta técnica de vectorización **convierte el contenido del texto en vectores de características numéricas.**

Bag of Words toma un documento de un corpus y lo convierte en un vector numérico al **asignar cada palabra del documento a un vector de características** para el modelo de aprendizaje automático.

- Requiere de 2 operaciones:
  1) Tokenización
  2) Creación de vectores de características

	and	affordable	delicious	is	not	pasta	tasty	this	very
this pasta is very tasty and affordable.	1	1	0	1	0	1	1	1	1
this pasta is not tasty and is affordable	1	1	0	2	1	1	1	1	0
this pasta is very very delicious.	0	0	1	1	0	1	0	1	2

INCRUSTACIONES DE PALABRAS O WORD EMBEDDINGS

4. Vectorización de N-gramas

Genera una matriz de términos del documento, y cada celda representa un recuento.

Las columnas representan todas las columnas de palabras adyacentes de longitud n.

La vectorización por recuento es un caso especial de N-Gram donde n=1.

Los N-gramas consideran la secuencia de n palabras en el texto; donde n es (1,2,3..) como 1- grama, 2-grama. para el par de tokens. A diferencia de BOW, mantiene el orden de las palabras.

Ejemplo:

Estoy estudiando Python NLP →Tiene 4 palabras y n=4

- *Si n=2, pej bigrama → ['Estoy estudiando', 'estudiando Python', 'Python NLP']*
- *Si n=3, pej trigrama → ['Estoy estudiando Python', 'estudiando Python NLP']*
- *Si n=4, pej cuatrigrama →['Estoy estudiando Python NLP']*

INCRUSTACIONES DE PALABRAS O WORD EMBEDDINGS

5. Vectorización TF-IDF

TF → Frecuencia de término o palabra. Frecuencia con la que ocurre la palabra en un documento.

$$tf(t,d) = \frac{n_t}{\sum_k n_k}$$

dónde:
n_t corresponde al número de veces que aparece un término en el documento
Sum n_k corresponde al total de términos que hay en el documento.

- Como cada documento puede tener una longitud diferente, un término puede aparecer más en documentos grandes que en cortos y todos los términos tienen la misma importancia.

10. Incrustaciones de palabras o Word Embeddings

- **Vectorización TF-IDF**

2. Idf → Frecuencia inversa de documento. **Permite evaluar cómo de importante es un determinado término en el documento.**

   Este valor no varía entre documentos, sólo entre palabras.

   Necesitamos saber cuántos documentos hay en nuestro corpus (representados en la fórmula con la letra D) y en cuántos aparece la palabra.

$$idf(t, D) = \log \frac{|D|}{|\{ d_i \in D \mid t \in d_i \}|}$$

Elementos como ‘el’, ‘la’, ‘eso’, … pueden aparecer muchas veces en un documento, sin embargo tienen poca importancia.

10. Incrustaciones de palabras o Word Embeddings

- **Vectorización TF-IDF**

3. Frecuencia inversa del documento:

Supongamos que tenemos un documento con 100 palabras, donde la palabra gato aparece 3 veces.

```
TF = 3 / 100 = 0.03
```

Ahora supongamos que tenemos 10 millones de documentos y gato aparece en 1000 de ellos.

```
idf = log (10.000.000 / 1.000) = 4
```

El peso que se asignaría a la palabra gato sería:

```
0.03 * 4 = 0.12
```

¿QUÉ ES EL PROCESAMIENTO DEL LENGUAJE NATURAL?

- Aplicaciones del PLN

Cortana

Hey Siri

VS

Un ejemplo

- Red Metal:
  - https://redmetalv2-nahikiv2.flywheelsites.com/

PROBLEMAS CHAT Y SISTEMAS MASIVOS

Un abogado usó ChatGPT en un juicio. Ahora es él quien debe dar explicaciones a un juez por incluir citas falsas

https://www.xataka.com/legislacion-y-derechos/abogado-uso-chatgpt-juicio-ahora-quien-debe-dar-explicaciones-a-juez-incluir-citas-falsas

Los datos se están agotando e inteligencias artificiales como ChatGPT tendrían un fin

Plataformas como DALL·E 2 y Midjourney también estarían afectadas

https://www.infobae.com/america/tecno/2023/01/11/los-datos-se-estan-agotando-e-inteligencias-artificiales-como-chatgpt-tendrian-un-fin/

Alternativa: Usar la tecnología y no la herramienta

HERRAMIENTAS IA GENERATIVA

- Generación de Imágenes:
  - https://www.freepik.com/pikaso/sketch?sign-up=google
- Transcripcion de Textos:
  - https://playground.deepgram.com/?endpoint=listen&file=4&smart_format=true&language=en&model=nova-2
- Generacion de voz:
  - https://elevenlabs.io/app/speech-synthesis/text-to-speech

HERRAMIENTAS IA GENERATIVA

- Generador de canciones:
  - https://suno.com/create
- Edición de videos:
  - https://app.runwayml.com/video-tools/teams/borjabalparda50/dashboard
- Creacion de videos:
  - https://www.synthesia.io/es#free-ai-video

DATA VALUE MANAGEMENT

Borja Balparda de Marco
CEO de Data Value Management

650.02.13.17
bbalparda@datavaluemanagement.es
@borja-balparda-de-marco-dvm
www.datavaluemanagement.es