ntic master
revolucionamos la comunicación

UNIVERSIDAD COMPLUTENSE MADRID

Natural Language Processing

Intro

Eduardo Fernández Carrión

¿Qué es NLP?

NLP es un subcampo de la lingüística, la informática, la ingeniería de información 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 (acceso en marzo de 2021)

Aplicaciones populares

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.

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.

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

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.

Aplicaciones populares

Voicebots: Agentes de voz automatizados que ahora pueden manejar interacciones más complejas, utilizados en ventas, marketing y atención al cliente.

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.

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

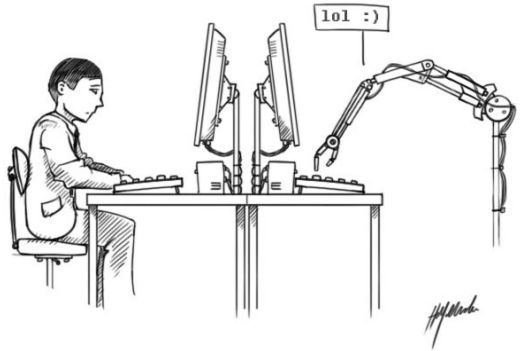
Extracción de Información: Creación de datos estructurados a partir de documentos no estructurados, como la extracción de entidades y relaciones de textos largos, por ejemplo, en noticias.

Breve historia

Test de Turing (1950)

El **Test de Turing** es una prueba de la capacidad de una máquina para demostrar inteligencia indistinguible de la humana.

Para que una máquina apruebe el Test de Turing, debe generar respuestas similares a las humanas de tal manera que un evaluador humano no pueda diferenciar si las respuestas fueron generadas por una persona o una máquina (es decir, las respuestas de la máquina son de calidad humana).



https://en.wikipedia.org/wiki/Turing_test

Breve historia

Hitos importantes de los 80

- En la década de 1980, el NLP cobró importancia con sistemas de traducción automática estadística, liderados por **IBM** (Thomas Watson). Antes de esto, la traducción automática se basaba en reglas hechas a mano, un proceso laborioso y limitado. La traducción automática estadística, que aprende de textos bilingües, redujo la necesidad de estas reglas y mejoró con más datos.
- A mediados de la década de 1980, **IBM** aplicó un enfoque estadístico al reconocimiento de voz y lanzó una máquina de escribir activada por voz llamada Tangora, que podía manejar un vocabulario de 20,000 palabras.
- **DARPA**, **Bell Labs** y la **Universidad Carnegie Mellon** también lograron éxitos similares a finales de los 80. Para entonces, los sistemas de software de reconocimiento de voz tienen vocabularios más amplios que el humano promedio y podían manejar el reconocimiento de habla continua, un hito importante en la historia del reconocimiento de voz.

Breve historia

Hitos importantes de los 90-00-10

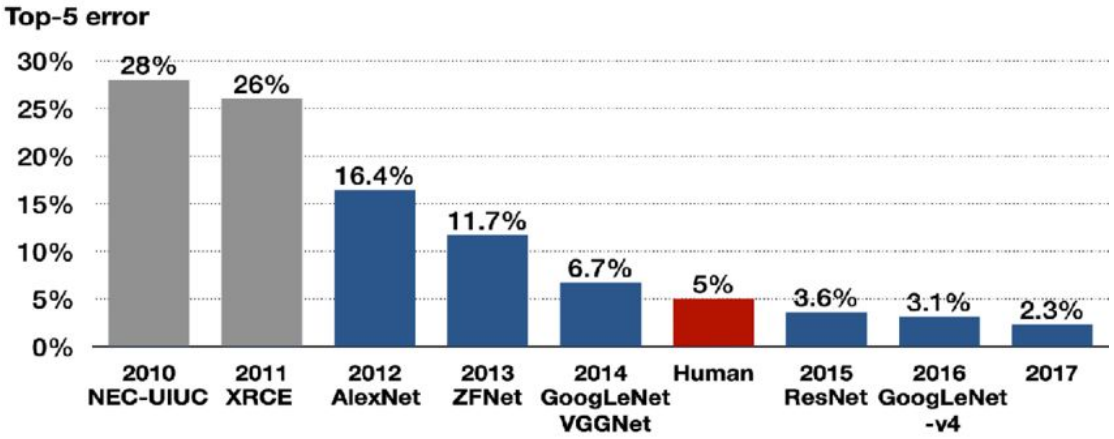
- En la década de 1990, varios investigadores del campo abandonaron laboratorios y universidades para trabajar en la industria, lo que llevó a más aplicaciones comerciales del reconocimiento de voz y la traducción automática.
- En 2007, **Google** empieza a contratar expertos en speech recognition. El gobierno de EE.UU. también se involucró en esa época; la Agencia de Seguridad Nacional (NSA) comenzó a etiquetar grandes volúmenes de conversaciones grabadas en busca de palabras clave específicas, facilitando el proceso de búsqueda para los analistas de la NSA.
- A principios de la década de 2010, los investigadores de NLP, tanto en el ámbito académico como en la industria, comenzaron a experimentar con **redes neuronales profundas** para tareas de NLP.

Reglas estadísticas → Redes neuronales

Momentos clave

Alexnet (2012)

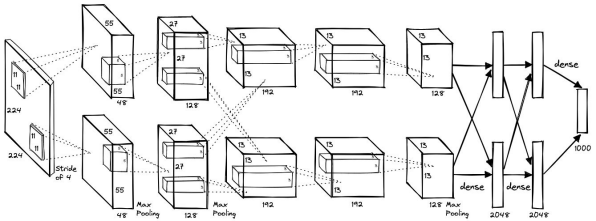
Computer Vision alcanzaron su punto de inflexión en 2012 cuando la solución basada en aprendizaje profundo, **AlexNet**, redujo drásticamente la tasa de error de los modelos de visión artificial en **Large Scale Visual Recognition Challenge** de **ImageNet (ILSVRC)**.



https://image-net.org/challenges/LSVRC/

Momentos clave

Característics 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:

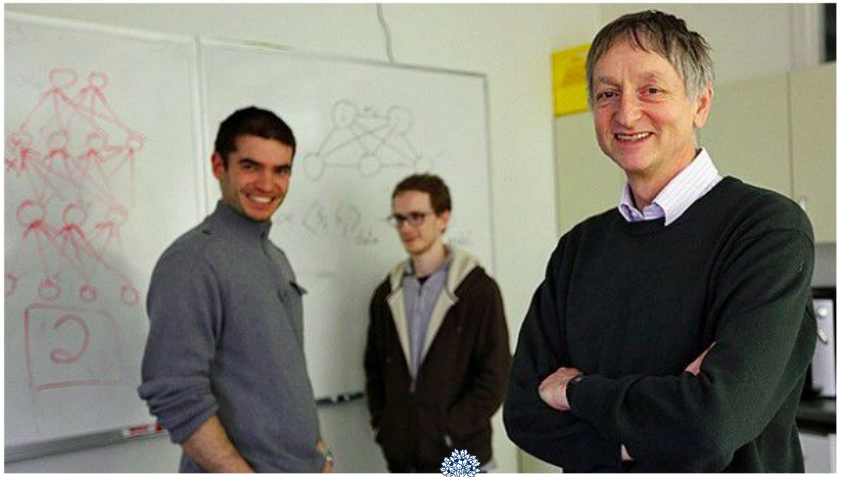
- Google Brain

Ilya Sutskever:

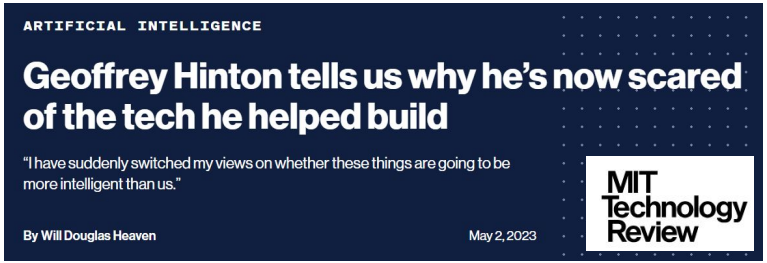
- Google Brain
- Co-Fundador y chief scientist en OpenAI

Geoffrey Hinton:

- Google DeepMind
- Creador del Backward propagation algorithm



UNIVERSITY OF TORONTO



Momentos clave

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 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.

Google Brain

Esto lo veremos con mayor detalle en este curso

Momentos clave

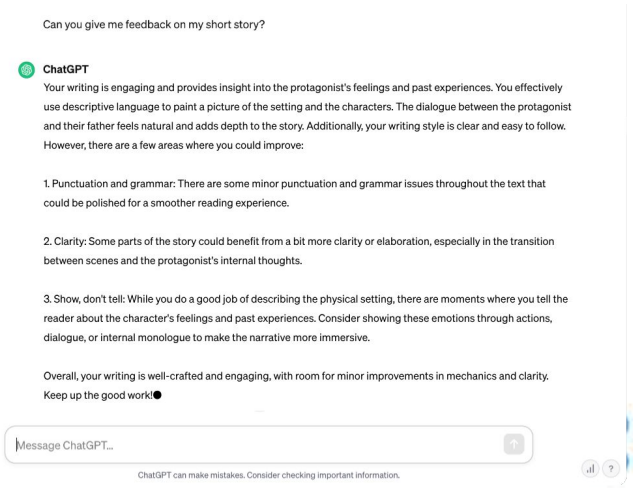
chatGPT (2022)

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



OpenAI

Microsoft

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.

DeepMind

Google

2010, Londres.

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

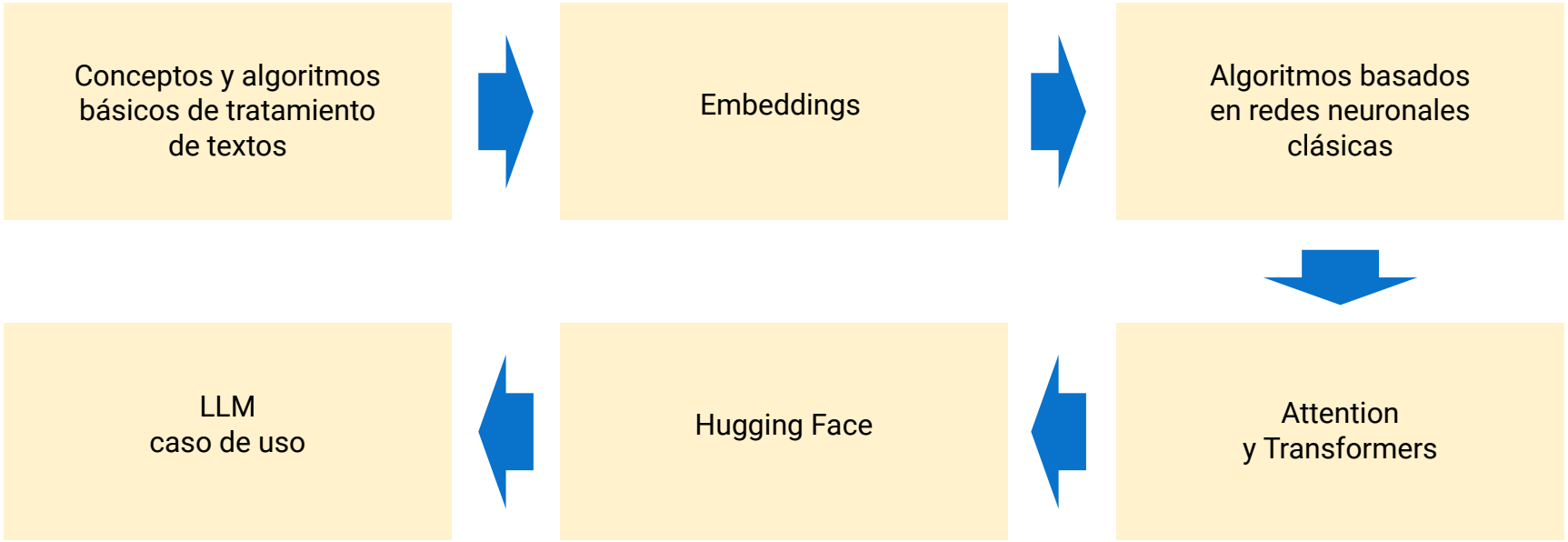
- AlphaGo, AlphaFold
- T5, Lambda, Gemini

Adquisición por Google en 2014.

NVIDIA

- 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


Conceptos y algoritmos básicos de tratamiento de textos
Embeddings
Algoritmos basados en redes neuronales clásicas
Attention y Transformers
Hugging Face
LLM caso de uso

Entorno Virtual

En nuestro entorno virtual para NLP, instalaremos las siguientes librerías:

```
conda create -n nlp_course python=3.10.13
```

```
pip install transformers
pip install datasets
pip install torch
pip install accelerate
pip install scikit-learn
pip install tensorflow==2.15.0
pip install nltk
pip install bokeh
pip install gensim
pip install spacy
Pip install sentence-transformers==2.7.0
pip install faiss-cpu==1.7.4
Pip install PyPDF2==3.0.1
pip install openai
pip install llama-index==0.10.31
python -m spacy download es_core_news_sm
```

El comando de instalación de `pytorch` se configura en la web oficial en función de características hardware y software

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:	`pip3 install torch torchvision torchaudio`			

```
tensorflow>2.0
```

```
NVIDIA-SMI 546.33                 Driver Version: 546.33       CUDA Version: 12.3
+-----------------------------------------+----------------------+----------------------+
| GPU  Name                     TCC/WDDM  | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA GeForce RTX 2060      WDDM  | 00000000:01:00.0  On |                  N/A |
| N/A   45C    P8              6W /  80W |    251MiB /  6144MiB |      1%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    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_cw5n1h2txyewy\SearchApp.exe    N/A      |
|    0   N/A  N/A      8948    C+G   ...t.LockApp_cw5n1h2txyewy\LockApp.exe    N/A      |
|    0   N/A  N/A      9876    C+G   ...siveControlPanel\SystemSettings.exe    N/A      |
|    0   N/A  N/A     11956    C+G   ...CBS_cw5n1h2txyewy\TextInputHost.exe    N/A      |
|    0   N/A  N/A     13180    C+G   ...64__8wekyb3d8bbwe\CalculatorApp.exe    N/A      |
+---------------------------------------------------------------------------------------+
```

nvidia-smi

Datasets

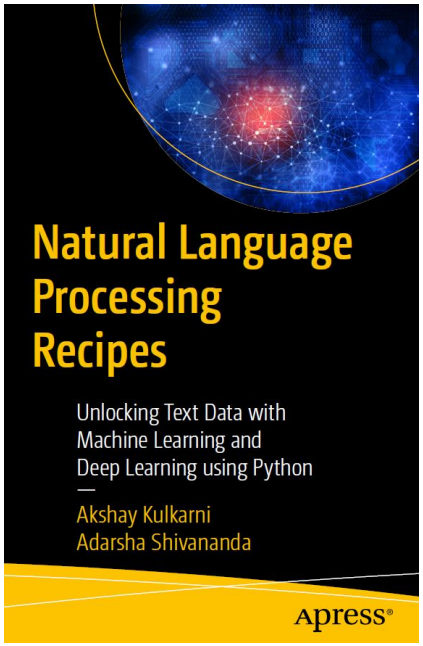
Phillip Keung, Yichao Lu, György Szarvas and Noah A. Smith. "*The Multilingual Amazon Reviews Corpus*". In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

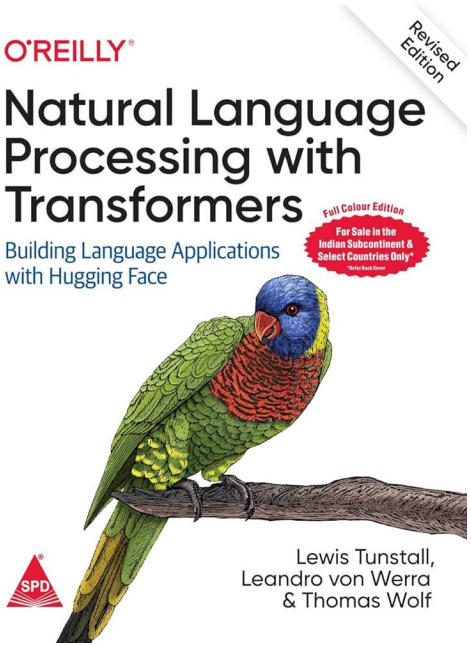
- https://raw.githubusercontent.com/eduardofc/data/main/amazon_sports.csv
- https://raw.githubusercontent.com/eduardofc/data/main/amazon_electronics.csv
- https://raw.githubusercontent.com/eduardofc/data/main/amazon_home.csv

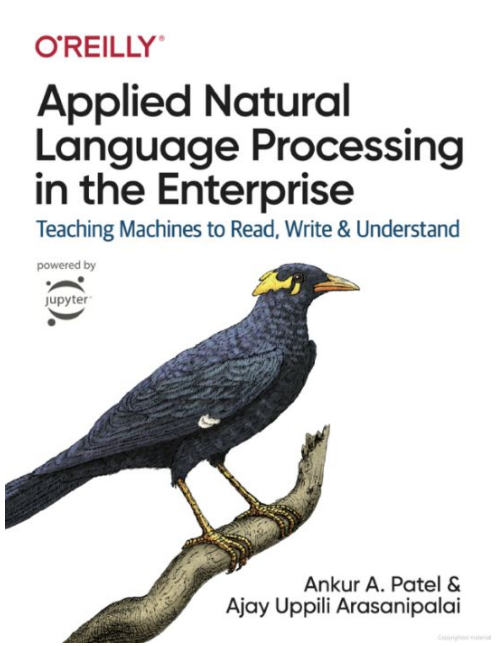
Anki cards para traducir frases entre español e inglés: https://www.manythings.org/anki/

- https://raw.githubusercontent.com/eduardofc/data/main/es_en.csv

Bibliografía







ntic master
revolucionamos la comunicación

UNIVERSIDAD COMPLUTENSE MADRID