



Puertas abiertas a la excelencia



Natural Language Processing (NLP)

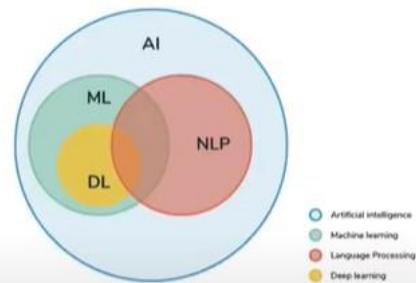
Carlos Manuel Rueda Ramírez
25 de marzo de 2023

► Introducción

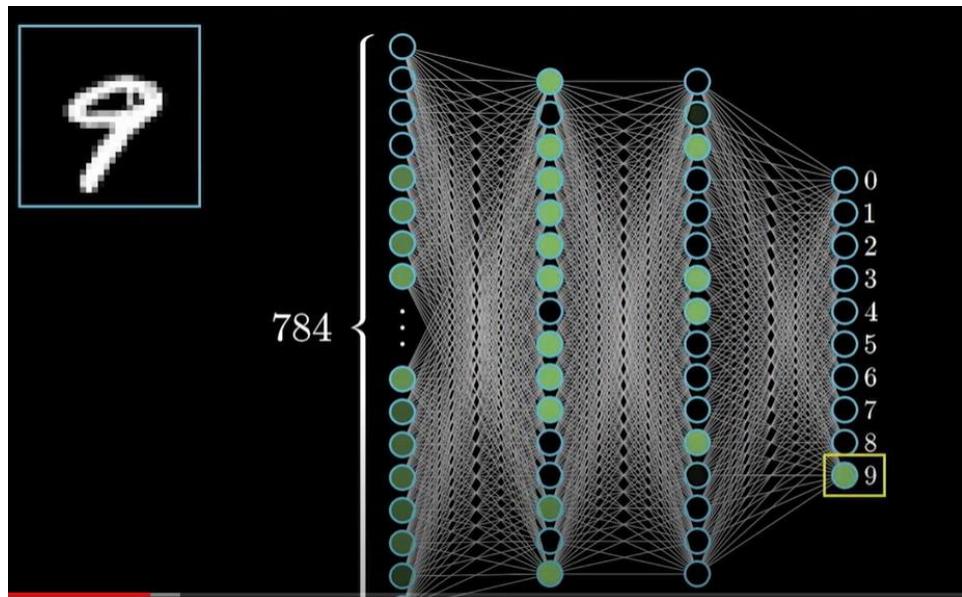
Definición de PLN

El **Procesamiento del Lenguaje Natural (PLN)** es una disciplina que busca procesar mediante ordenadores la información almacenada en textos (orales y escritos) generados mediante lenguajes naturales.

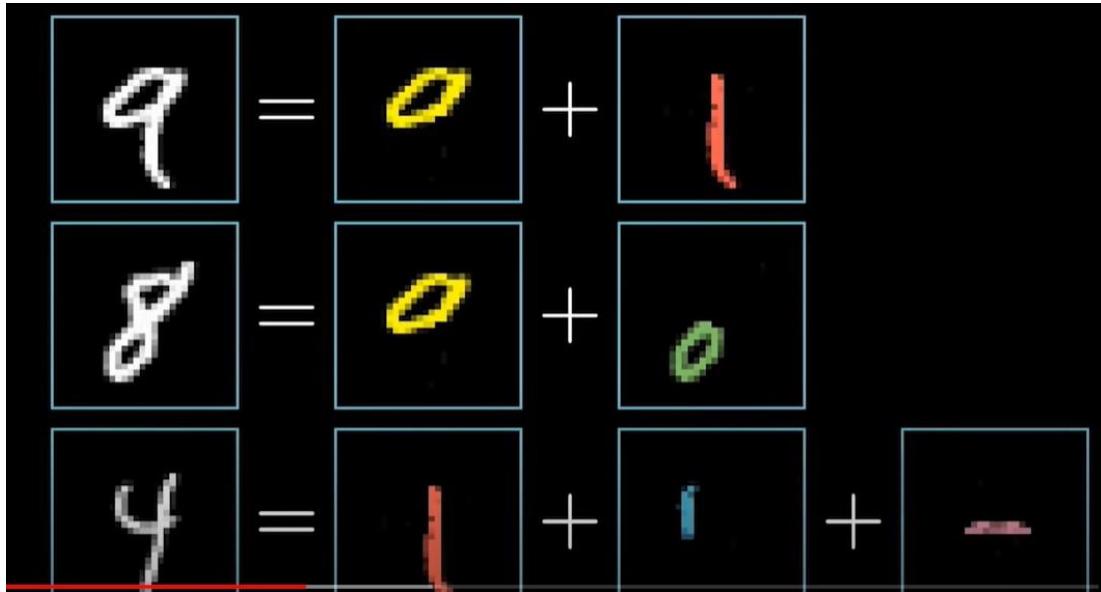
Es un subcampo de la Inteligencia Artificial pero no todo el Procesamiento de Lenguaje Natural está englobado en el Aprendizaje Automático (Machine Learning).



Redes neuronales



Redes neuronales



Redes neuronales

- Reconocimiento de caracteres, de imágenes, de voz
- Generación de texto
- Traducción de idiomas
- **Detección de fraude**
- Conducción autónoma
- Predicción bursátil
- Pronóstico de enfermedades
- Análisis genético
- **Clasificar pepinos**

Concepto de Embeddings

Hacemos la asociación de cada palabra con la misma etiqueta.

[0]

[0]

...

[0]

[0]

[0]

[0]

...

[0]

[0]

[0]

[0]

...

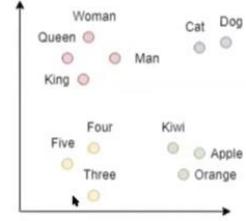
[0]

[0]

Concepto de Embedding

Concepto de Embedding

| | Humano | Metálico | Inteligente |
|------------|--------|----------|-------------|
| Celular | 0.07 | 0.94 | 0.75 |
| Terminator | 0.5 | 0.5 | 0.8 |
| Abuela | 0.95 | 0.1 | 0.85 |



► Modelos relevantes

BERT (Bidirectional Encoder Representations from Transformers)

- Primer modelo que emplea bidireccionalidad.
- Amplio rango de capacidades tan solo con añadir una capa extra (entrenado en BookCorpus y Wikipedia).
- Desaparece la direccionalidad.
- Uso de mecanismos de atención.



Ejemplo Modelo BERT (BETO)

The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** bert_example.ipynb
- Header:** Archivo, Editar, Ver, Insertar, Entorno de ejecución, Herramientas, Ayuda, Guardado el ~~última vez: 1:19~~
- Toolbar:** Comentario, Arturo Sánchez Pal... (with a profile picture), Inicializando, Editar
- Section:** BERT Example
- Description:** En este cuaderno vamos a ver un ejemplo de como el modelo BERT puede completar la palabra que falta en una oración, en ocasiones, con el mismo atino que un ser humano:
- Code Cell 4:** [4] unmasker("This is a webinar about [MASK] Language Processing.")

```
[4]: [{"score": 0.9825870394706726, "sequence": "[CLS] this is a webinar about natural language processing. [SEP]", "token": 3019, "token_str": "natural"}, {"score": 0.004522316623479128, "sequence": "[CLS] this is a webinar about machine language processing. [SEP]", "token": 3099, "token_str": "machine"}, {"score": 0.002148685045540333, "sequence": "[CLS] this is a webinar about assembly language processing. [SEP]", "token": 3320, "token_str": "assembly"}, {"score": 0.0016864982899278402, "sequence": "[CLS] this is a webinar about programming language processing. [SEP]", "token": 4730, "token_str": "programming"}, {"score": 0.0009078820585273206, "sequence": "[CLS] this is a webinar about human language processing. [SEP]", "token": 2529, "token_str": "human"}]
```

Modelos relevantes

MUSE (Multilingual Unsupervised and Supervised Embedding)

- Creado por Facebook en 2017.
- Contiene más de 30 idiomas (entre ellos el español).
- Se basa en el uso de FastText y en diccionarios bilingües.
- MUSE se utiliza a menudo para evaluar embeddings propios.



Ejemplo Modelo MUSE

```
[20] ejemplo = embed_large('Esto es un ejemplo')

In [20]: print(ejemplo)

Out[20]:
tf.Tensor(
[[-5.53136691e-02  7.51734152e-03 -1.33971944e-02  8.54406059e-02
 4.89021838e-02  2.73253657e-02 -6.46049529e-03 -9.66204517e-03
 3.59891616e-02  2.02764651e-02  3.21702845e-02 -5.52996546e-02
-2.79748995e-02 -3.26707211e-02  2.09495369e-02 -1.66090783e-02
-2.36730166e-02 -1.10793812e-02 -3.42288539e-02 -7.36196898e-03
-6.04444193e-03  1.06974415e-01  1.51025467e-02  2.22048932e-03
5.11200354e-02  3.16248537e-02  2.68503912e-02  3.38718295e-03
6.07594810e-02 -1.52052207e-02 -1.25000000e-01  2.91699674e-02
-5.17448202e-02 -5.86271676e-02  1.30695375e-02  2.000031164e-02
7.55034834e-02 -2.40348149e-02  5.63987643e-02 -4.88163000e-02
1.03335920e-02  3.82852480e-02  5.19630797e-02 -4.732464651e-03
5.087989324e-02 -3.47664866e-03  3.10360100e-02  7.41312727e-03
1.49562387e-02  3.36607127e-03 -2.59210709e-02  4.39192588e-03
-2.23776251e-02  2.75029894e-02  5.00163659e-02  3.55228558e-02
2.59855520e-02  5.96970040e-03 -4.88507897e-02  4.87358272e-02
-2.52136011e-02  2.76784040e-02  5.98622439e-03 -5.40143773e-02
-1.97089426e-02  2.79733352e-02 -3.33456360e-02 -1.33442087e-02
-2.89916340e-02  5.82487276e-03  3.77151109e-02  3.74950692e-02
4.34260396e-03  5.22547774e-02  5.46167232e-03 -1.79926977e-02
7.61899864e-03 -7.94465616e-02 -6.53714240e-02 -5.04223332e-02
-1.37146171e-02 -1.19473995e-03 -3.17683257e-02  1.74814928e-03]
```

```
In [21]: np.inner(embed_large('Hola'), embed_large('Saluts'))

Out[21]:
array([0.92447877], dtype=float32)

In [22]: np.inner(embed_large('Hola'), embed_large('dog'))

Out[22]:
array([0.36329132], dtype=float32)
```

Modelos relevantes

GPT-3 (Generative Pretrained Transformer)

- Creado por OpenAI en junio 2020.
- Es un modelo generador de textos.
- Realmente potente, 175.000.000.000 parámetros.
- Genera frases pero también código informático e incluso diálogos.



Ejemplo Modelo GPT-3

The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** gpt_example.ipynb
- Toolbar:** Archivo, Editar, Ver, Insertar, Entorno de ejecución, Herramientas, Ayuda, Se han guardado todos los cambios.
- Code Cells:**
 - Cell 1:

```
generator("This is a conference about Natural Language Processing", max_length=50, num_return_sequences=3)
```

Output:

```
Setting 'pad_token_id' to 50256 (first 'eos_token_id') to generate sequence
[{'generated_text': "This is a conference about Natural Language Processing, we've brought your software to this stage in the recent years, where we talk abo", 'generated_text': "This is a conference about Natural Language Processing (NNP) and I think that some of you may be curious as to why this is happening. In", 'generated_text': "This is a conference about Natural Language Processing.\n\n\nThis is an attempt to document and publish Natural Language Processing (NLP"}]
```
 - Cell 2:

```
generator("How are you doing?", max_length=50, num_return_sequences=3)
```

Output:

```
g 'pad_token_id' to 50256 (first 'eos_token_id') to generate sequence
[{'generated_text': "How are you doing? Have you ever slept on a treadmill? You do not care, your eyes are open? You're not listening to my voice, aren't you? How i", 'generated_text': "erated_text": "How are you doing? I can see how everyone's looking at the same thing. It's not that anybody should be scared of us. What are we waiting for?", 'generated_text': "How are you doing?\n\nTotally. I've been working on some really cool stuff for a while now - like the music project I'm doing. We've worked for", 'generated_text': "erated text": "How are you doing?\n\nNo I am not. Yes I need to be doing more in my life. I want to be happy and healthy. But that is hard. That is life.\n\n"}]
```
- User Information:** Arturo Sánchez Pal...

Ejemplo Modelo GPT-3

Actualmente este modelo solo está disponible en inglés. Observamos que si probamos en español los resultados son terribles:

```
generator("Esto es una charla sobre Procesamiento de Lenguaje Natural", max_length=60, num_return_sequences=5)

Setting 'pad_token_id' to 50256 (first 'eos_token_id') to generate sequence
[{'generated_text': 'Esto es una charla sobre Procesamiento de Lenguaje Natural del Marques Natural del Amelario Naturalized Naturalized El Marques Osteria.', 'generated_text': 'Esto es una charla sobre Procesamiento de Lenguaje Naturales para los Más Sárpas que sobre la brasilian brasiliano.\n\nAproposas las nuev', 'generated_text': 'Esto es una charla sobre Procesamiento de Lenguaje Naturales (5 min) [Spanish translation]\n\nThe greatest achievement of my youth h', 'generated_text': 'Esto es una charla sobre Procesamiento de Lenguaje Natural. También porque susa sua sua susa.\n\nThe caras are not painted at all. I bought', 'generated_text': 'Esto es una charla sobre Procesamiento de Lenguaje Natural de los Avantados.\n\nCómo más la propería de las esximentiros perdus con de l'}
```

Referencias

Sanchez, A. [Databits].(2 de octubre de 2020).*Procesamiento del Lenguaje Natural*[Video].
<https://www.youtube.com/watch?v=cLLpyQQebF8>.

[Aprende IA]. (27 de abril de 2021).*¿QUÉ ES EL PROCESAMIENTO DEL LENGUAJE NATURAL? 06 Inteligencia Artificial 101*[Video]. <https://www.youtube.com/watch?v=LK9ftWUuw>.

[3Blue1Brown Español].*¿Qué es una Red Neuronal? Aprendizaje Profundo. Capítulo*[Video].
<https://www.youtube.com/watch?v=jKCQsndqEGQ>.

[Sensio Coders]. (7 de septiembre de 2020).*Procesamiento de Lenguaje Natural Generación de Texto*[Video]. https://www.youtube.com/watch?v=uZ2bH5O_8f0.

Universitat Politècnica de València. [Dot CSV].].(15 de julio de 2020).*INTRO al NaturalLanguage Processing (NLP) #2- ¿Qué es un EMBEDDING?* [Video].<https://www.youtube.com/watch?v=RkYuHK7Fx4>.

[AMP Tech].(28 de julio de 2017). ¿Cómo funcionan las redes neuronales? [Video]
<https://www.youtube.com/watch?v=IQMoglfpBk>.