

# PROYECTO 3

# RESÚMEN DE CURRICULUM

*Jenifer Leiva - Astrid Melo - Juanita Palacios*

# Introducción

La tecnología actual permite que muchos currículums sean enviados a las empresas a través de medios digitales, en consecuencia, el tiempo necesario para que los humanos extraigan información útil de estos documentos aumenta.

El resumen de texto funciona como una tarea de Procesamiento de Lenguaje Natural (NLP) permite obtener información clave mediante resúmenes coherentes y concisos de documentos extensos sin intervención humana.

Abstractive Text  
Summarization for Resumes  
With Cutting Edge NLP  
Transformers and LSTM by  
Mercon et al. (2023)

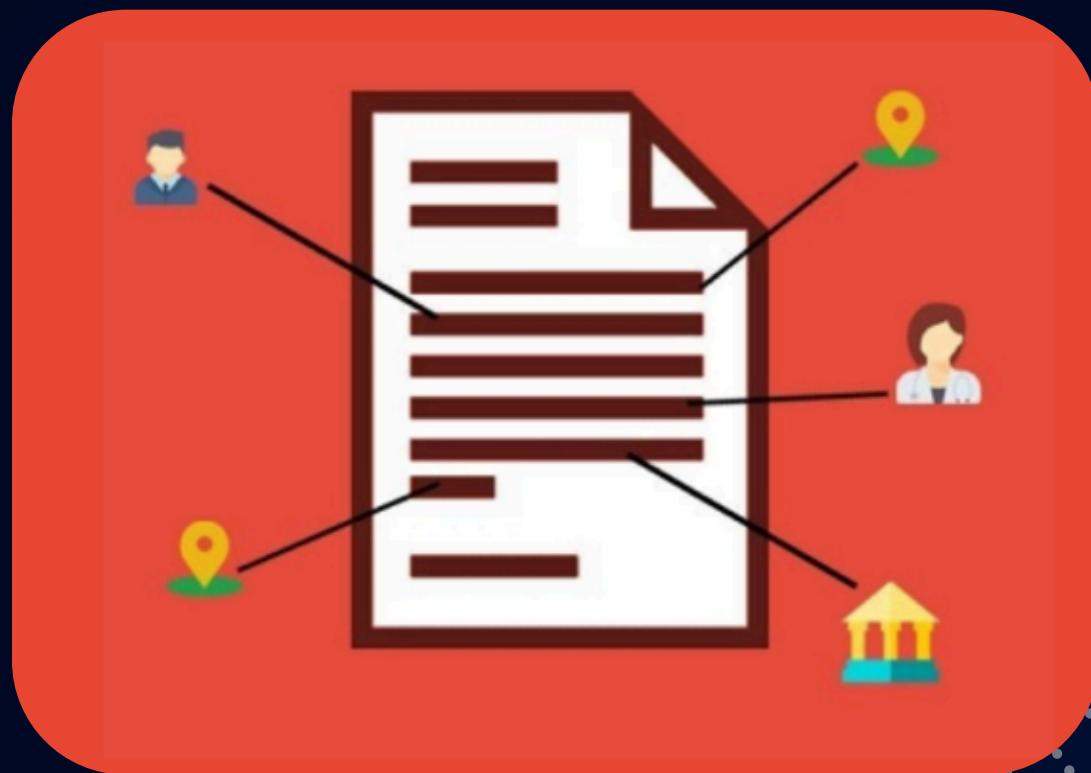
# DATASET

Resume Dataset compila más de 2.000 currículums del sitio web LiveCareer y los organiza en categorías de trabajo definidas en formato PDF.

Los trabajos que se pueden encontrar son: HR, Designer, Information-Technology, Teacher, Advocate, Business-Development, Healthcare, Fitness, Agriculture, BPO, Sales, Consultant, Digital-Media, Automobile, Chef, Finance, Apparel, Engineering, Accountant, Construction, Public-Relations, Banking, Arts and Aviation.

[Kaggle - Resume Dataset](#)

**Snehaan Bhawal**



# BART-LARGE-CNN

- BART (Bidirectional and Auto-Regressive Transformer)
- Fue publicado por **Facebook** en 2019
- Se preentrena usando **denoising autoencoding**
- Pre-trained **seq2seq** transformer model que combina las fortalezas de **BERT**, como encoder para comprender el contexto completo, y **GPT**, como decoder para la generación de texto
- BART Large CNN es una versión de BART que viene preentrenada con finetuning en el dataset **CNN/DailyMail**

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Lewis et al.

# Planteamiento

*Aplicar Fine-Tuning a un modelo BART para generar resúmenes automáticos de hojas de vida (en formato pdf)*

[https://colab.research.google.com/drive/1E5J57e7YsMrq488OlP-QjGFUWP9ihDH\\_#scrollTo=jD5RZp1Gx47\\_](https://colab.research.google.com/drive/1E5J57e7YsMrq488OlP-QjGFUWP9ihDH_#scrollTo=jD5RZp1Gx47_)

# Extracción y procesamiento de texto

1. Se descomprime el ZIP que contiene el dataset
2. Los archivos PDF se convierten a texto plano → PyMuPDF  
(libreria fitz)\
3. Se almacenan en un csv generado con columnas:  
filename, category y resume\_text

*El nuevo csv (all\_resume\_data.csv) Se convierte en el dataset base para todas las tareas posteriores.*



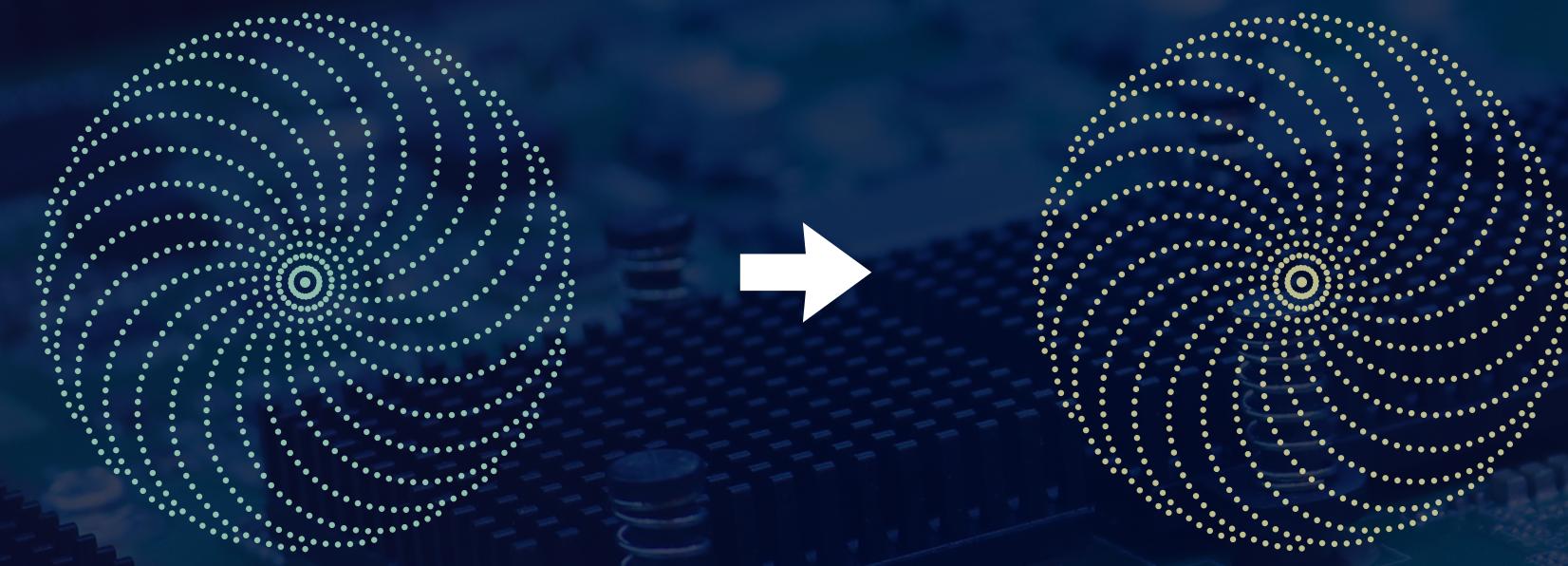
# Generación de resúmenes: Baseline

Se genera una nueva columna de resúmenes (generated\_summary) usando BART preentrenado, sin fine tuning aún.



# Preparación para Fine-Tuning

Se tokenizan los valores del data frame (data)  
de la siguiente manera:



**Input**  
resume\_text → texto puro a  
partir de pdf's

**Label (objetivo)**  
generated\_summary → resumen  
generado sin FT

# Dataset personalizado

Se construye un dataset personalizado en  
*PyTorch* para agrupar los tensores para  
posterior/ alimentar al modelo

- input ids
- attention mask
- labels

# Fine-Tuning

Se realiza Fine-Tuning de BART aplicando *Forward* y *Backward Pass*.

el modelo genera  
predicciones a partir  
de input\_ids y  
compara con labels.



se retropropaga el  
error y se actualizan  
los pesos con  
optimizer.step().

# Generación de resúmenes con el modelo Fine-Tuned

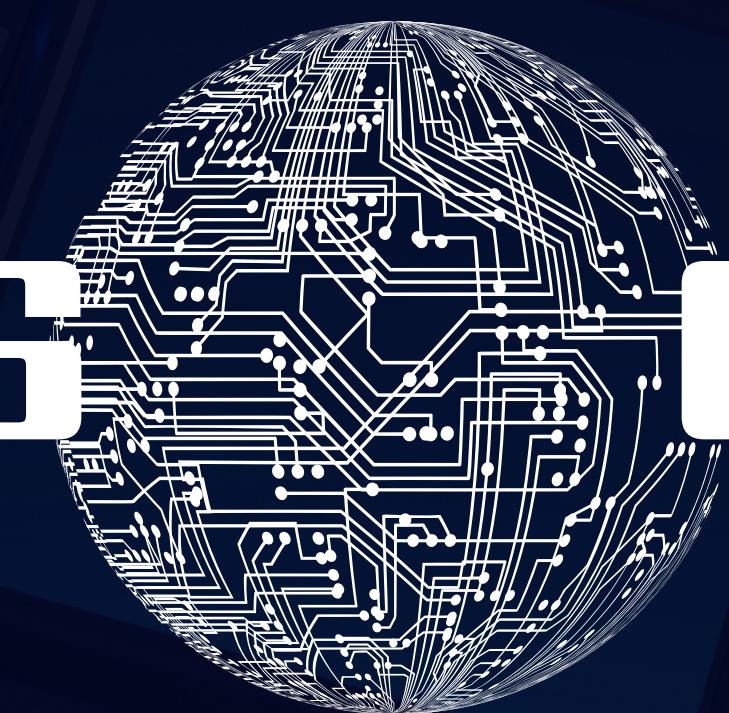
El modelo entrenado se pone en modo evaluación y genera nuevos resúmenes *más adaptados al dominio*.

[https://colab.research.google.com/drive/1hRHSyxD\\_GMJPPDeXMQLZk5PGD0N19F95?  
authuser=1#scrollTo=tJwWI7EzLLqG](https://colab.research.google.com/drive/1hRHSyxD_GMJPPDeXMQLZk5PGD0N19F95?authuser=1#scrollTo=tJwWI7EzLLqG)

[https://drive.google.com/drive/folders/1sOs2c2xGD2gU7NU5sYn9xWJW3KFitej5?  
usp=drive\\_link](https://drive.google.com/drive/folders/1sOs2c2xGD2gU7NU5sYn9xWJW3KFitej5?usp=drive_link)

# Referencias

- Bhawal, S. (2021). Resume dataset. Kaggle. <https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset/data>
- Dharrao, D., Mishra, M., Kazi, A., Pangavhane, M., Pise, P., & Bongale, A. M. (2024). Summarizing business news: Evaluating BART, T5, and PEGASUS for effective information extraction. Research in Artificial Intelligence. <https://doi.org/10.18280/ria.380311>
- Dafda, C. (n.d.). Text summarization using Facebook BART-large-CNN. Techblog GeekyAnts. <https://techblog.geekyants.com/text-summarization-using-facebook-bart-large-cnn>
- Hugging Face. (2020). BART — Transformers documentation. Hugging Face. [https://huggingface.co/docs/transformers/model\\_doc/bart](https://huggingface.co/docs/transformers/model_doc/bart)
- 
- Mercan, Ö., Cavsak, S., Deliahmetoglu, A., & Tanberk, S. (2023). Abstractive text summarization for resumes with cutting edge NLP transformers and LSTM. arXiv. <https://doi.org/10.48550/arXiv.2306.13315>



**MUCHAS GRACIAS**