**NAME: Nicolás Echeverri Rojas**

**Instructions:**

In points 1 and 2, you will be asked to develop a model and a Python script. For the submission of the source code, please create a GitHub repository and add all the source codes used for exploration, training, testing, and the computer vision script. If the trained model artifact is too large and doesn't fit in the repository, kindly deliver it through a Google Drive link. Finally, ensure that the repository contains a README file explaining the content of each source file and specifying which corresponds to each point of the practical test.

1. **Natural Language Processing**

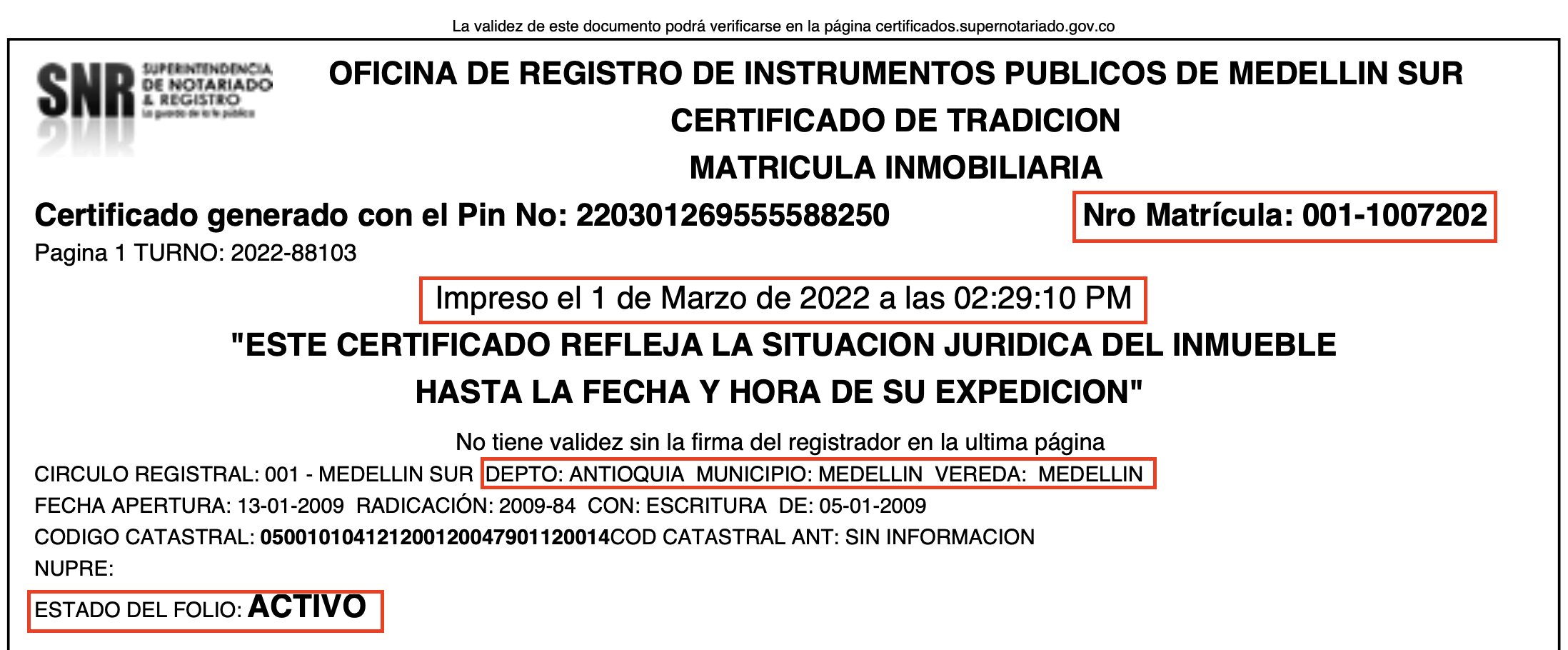
Based on the Diplomacy dataset (<https://sites.google.com/view/qanta/projects/diplomacy?pli=1>) train a deep learning model to predict if a message is a lie or not.

* This model must be trained with transfer learning using a Python framework.
* The model must be a transformer architecture and **you can’t use allenlp library** as it was used on the paper’s code.
* Compare the model you trained with the benchmark provided on the ACL’20 paper (Peskov et al.).

1. **Computer Vision**

In the "files" folder, you have two PDF files corresponding to the property certificates of two Colombian real estate properties. Likewise, for each .pdf file, there is a .json file representing the results of applying an OCR model with AWS Textract to the first page of each PDF. These .json files contain the extracted texts and coordinates from each document.

Based on these inputs, the task is to develop a Python code that receives the JSON file route, loads it, and extracts the Registration Number (Nro Matrícula), the print date (in YYYY-MM-DD format), the department, municipality, and locality (vereda) of the certificate or property, as well as the status of the folio. This information should be extracted from the .json file corresponding to the first page of each document.



The code should be singular and generic, meaning it should work for either of the two provided certificates. During the development evaluation, additional certificates will be provided to validate whether the code generalizes well for the extraction logic with other certificates having the same structure and arrangement of the mentioned fields but different values.

For more information on how AWS Textract delivers OCR, you can visit:

<https://docs.aws.amazon.com/textract/latest/dg/API_DetectDocumentText.html>

1. **Architecture and AWS**

Let's imagine you have a machine learning model, and you want to create a cloud architecture to serve this model through an API.

Please, create a simple architecture diagram on <http://app.diagrams.net> (draw.io) that utilizes the AWS resources you consider suitable for this requirement.

**Solución:**

**Explicación diagrama**

**Amplify:** Es el servicio encargado de crear y alojar la interfaz gráfica por la cual los usuarios podrán interactuar con el modelo.

**Cognito:** Servicio por el cual se podrá controlar el acceso mediante creación de usuarios.

**API Gateway:** Endpoint que al ser llamado desencadena AWS Lambda.

**Lambda:** Es el servicio que al recibir la petición de API Gateway acciona el endpoint del modelo en Sagemaker además de que almacena en DynamoDB

**Sagemaker:** Servicio el cual contiene el endpoint del modelo encapsulado previamente entrenado y probado.

**DynamoDB:** Base de datos no relacional en la cual se va a almacenar diferente información del usuario, del input y output del modelo con el fin de posterior seguimiento de métricas y reentrenamiento.

**IAM:** Manejo de permisos. En este caso es para poder escribir en la base de datos y llamar al endpoint de Sagemaker.

Diagrama

Descripción generada automáticamente