

# Natural Language Processing Intro

# Profesor

Eduardo Fernández Carrión



Lead Data Scientist | ML/AI Engineer

PhD en Métodos Estadístico-Matemáticos para el Tratamiento Computacional de la Información (UCM)

Ciencias Matemáticas (UCM)

Ingeniería Informática (URJC)



# ¿Qué es NLP?

NLP es un subcampo de la lingüística, la informática y la inteligencia artificial que se ocupa de las **interacciones entre las computadoras y los lenguajes humanos (naturales)**.

Específicamente, se enfoca en **cómo programar las computadoras** para procesar y analizar grandes cantidades de datos de lenguaje natural. Los desafíos en el procesamiento del lenguaje natural a menudo involucran el reconocimiento de voz, la comprensión del lenguaje natural y la generación de lenguaje natural.

Wikipedia



# Aplicaciones populares

**Análisis de Sentimientos:** Análisis automático del sentimiento de los clientes en redes sociales, clasificando publicaciones como positivas, negativas, neutras o identificando emociones específicas.

**Extracción de Información:** Creación de datos estructurados a partir de documentos no estructurados, como la extracción de entidades y relaciones de textos largos, por ejemplo, en noticias.

**Traducción de Textos:** Textos en un idioma que son traducidos a otro idioma (en lote o en tiempo real).

**Respuesta a Preguntas (QA):** Los asistentes digitales no sólo reconocen el habla, sino que también buscan y entregan respuestas adecuadas a las preguntas formuladas por los usuarios.

**Resumen de Texto:** Las máquinas pueden resumir documentos largos, facilitando a profesionales como abogados, analistas de negocios y estudiantes, el proceso de revisión y selección de documentos relevantes.



# Aplicaciones populares

**Chatbots:** Bots en sitios web que interactúan automáticamente con los usuarios, determinando el propósito de la visita y respondiendo a preguntas sin intervención humana.

**Generación de Texto y Audio:** Software que utiliza aprendizaje automático para generar texto y audio, como la sugerencia de oraciones completas en Gmail o la generación de resúmenes textuales de bases de datos.

**Conversión de Texto a Voz y Voz a Texto:** Software capaz de convertir texto en audio de alta calidad y viceversa, en múltiples idiomas y dialectos.

**Asistentes Digitales de Voz:** Dispositivos como Amazon Alexa, Apple Siri, Google Assistant, y otros, que reconocen la voz humana con alta precisión y responden en tiempo real.

**Agentes Autónomos Inteligentes:** Ejecutar tareas de manera autónoma, como agentes de búsqueda de información, planificación de proyectos o asistentes personales mejorados que pueden interactuar con múltiples aplicaciones y servicios sin intervención humana.



# Breve historia

## Timeline

1950s

**1950** – Alan Turing introduce el **Test de Turing**.

**1954** – Primer experimento de **traducción automática**: Georgetown-IBM traduce 60 frases del ruso al inglés (Thomas Watson). Sistema basado en reglas.

1960s

**1960s** – Crecen los modelos **basados en reglas** y **gramática formal** (lingüistas).

**1966** – El informe de **DARPA** concluye que la traducción automática no era viable en ese momento, reduciendo la financiación en EE.UU.

1990s

**1990** – **IBM** desarrolla **Hidden Markov Models (HMMs)** para el **reconocimiento de voz** y **etiquetado de texto**.

**1998** – Se publican **WordNet** y los primeros **corpus** de referencia para NLP.

2000s

**2001** – **LSA** (Latent Semantic Analysis) y **TF-IDF** se usan ampliamente para **búsqueda** y **clasificación de documentos**.



# Breve historia

## Timeline

2000s

**2003** – Surge el **Perceptrón multicapa (MLP)** y las primeras aplicaciones de **SVM** en NLP.

**2008** – Se popularizan los **modelos de tópicos**, como **LDA** (Latent Dirichlet Allocation).

2010s

**2013** – Google introduce **Word2Vec**, revolucionando la representación de palabras con embeddings.

**2014** – Facebook y Google implementan **seq2seq** y modelos LSTM para traducción automática.

**2015** – Aparecen modelos basados en **memoria de atención**.

**2017** – Google publica "*Attention is All You Need*", introduciendo **Transformers** (base de los LLM modernos).

**2018** – OpenAI lanza **GPT-1**, usando pre-entrenamiento + ajuste fino.

**2019** – Google introduce **BERT**, que mejora la comprensión contextual de las palabras.



# Breve historia

## *Timeline*

2020s

**2020** – OpenAI lanza **GPT-3** con 175B de parámetros, llevando NLP a otro nivel.

**2021** – DeepMind introduce **Gopher**, un modelo de 280B parámetros.

**2022** – OpenAI lanza **ChatGPT (basado en GPT-3.5)**, democratizando el NLP conversacional.





# Breve historia

## *Evolución*

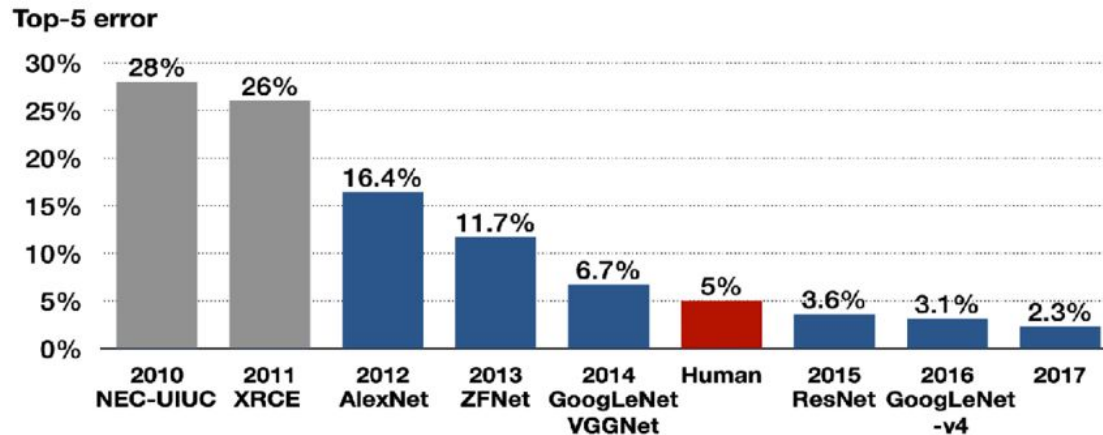
Modelos basados en reglas y gramática formal	Procesos muy manuales
Modelos estadísticos y probabilísticos	Algoritmos clásicos basados en probabilidad
Modelos basados en aprendizaje automático	Las redes neuronales son populares en problemas clásicos de ML y Time Series, pero no en NLP
Modelos basados en redes neuronales	Las redes neuronales profundas comienzan a dar resultado en Computer Vision, pero necesitarían un impulso adicional para que consiguieran converger en NLP
Modelos basados inteligencia artificial generativa	LLM dedicados a generación de texto
Futuro de NLP	Modelos más eficientes, alineados y multimodales que combinan texto, imágenes y voz.



# Momentos clave

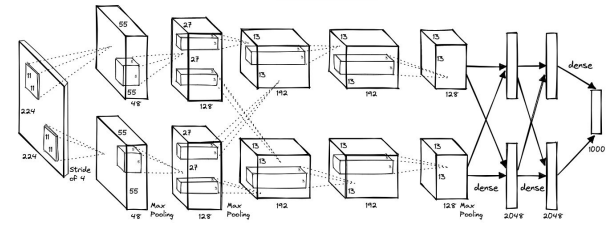
*Alexnet (2012)*

Computer Vision alcanzaron su punto de inflexión en 2012 cuando la solución basada en aprendizaje profundo, **AlexNet**, redujo drásticamente la tasa de error de los modelos de visión artificial en **Large Scale Visual Recognition Challenge** de ImageNet (**ILSVRC**).



# Momentos clave

## Características de Alexnet



- **Profundidad de la Red:** AlexNet estaba compuesta por 8 capas aprendibles, significativamente más profundas que las CNNs anteriores. Esta profundidad permitió que la red aprendiera características más complejas a diferentes niveles de abstracción.
- **Uso de ReLU (Rectified Linear Units):** Aunque las unidades ReLU ya existían, AlexNet fue una de las primeras redes neuronales en utilizarlas ampliamente. Las ReLUs ayudaron a acelerar el entrenamiento de la red, ya que resuelven el problema del gradiente desvaneciente que ocurre con las funciones de activación sigmoidales o hiperbólicas.
- **Overlap Pooling:** AlexNet introdujo el concepto de "overlap pooling", donde los pasos del "pooling" (agrupación) se superponen. Esto ayudó a reducir el tamaño de la red y a mejorar el rendimiento.
- **Uso de GPUs para el Entrenamiento:** AlexNet fue entrenada usando dos GPUs (NVIDIA), lo que fue innovador en ese momento. Esto permitió un entrenamiento más rápido y eficiente, y abrió el camino para el desarrollo de redes más grandes y complejas.
- **Regularización con Dropout:** AlexNet utilizó la técnica de *dropout* para reducir el sobreajuste en las capas completamente conectadas. Dropout "apaga" aleatoriamente un conjunto de neuronas durante el entrenamiento, lo que ayuda a que la red sea más robusta y menos propensa a sobreajustarse a los datos de entrenamiento.
- **Data Augmentation:** AlexNet aplicó técnicas de aumento de datos, como el recorte aleatorio y las rotaciones de las imágenes, para ampliar artificialmente el conjunto de datos de entrenamiento. Esto ayudó a mejorar la generalización del modelo.

# Momentos clave

*Alex Krizhevsky, Ilya Sutskever y Geoffrey Hinton*

Alex Krizhevsky:

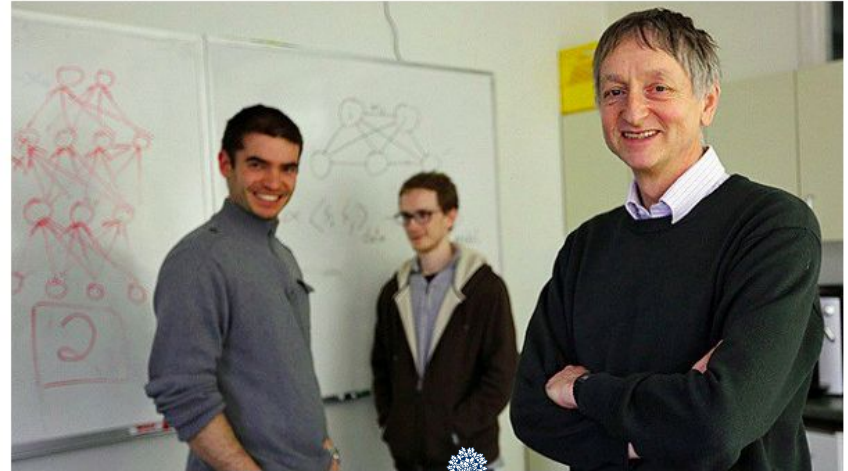
- Google Brain

Ilya Sutskever:

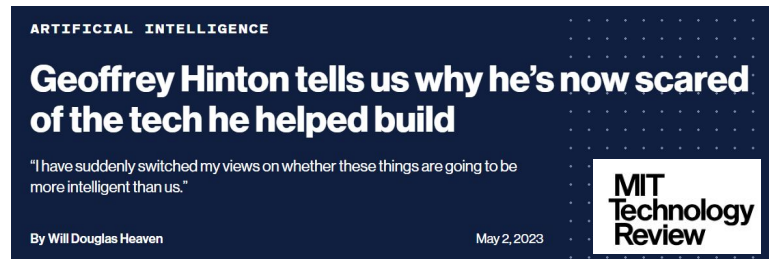
- Google Brain
- Co-Fundador y chief scientist en OpenAI

Geoffrey Hinton:

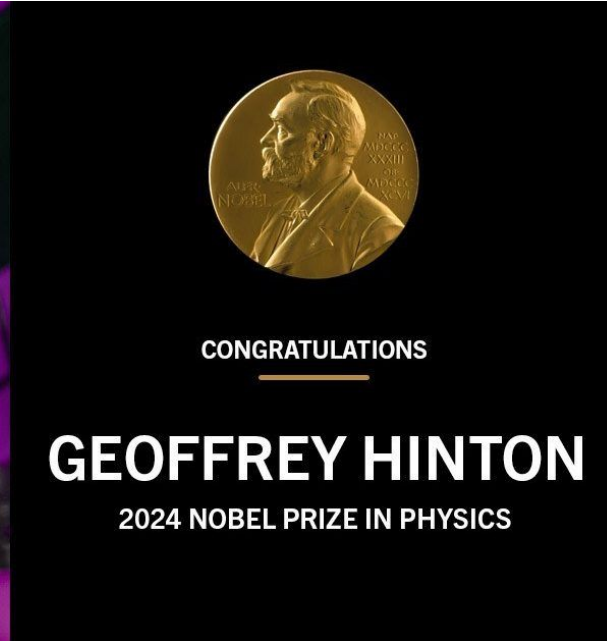
- Google DeepMind
- Creador del Backward propagation algorithm



UNIVERSITY OF  
TORONTO



# Momentos clave



# Momentos clave

*Embeddings (2013-14)*

- En 2013, se publicó el algoritmo de **Word2Vec** por Google.
- En 2014, se publicó el algoritmo de **GloVe** por la Universidad de Stanford.

Ambos artículos mostraron un algoritmo de aprendizaje no supervisado para obtener representaciones vectoriales de palabras dando al llamado **Espacio de Embeddings**.

El uso de Embeddings fomentó el uso de Redes Neuronales en los casos de uso de NLP, dado que, hasta la fecha, las soluciones de NLP basadas en algoritmos basados en frecuencias eran mucho más eficientes computacionalmente y precisos, en general.



Esto lo veremos con mayor detalle en este curso



# Momentos clave

*Attention is all you need (2017)*

En 2017 se publicó un paper llamado *Attention is all you need* en el que se describió la arquitectura **Transformers**, dando luz a el modelo **BERT**.

Los modelos basados en Transformers establecieron nuevos estándares en una amplia gama de tareas de NLP, incluyendo traducción automática, resumen de texto, generación de texto, comprensión lectora, y más.

Han demostrado ser superiores en calidad y precisión en comparación con las arquitecturas anteriores.

Esto lo veremos con mayor detalle en este curso



# Momentos clave

*chatGPT (2022)*

En 2019, modelos generativos como **GPT-2** de **OpenAI** causaron sensación, generando nuevo contenido al instante basado en contenido previo, una hazaña previamente insuperable.

En 2020, OpenAI lanzó una versión aún más grande e impresionante, **GPT-3**, basándose en sus éxitos anteriores y con un total de **175 mil millones de parámetros**.

Podría considerarse de los primeros **LLM** de la historia

**En noviembre 2022, OpenAI lanzó chatGPT**

Can you give me feedback on my short story?

**ChatGPT**

Your writing is engaging and provides insight into the protagonist's feelings and past experiences. You effectively use descriptive language to paint a picture of the setting and the characters. The dialogue between the protagonist and their father feels natural and adds depth to the story. Additionally, your writing style is clear and easy to follow. However, there are a few areas where you could improve:

1. Punctuation and grammar: There are some minor punctuation and grammar issues throughout the text that could be polished for a smoother reading experience.

2. Clarity: Some parts of the story could benefit from a bit more clarity or elaboration, especially in the transition between scenes and the protagonist's internal thoughts.

3. Show, don't tell: While you do a good job of describing the physical setting, there are moments where you tell the reader about the character's feelings and past experiences. Consider showing these emotions through actions, dialogue, or internal monologue to make the narrative more immersive.

Overall, your writing is well-crafted and engaging, with room for minor improvements in mechanics and clarity. Keep up the good work! ●

Message ChatGPT...

ChatGPT can make mistakes. Consider checking important information.







2015, San Francisco.

Investigación de IA avanzada,  
especialmente en modelos de lenguaje.

- Desarrollo de GPTs
- ChatGPT, Dall-E y Whisper

Microsoft acuerda una inversión de 10  
mil millones.



2010, Londres.

Aprendizaje profundo y aprendizaje por  
refuerzo para sistemas de IA autónomos.

- AlphaGo, AlphaFold
- T5, Lambda, Gemini

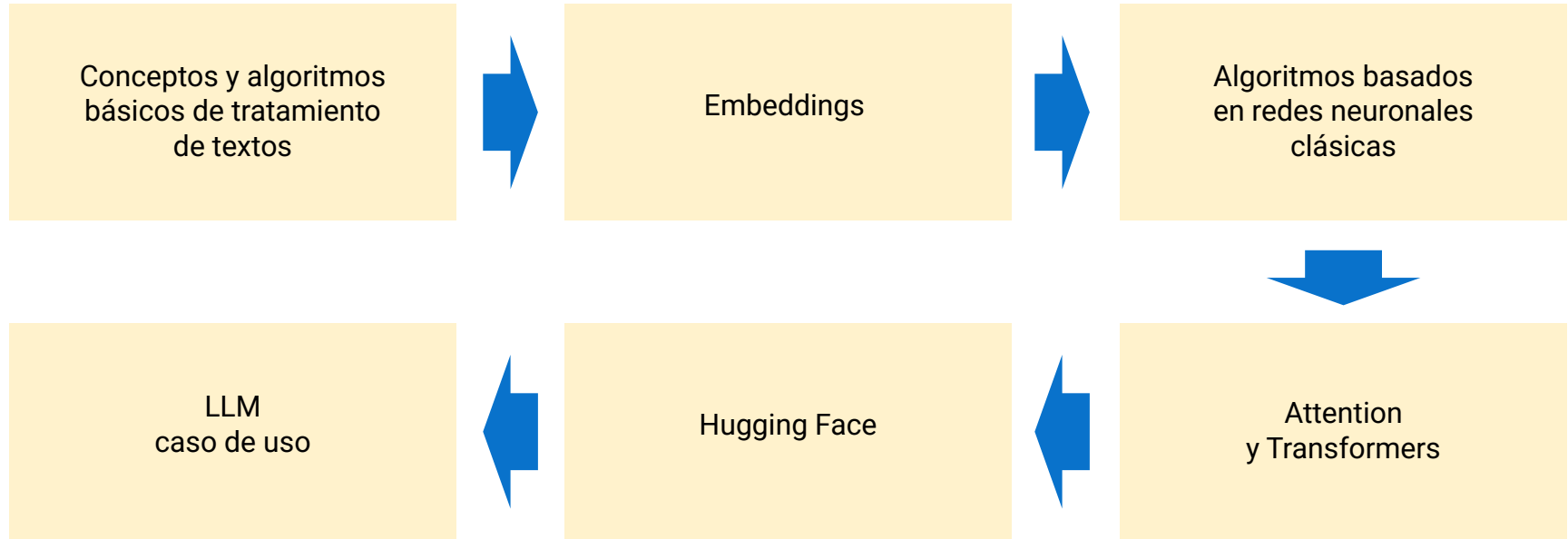
Adquisición por Google en 2014.



- Las GPUs de NVIDIA han sido fundamentales para acelerar el entrenamiento de modelos de NLP.
- NVIDIA no solo proporciona el hardware, sino también una serie de herramientas y bibliotecas de software (como CUDA, cuDNN, y TensorRT)



# Este curso



# Entornos Virtuales

```
# python=3.10.13
```

```
scikit-learn==1.5.2  
matplotlib==3.10.1  
seaborn==0.13.2  
tensorflow==2.15.0  
nltk==3.9.1  
bokeh==3.6.1  
gensim==4.3.3  
sentence-transformers==2.7.0  
spacy==3.7.5  
notebook==6.5.6
```

```
python -m spacy download es_core_news_sm
```

1

Se puede usar Google  
Colab o similar



```
# python=3.11.9
```

```
scikit-learn==1.5.2  
matplotlib==3.10.1  
seaborn==0.13.2  
tensorflow==2.15.0  
transformers==4.50.1  
datasets==3.4.1  
torch==2.6.0  
accelerate==1.5.2  
notebook==6.5.6  
sentencepiece==0.2.0
```

2

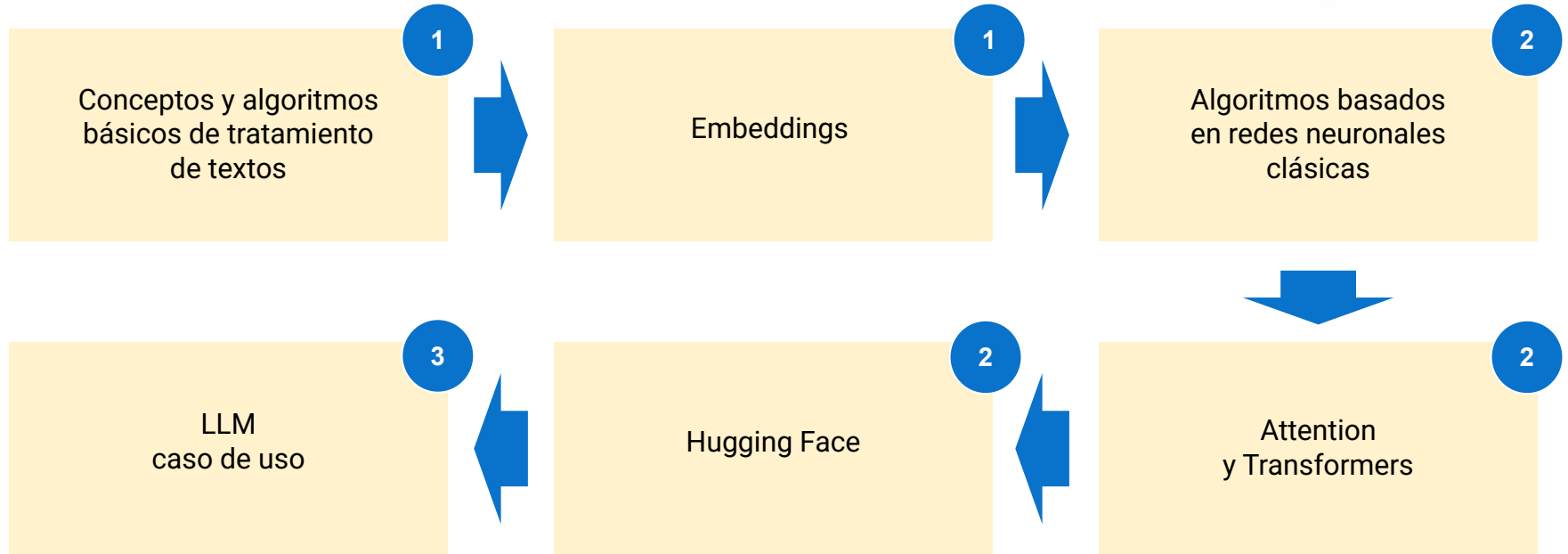
```
# python=3.11.9
```

```
PyPDF2==3.0.1  
openai==1.69.0  
llama-index==0.12.4  
llama-index-core==0.12.4  
sentence-transformers==2.7.0  
faiss-cpu==1.10.0  
notebook==6.5.6  
langchain==0.3.22  
langgraph==0.3.24  
langchain-openai==0.3.12  
langgraph-prebuilt==0.1.8
```

3



# Entornos Virtuales



# Instalación de Pytorch (con GPU Nvidia)

Si tenemos GPU de NVIDIA, podemos instalar pytorch para que ejecute con CUDA y el rendimiento sea mayor.

Con NLP, sin GPU los procesos de entrenamiento pueden ser muy largos

PyTorch Build	Stable (2.1.1)		Preview (Nightly)	
Your OS	Linux		Mac	Windows
Package	Conda	Pip	LibTorch	Source
Language	Python		C++ / Java	
Compute Platform	CUDA 11.8	CUDA 12.1	ROCm 5.6	CPU
Run this Command:	<pre>pip3 install torch torchvision torchaudio</pre>			

## ¿Como sabemos si tenemos CUDA?

nvidia-smi



NVIDIA-SMI 546.33		Driver Version: 546.33		CUDA Version: 12.3	
GPU Name	TC/MDM	Bus-ID	Disp.A	Volatile	Uncorr. ECC
ran Temp	Pwr:Usage/Cap		Memory-Usage	GPU-Util	Compute M. MIG M.
0 NVIDIA GeForce RTX 2060	MDM	00000000:01:00.0	On		N/A
N/A 45C	6W / 80W	251MiB / 6144MiB		1%	Default N/A

Processes:						
GPU	GI	CI	PID	Type	Process name	GPU Memory Usage
ID	ID	ID				
0	N/A	N/A	1964	C+G	...Brave-Browser\Application\brave.exe	N/A
0	N/A	N/A	4196	C+G	C:\Windows\explorer.exe	N/A
0	N/A	N/A	4840	C+G	...Search_cw5n1h2xyeyw\SearchApp.exe	N/A
0	N/A	N/A	8948	C+G	...t.LockApp_cw5n1h2xyeyw\LockApp.exe	N/A
0	N/A	N/A	9876	C+G	...sliveControlPanel\SystemSettings.exe	N/A
0	N/A	N/A	11956	C+G	...CBS_cw5n1h2xyeyw\TextInputHost.exe	N/A
0	N/A	N/A	13100	C+G	...64_bwek3j3d8bwe\CalculatorApp.exe	N/A
0	N/A	N/A	13180	C+G		N/A

# Datasets

Phillip Keung, Yichao Lu, György Szarvas and Noah A. Smith. “*The Multilingual Amazon Reviews Corpus*”. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

- [https://raw.githubusercontent.com/eduardofc/data/main/amazon\\_sports.csv](https://raw.githubusercontent.com/eduardofc/data/main/amazon_sports.csv)
- [https://raw.githubusercontent.com/eduardofc/data/main/amazon\\_electronics.csv](https://raw.githubusercontent.com/eduardofc/data/main/amazon_electronics.csv)
- [https://raw.githubusercontent.com/eduardofc/data/main/amazon\\_home.csv](https://raw.githubusercontent.com/eduardofc/data/main/amazon_home.csv)

Anki cards para traducir frases entre español e inglés: <https://www.manythings.org/anki/>

- [https://raw.githubusercontent.com/eduardofc/data/main/es\\_en.csv](https://raw.githubusercontent.com/eduardofc/data/main/es_en.csv)

PDFs de Coberturas de pólizas de Santalucía: <https://www.santalucia.es/seguros-hogar/coberturas-y-garantias>

- <https://github.com/eduardofc/data/blob/main/AF-SL-TablasGarantiasHogar-Completo-Modalidad14-v11-CAS.pdf>
- <https://github.com/eduardofc/data/blob/main/AF-SL-TablasGarantiasHogar-Eficaz-Modalidad13-v5-CAS.pdf>
- <https://github.com/eduardofc/data/blob/main/AF-SL-TablasGarantiasHogar-Premium-Modalidad15-v13-CAS.pdf>

# Bibliografía

