Cápsula Ayudantía 14

Procesamiento de Lenguaje Natural

Martín Castillo y Kaina Galdames

1 de julio 2024

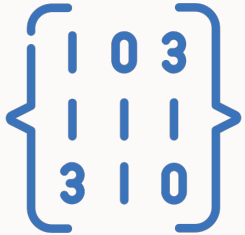
Contenidos

¿Cómo funciona la **codificación** de textos?

¿Cómo funciona la **generación** de texto?

Otras tareas de Procesamiento de Lenguaje Natural

Ejemplos en **código**



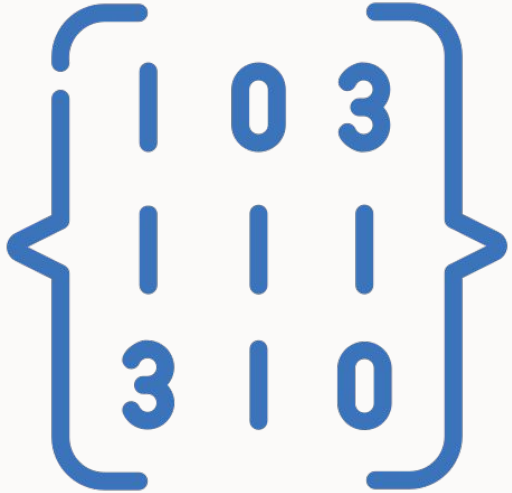
¿Cómo funciona la codificación de textos?

Codificación de textos

Alfabeto → Números

Mantiene significado

Redes Neuronales ✓



Codificación de textos

***Tokens*: Fragmentos de texto**

Tokenización en distintos niveles

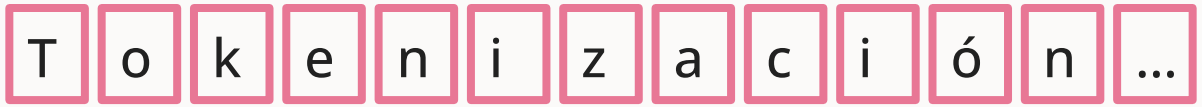
Caracteres

Palabras

Sub Palabras

Codificación de textos

Tokens: Fragmentos de texto



Caracteres Palabras Sub Palabras

Codificación de textos

Tokens: Fragmentos de texto

Tokenización	en	distintos	niveles

Caracteres **Palabras** Sub Palabras

Codificación de textos

Tokens: Fragmentos de texto

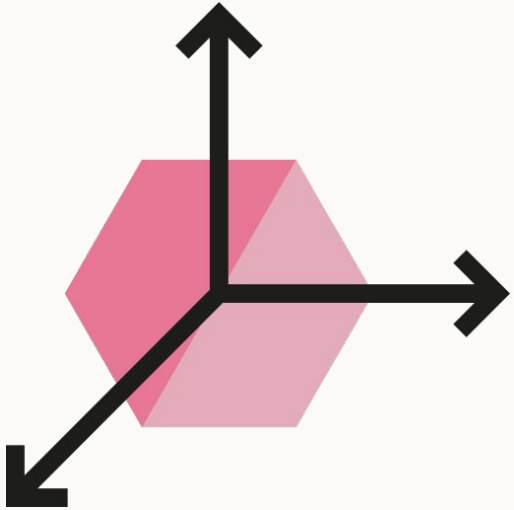
Tokeniza	ción	en	distintos	niveles

Caracteres Palabras **Sub Palabras**

Codificación de textos

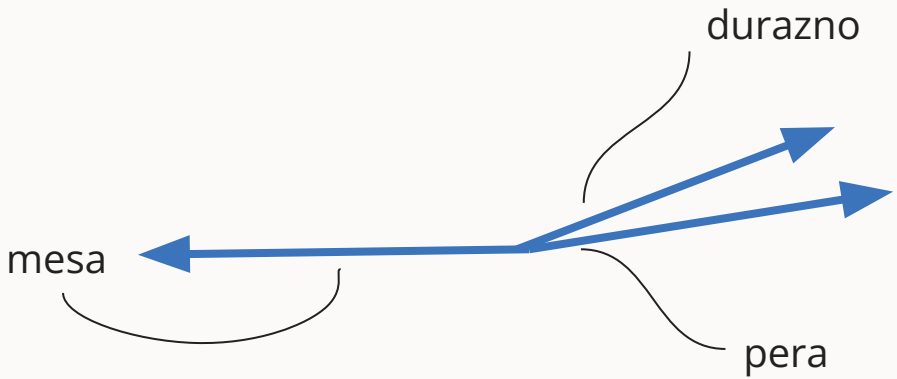
¿1 token = 1 número ?

¿1 token = 1 vector de números?



Codificación de textos

Vectores cercanos = tokens similares



Codificación de textos

¿Cómo escoger los números?

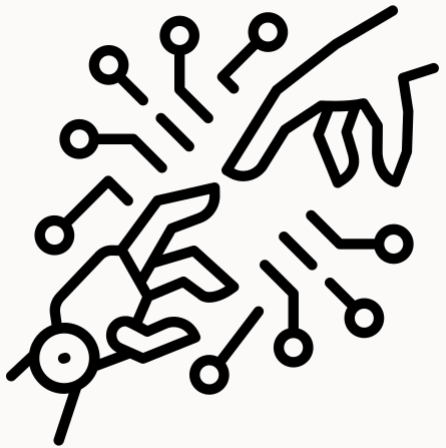
☝ Lenguaje es complejo

☝ Capturar patrones → Gran volumen

Codificación de textos

¿Cómo escoger los números?

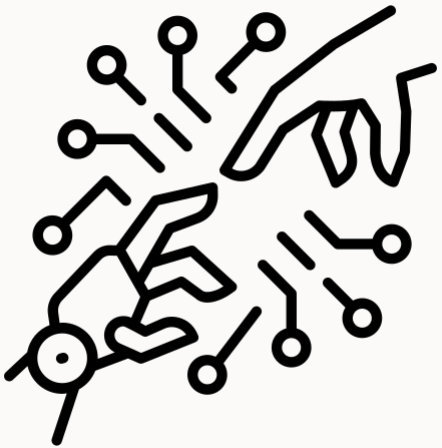
¡Redes Neuronales!



Codificación de textos

Bidirectional
Encoder
Representation
from
Transformers

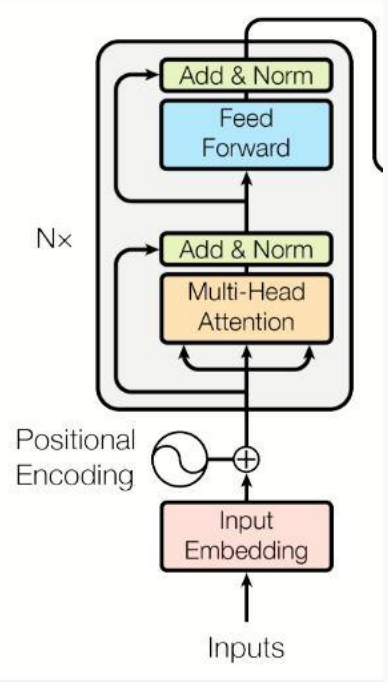
BERT



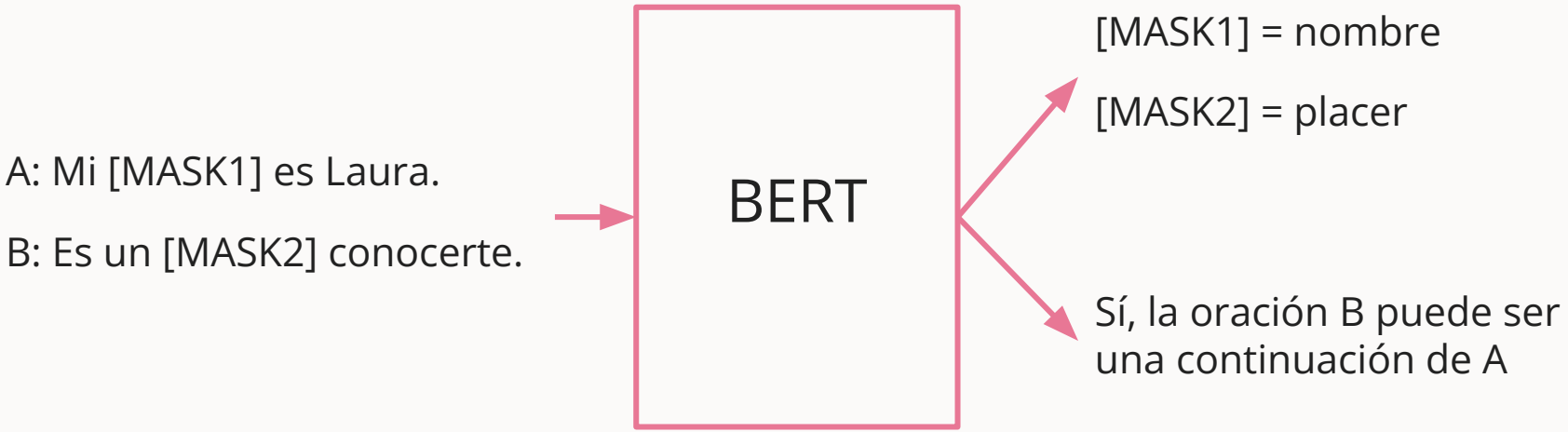
Codificación de textos

BERT

Red compuesta de capas
de codificadores de transformer



Codificación de textos

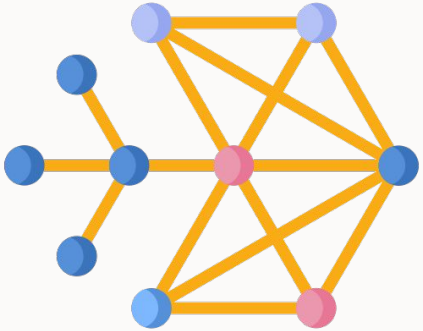


Codificación de textos

Fine tunning de BERT

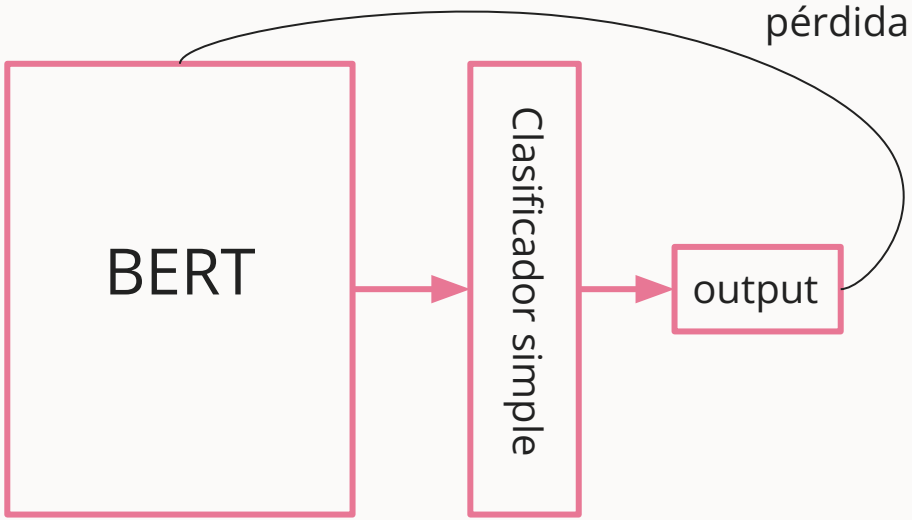
Ajusta los parámetros

Para tareas específicas



Codificación de textos

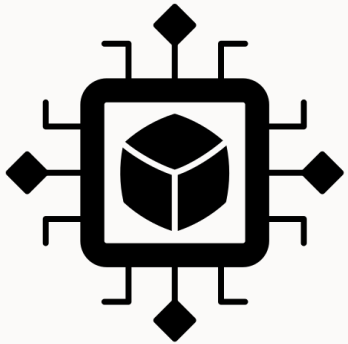
Fine tunning de BERT



Codificación de textos

Fine tunning de BERT

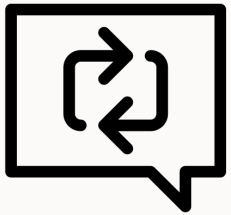
Dataset etiquetado para clasificador



¿Cómo funciona la generación de textos?

Generación de textos

1. Tokenización
2. Vectorización
3. Pasar por la red
4. Predicción
5. Selección
6. Repetición



Generación de textos

1. **Tokenización**
2. Vectorización
3. Pasar por la red
4. Predicción
5. Selección
6. Repetición

Texto se fragmenta en tokens

Generación de textos

1. Tokenización
2. **Vectorización**
3. Pasar por la red
4. Predicción
5. Selección
6. Repetición

Tokens a vectores

Generación de textos

1. Tokenización
2. Vectorización
3. **Pasar por la red**
4. Predicción
5. Selección
6. Repetición

Secuencia
se procesa

Generación de textos

1. Tokenización
2. Vectorización
3. Pasar por la red
4. **Predicción**
5. Selección
6. Repetición

Probabilidades siguiente token

Generación de textos

1. Tokenización
2. Vectorización
3. Pasar por la red
4. Predicción
5. **Selección**
6. Repetición

Elección probabilidad más alta

Generación de textos

1. Tokenización
2. Vectorización
3. Pasar por la red
4. Predicción
5. Selección
6. **Repetición**

secuencia = secuencia + token

Otras tareas de Procesamiento de Lenguaje Natural

Otras tareas de PLN

- Resumen de texto
- Traducción automática
- Análisis de sentimiento
- Clasificación de textos

Otras tareas de PLN

- Resumen de texto
- Traducción automática
- **Análisis de sentimiento**
- Clasificación de textos

Sentimientos de reseñas de productos

Otras tareas de PLN

- Resumen de texto
- Traducción automática
- Análisis de sentimiento
- **Clasificación de textos**

Detección de spam en correo

Ejemplos en código

Vamos a Colab

Cápsula Ayudantía 14

Procesamiento de Lenguaje Natural

Martín Castillo y Kaina Galdames

1 de julio 2024

Fuentes de información

- ¿Qué es tokenizar?
  - https://huggingface.co/learn/nlp-course/es/chapter2/4
  - https://medium.com/escueladeinteligenciaartificial/procesamiento-de-texto-para-nlp-1-tokenizaci%C3%B3n-4d533f3f6c9b
- Encoding:
  - https://bishalbose294.medium.com/nlp-text-encoding-a-beginners-guide-fa332d715854
- ¿Qué es BERT?
  - https://www.youtube.com/watch?v=MdEYUliufmk
  - https://www.youtube.com/watch?v=xI0HHN5XKDo

Fuentes de información

- Tareas de NLP
  - https://www.deeplearning.ai/resources/natural-language-processing/
  - https://datascientest.com/es/nlp-introduccion
  - https://www.geeksforgeeks.org/top-7-applications-of-natural-language-processing/
- Predicción del siguiente token desde 1:40
  - https://www.youtube.com/watch?v=wl3mbqOtlmM

Créditos vectores e imágenes

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. → encoder del transformer
- https://www.flaticon.com/free-icons/matrix created by Mayor Icons
- https://www.flaticon.com/free-icons/tick created by Maxim Basinski Premium
- https://www.flaticon.com/free-icons/vector created by Freepik
- https://www.flaticon.com/free-icons/code-review created by juicy_fish
- https://www.flaticon.com/free-icons/robot created by shmai
- https://www.flaticon.com/free-icons/neural-network created by Dewi Sari
- https://www.flaticon.com/free-icons/embedded created by zafdesign
- https://www.flaticon.com/free-icons/generative created by HideMaru
- https://www.flaticon.com/free-icons/font
- https://www.flaticon.com/free-icons/cycle created by TravisAvery
- https://www.flaticon.com/free-icons/nlp created by Freepik
- https://www.flaticon.com/free-icons/code created by juicy_fish