NAME:

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 or **free** LLMs 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.).

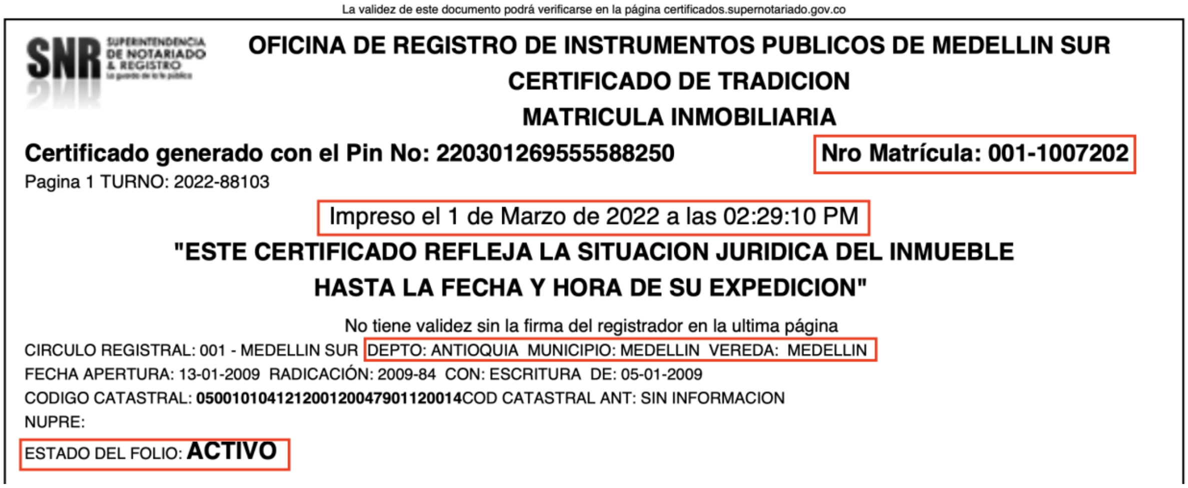
Grading Criteria for NLP Task:

- Analysis, EDA, and Explanation (40%):
  - Perform a thorough analysis prior to tackling the exercise.
  - Conduct Exploratory Data Analysis (EDA).
  - Explain your approach in a PowerPoint presentation (ppt).
- Implementation and Documentation (40%):
  - Diagram the implemented code.
  - Document the code and explain it in a PowerPoint presentation (ppt).
- Metric Achievement (20%):
  - Achieve a satisfactory metric on the exercise, specifically a score above 0.7.

2. 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.



La validez de este documento podrá verificarse en la página certificados.supernotariado.gov.co

OFICINA DE REGISTRO DE INSTRUMENTOS PUBLICOS DE MEDELLIN SUR

CERTIFICADO DE TRADICION

MATRICULA INMOBILIARIA

Certificado generado con el Pin No: 220301269555588250 **Nro Matrícula: 001-1007202**

Pagina 1 TURNO: 2022-88103

Impreso el 1 de Marzo de 2022 a las 02:29:10 PM

"ESTE CERTIFICADO REFLEJA LA SITUACION JURIDICA DEL INMUEBLE

HASTA LA FECHA Y HORA DE SU EXPEDICION"

No tiene validez sin la firma del registrador en la ultima página

CIRCULO REGISTRAL: 001 - MEDELLIN SUR DEPTO: ANTIOQUIA MUNICIPIO: MEDELLIN VEREDA: MEDELLIN

FECHA APERTURA: 13-01-2009 RADICACIÓN: 2009-84 CON: ESCRITURA DE: 05-01-2009

CODIGO CATASTRAL: **0500101041212001200479011200014**COD CATASTRAL ANT: SIN INFORMACION

NUPRE:

ESTADO DEL FOLIO: **ACTIVO**

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

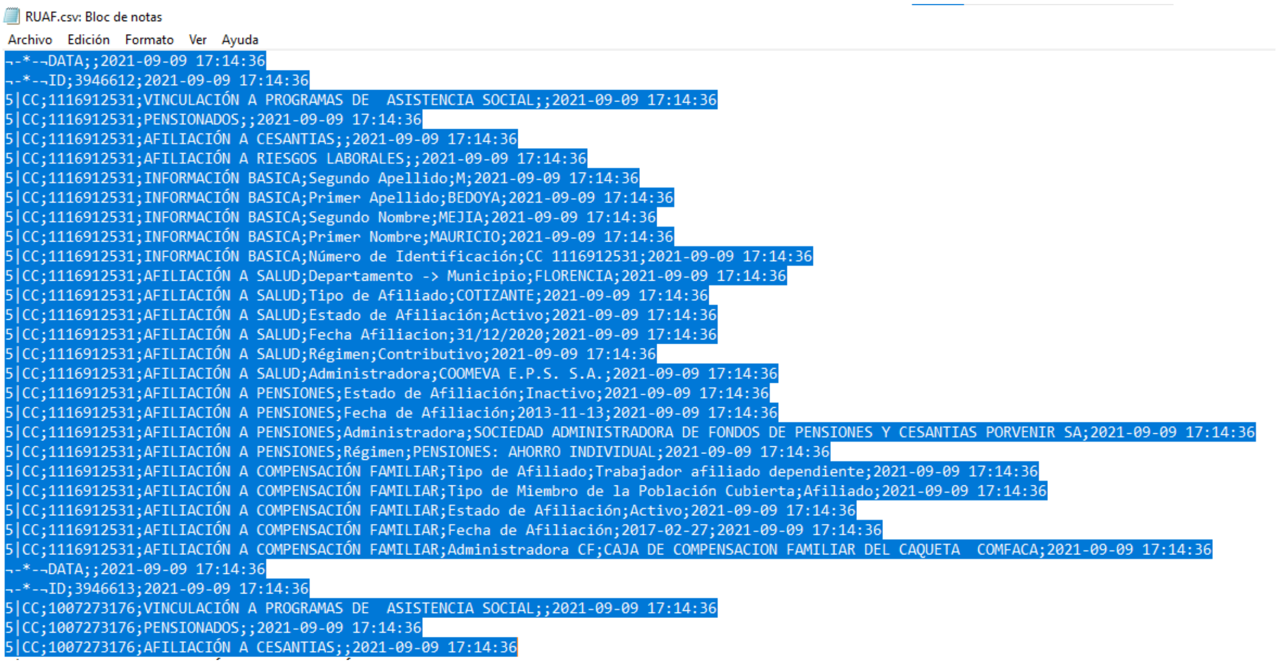
Grading Criteria for Computer Vision Task:

- Explanation and Approach (20%):
  - Provide an explanation on how you approached the exercise.
  - Present your approach in a PowerPoint presentation (ppt).
- Implementation and Documentation (50%):
  - Document the solution.
  - Diagram the code.
  - Explain the solution in a PowerPoint presentation (ppt).
- Result Accuracy (30%):
  - Achieve satisfactory results with the extraction logic.

3. Programming or SQL:

The following problem can be solved using only SQL, using Python or SQL + Python and you can use the libraries and functions that each tool provides.

For the attached file RUAF.csv there is the following structure:



```
¬-*-¬DATA;;2021-09-09 17:14:36
¬-*-¬ID;3946612;2021-09-09 17:14:36
5|CC;1116912531;VINCULACIÓN A PROGRAMAS DE  ASISTENCIA SOCIAL;;2021-09-09 17:14:36
5|CC;1116912531;PENSIONADOS;;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A CESANTIAS;;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A RIESGOS LABORALES;;2021-09-09 17:14:36
5|CC;1116912531;INFORMACIÓN BASICA;Segundo Apellido;M;2021-09-09 17:14:36
5|CC;1116912531;INFORMACIÓN BASICA;Primer Apellido;BEDOYA;2021-09-09 17:14:36
5|CC;1116912531;INFORMACIÓN BASICA;Segundo Nombre;MEJIA;2021-09-09 17:14:36
5|CC;1116912531;INFORMACIÓN BASICA;Primer Nombre;MAURICIO;2021-09-09 17:14:36
5|CC;1116912531;INFORMACIÓN BASICA;Número de Identificación;CC 1116912531;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A SALUD;Departamento -> Municipio;FLORENCIA;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A SALUD;Tipo de Afiliado;COTIZANTE;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A SALUD;Estado de Afiliación;Activo;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A SALUD;Fecha Afiliacion;31/12/2020;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A SALUD;Régimen;Contributivo;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A SALUD;Administradora;COOMEVA E.P.S. S.A.;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A PENSIONES;Estado de Afiliación;Inactivo;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A PENSIONES;Fecha de Afiliación;2013-11-13;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A PENSIONES;Administradora;SOCIEDAD ADMINISTRADORA DE FONDOS DE PENSIONES Y CESANTIAS PORVENIR SA;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A PENSIONES;Régimen;PENSIONES: AHORRO INDIVIDUAL;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A COMPENSACIÓN FAMILIAR;Tipo de Afiliado;Trabajador afiliado dependiente;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A COMPENSACIÓN FAMILIAR;Tipo de Miembro de la Población Cubierta;Afiliado;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A COMPENSACIÓN FAMILIAR;Estado de Afiliación;Activo;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A COMPENSACIÓN FAMILIAR;Fecha de Afiliación;2017-02-27;2021-09-09 17:14:36
5|CC;1116912531;AFILIACIÓN A COMPENSACIÓN FAMILIAR;Administradora CF;CAJA DE COMPENSACION FAMILIAR DEL CAQUETA  COMFACA;2021-09-09 17:14:36
¬-*-¬DATA;;2021-09-09 17:14:36
¬-*-¬ID;3946613;2021-09-09 17:14:36
5|CC;1007273176;VINCULACIÓN A PROGRAMAS DE  ASISTENCIA SOCIAL;;2021-09-09 17:14:36
5|CC;1007273176;PENSIONADOS;;2021-09-09 17:14:36
5|CC;1007273176;AFILIACIÓN A CESANTIAS;;2021-09-09 17:14:36
```

- The information must be organized horizontally (columns)
- The beginning of the record is identified with “¬-*-¬DATA”
- The information is organized by sections, but these should not remain in the final table
- If the record does not record information, create a field “marca_sin_informacion” with the indicator 0=with information, 1=without information.
- The fields must be with the snake_case notation
- Valid sections and related fields are as follows:

Sección/Campo
INFORMACIÓN BASICA;Primer Nombre
INFORMACIÓN BASICA;Segundo Nombre
INFORMACIÓN BASICA;Primer Apellido
INFORMACIÓN BASICA;Segundo Apellido
INFORMACIÓN BASICA;Sexo
AFILIACIÓN A SALUD;Administradora
AFILIACIÓN A SALUD;Régimen
AFILIACIÓN A SALUD;Fecha Afiliacion|Fecha de Afiliacion
AFILIACIÓN A SALUD;Estado de Afiliación
AFILIACIÓN A SALUD;Tipo de Afiliado

AFILIACIÓN A SALUD;Departamento -> Municipio
AFILIACIÓN A PENSIONES;Régimen
AFILIACIÓN A PENSIONES;Administradora
AFILIACIÓN A PENSIONES;Fecha de Afiliación
AFILIACIÓN A PENSIONES;Estado de Afiliación
AFILIACIÓN A RIESGOS LABORALES;Administradora
AFILIACIÓN A RIESGOS LABORALES;Fecha de Afiliación
AFILIACIÓN A RIESGOS LABORALES;Estado de Afiliación
AFILIACIÓN A RIESGOS LABORALES;Actividad Economica
AFILIACIÓN A RIESGOS LABORALES;Municipio Labora
AFILIACIÓN A COMPENSACIÓN FAMILIAR;Administradora CF
AFILIACIÓN A COMPENSACIÓN FAMILIAR;Fecha de Afiliación
AFILIACIÓN A COMPENSACIÓN FAMILIAR;Estado de Afiliación
AFILIACIÓN A COMPENSACIÓN FAMILIAR;Tipo de Miembro de la Población Cubierta
AFILIACIÓN A COMPENSACIÓN FAMILIAR;Tipo de Afiliado
AFILIACIÓN A COMPENSACIÓN FAMILIAR;Municipio Labora
AFILIACIÓN A CESANTIAS;Administradora
AFILIACIÓN A CESANTIAS;Fecha de Afiliación
AFILIACIÓN A CESANTIAS;Estado de Afiliación
AFILIACIÓN A CESANTIAS;Régimen
AFILIACIÓN A CESANTIAS;Municipio Labora
PENSIONADOS;Entidad que reconoce la pensión
PENSIONADOS;Fecha Resolución
PENSIONADOS;Estado
PENSIONADOS;Modalidad
PENSIONADOS;Número Resoluciòn Pension PG
PENSIONADOS;Tipo de Pensión
PENSIONADOS;Tipo de Pensionado
VINCULACIÓN A PROGRAMAS DE ASISTENCIA SOCIAL;Administradora
VINCULACIÓN A PROGRAMAS DE ASISTENCIA SOCIAL;Fecha de Vinculación
VINCULACIÓN A PROGRAMAS DE ASISTENCIA SOCIAL;Estado de la Vinculación
VINCULACIÓN A PROGRAMAS DE ASISTENCIA SOCIAL;Estado del Beneficio
VINCULACIÓN A PROGRAMAS DE ASISTENCIA SOCIAL;Fecha Ultimo Beneficio
VINCULACIÓN A PROGRAMAS DE ASISTENCIA SOCIAL;Programa
VINCULACIÓN A PROGRAMAS DE ASISTENCIA SOCIAL;Ubicación de Entrega del Beneficio

Attached you will find a file called “muestra estructurada RUAF.xlsx” to guide you as to how the information should be organized.

Grading Criteria for Programming or SQL task:

- Attach code. – 30%
- Code documentation and clean code. – 70%