**Índice**

[**Qué tipo data hay 1**](#_ek9dmtk3eq0y)

[**principales fuentes 1**](#_476eykz59tmi)

[**Herramientas adicionales: 1**](#_1fx8kqdx6rpf)

[**proyectos qa y datasets 2**](#_itinos8stku0)

[**Pregunta test: 2**](#_mjk3wuon5k5c)

[**fuentes 2**](#_wirwvxvouesz)

[**Observaciones Generales (laura sofia) 4**](#_gc8y1a5qcw3a)

Tema: Inteliagencia artificial aplicada al PLN

Modalidad: Proyecto de Grado

Director: [RAUL ERNESTO GUTIERREZ REYES](mailto:raul.gutierrez@correounivalle.edu.co)

Idea: Chatbot que permita consultar información acerca de los informes de la comisión de la verdad.

## Qué tipo data hay

* informes generales
* informes casos
* contexto de labores hechas - principalmente en las páginas web

## principales fuentes

* prensa: <https://www.comisiondelaverdad.co/prensa?field_categoria_target_id=All>
* archivo: <https://archivo.comisiondelaverdad.co/>
* informes finales: <https://www.comisiondelaverdad.co/hay-futuro-si-hay-verdad>
* lista de informes casos: <https://www.comisiondelaverdad.co/analitica-de-datos-informacion-y-recursos#c5>

## 

## Herramientas adicionales:

playground para desarrollar ias generativas:

<https://dashboard.cohere.com/welcome/register>

<https://huggingface.co/datasets/PlanTL-GOB-ES/SQAC>

<https://huggingface.co/datasets/rajpurkar/squad>

<https://huggingface.co/jorge-henao>

<https://medium.com/@kshitiz.sahay26/fine-tuning-llama-2-for-news-category-prediction-a-step-by-step-comprehensive-guide-to-fine-tuning-48c06dee28a9>

<https://github.com/dccuchile/beto?tab=readme-ov-file#references>

<https://github.com/KGQA/qald_9_plus>

blog <https://jaimesendraberenguer.medium.com/bert-para-responder-preguntas-sobre-squad-2-0-en-espa%C3%B1ol-5842748f051a>

como

estructurar un dataset

<https://www.flowrite.com/blog/dataset-engineering-llm-finetuning>

## 

## proyectos qa y datasets

<https://huggingface.co/consciousAI/question-answering-roberta-base-s-v2>

<https://huggingface.co/deepset/roberta-base-squad2>

<https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-sqac>

<https://huggingface.co/mrm8488/spanish-t5-small-sqac-for-qa>

<https://huggingface.co/p1gm1/beto-col-qa>

<https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne-sqac>

<https://github.com/ccasimiro88/TranslateAlignRetrieve/blob/master/SQuAD-es-v2.0/dev-v2.0-es.json>

<https://huggingface.co/deepset>

<https://huggingface.co/datasets/bertin-project/alpaca-spanish>

<https://huggingface.co/datasets/jorge-henao/ask2democracy-cfqa-salud-pension>

## Pregunta test:

¿cuales son las principales funciones de la comisión?

**context:**

**“**

La Comisión de la Verdad en Colombia, creada en el marco del Acuerdo de Paz de 2016, ha adelantado un arduo trabajo para esclarecer los hechos y patrones de graves violaciones a los derechos humanos y del Derecho Internacional Humanitario ocurridos durante el conflicto armado en el país. Entre sus principales labores se destacan:

Recepción de testimonios: La Comisión ha recibido más de 14.000 testimonios de víctimas, victimarios, agentes del Estado y otros actores del conflicto, recopilando un acervo invaluable de relatos que permiten comprender la complejidad del conflicto y sus múltiples dimensiones.

Investigación: A partir de los testimonios y otras fuentes de información, la Comisión ha realizado investigaciones exhaustivas sobre diferentes temáticas, incluyendo masacres, desapariciones forzadas, desplazamiento forzado, reclutamiento de menores, violencia sexual, entre otras.

Elaboración del Informe Final: La Comisión se encuentra en la fase final de elaboración de su Informe Final, en el que se presentarán los hallazgos de sus investigaciones, un análisis del conflicto armado y recomendaciones para la no repetición. El Informe Final será presentado al país el 28 de junio de 2024.

Contribución a la construcción de paz: La Comisión ha desarrollado diversas iniciativas para contribuir a la construcción de paz en Colombia, incluyendo programas de pedagogía para la paz, promoción de la convivencia y la reconciliación, y apoyo a las víctimas en su proceso de reparación.

“

## fuentes

<https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=80633>

<https://bibliotecadigital.univalle.edu.co/server/api/core/bitstreams/d0ee9e19-2069-467e-991e-3864be3e5a27/content?authentication-token=eyJhbGciOiJIUzI1NiJ9.eyJlaWQiOiI1NjFhZDZlYy04YjBlLTQyYTktYjBkMC1hMmQwNWQzOTg4ZmEiLCJzZyI6WyIxYjQwYzg4OC1hOTg2LTQ0ZWMtOWFjZi02OGJiMTU3MjFjODAiLCIxYjQwYzg4OC1hOTg2LTQ0ZWMtOWFjZi02OGJiMTU3MjFjODAiXSwiYXV0aGVudGljYXRpb25NZXRob2QiOiJvcmNpZCIsImV4cCI6MTcxMjE3NDc3N30.ZAxYVGqakfx0cYUiAHzQQ4K79zofpxVhSZyFrqzYpa0>

<https://academic.oup.com/jamia/article/18/5/544/829676>

<https://arxiv.org/abs/2402.06196>

<https://arxiv.org/pdf/2402.06196v2.pdf>

<https://arxiv.org/abs/2401.17197>

<https://repository.unilibre.edu.co/handle/10901/27487>

<https://ieeexplore.ieee.org/abstract/document/8884203>

<https://arxiv.org/abs/1706.03762>

pln y los llm

<https://ieeexplore.ieee.org/abstract/document/9998298>

chatbots y fine-tunning

<https://arxiv.org/pdf/2402.16142>

pln comision

anex nlp

<https://www.comisiondelaverdad.co/sites/default/files/descargables/2022-08/03_Anexo_NLP_07312022_4%2C9KB.pdf>

fuente de imagen

<https://www.deepset.ai/blog/llm-finetuning>

datasets definicion

<https://uniandes-co.libguides.com/administracion/datasets>

<https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/meet.14504701240> p 2

fuente de imagen

<https://pro.arcgis.com/es/pro-app/latest/help/data/feature-datasets/feature-datasets-in-arcgis-pro.htm>

caracteristicas de un buen dataset

<https://link.springer.com/chapter/10.1007/978-3-030-67024-5_4> p55

historia del los QA

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3996229> p1

fuente imagen

<https://x.com/hwchase17/status/1649428297816711174>

dataset henao

<https://huggingface.co/datasets/jorge-henao/historias_conflicto_colombia>

proyecto henao

<https://huggingface.co/jorge-henao/gpt2-small-spanish-historias-conflicto-col>

sqac

<https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne-sqac>

plantgob

<https://huggingface.co/PlanTL-GOB-ES>

redes neronales def

<https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414>

hacer los pendientes del trello

## Observaciones Generales (laura sofia)

Haga énfasis en la problemática no solo desde el contexto técnico, es importante que profundice en el problema raíz sobre el conflicto armado y el proceso de paz

• cuanto años duro el conflicto armado en Colombia? cuantas víctimas? que tipo de delitos se cometieron? quienes fueron los más afectados?

Consultar literatura del profesor Boris Salazar Trujillo Univalle y la profesora Raquel Méndez Villamizar de la UIS, entre otros.

• como fue el proceso de paz? cuanto duro? involucrados? consultar literatura, escuche los programas de diana Uribe sobre el proceso de paz

Seguridad:

Es importante establecer políticas y procedimientos de seguridad de la información para proteger los datos y recursos del sistema. Se debe implementar un sistema de monitoreo para detectar y responder a posibles problemas de seguridad, rendimiento o disponibilidad.

1. Protección de datos sensibles:

• ¿Qué medidas se implementan para proteger la información personal y confidencial contenida en los informes de la Comisión de la Verdad que se utiliza para entrenar y operar el sistema de QA?

• ¿Cómo se garantiza que el acceso a los datos sensibles esté restringido a usuarios autorizados y que se cumplan las normas de privacidad de datos?

• ¿Se han realizado pruebas de seguridad exhaustivas para identificar y mitigar posibles vulnerabilidades en el sistema, como ataques de inyección SQL, cross-site scripting (XSS) y otros ataques cibernéticos?

2. Gestión de accesos:

• ¿Qué mecanismos de autenticación y autorización se utilizan para controlar el acceso de usuarios y administradores al sistema de QA?

• ¿Cómo se gestionan los diferentes niveles de acceso en el sistema para garantizar que los usuarios solo puedan acceder a la información y funcionalidades que les corresponden?

• ¿Se implementan medidas de seguridad adicionales para proteger las cuentas de usuario, como contraseñas seguras, autenticación de dos factores y registros de auditoría?

Escalabilidad:

1. Arquitectura escalable:

• ¿Cómo se ha diseñado la arquitectura del sistema de QA para permitir un crecimiento en el número de usuarios y el volumen de datos sin afectar el rendimiento o la disponibilidad del sistema?

• ¿Qué tecnologías y herramientas se utilizan para garantizar la escalabilidad del sistema, como la computación en la nube, el almacenamiento distribuido y las bases de datos escalables?

• ¿Se ha considerado la implementación de mecanismos de autoscalado para ajustar los recursos del sistema en función de la demanda?

2. Disponibilidad y confiabilidad:

• ¿Qué medidas se toman para garantizar la alta disponibilidad del sistema de QA, incluyendo planes de contingencia ante fallos de hardware, software o red?

• ¿Se implementan estrategias de redundancia y tolerancia a fallos para minimizar el impacto de posibles interrupciones del servicio?

• ¿Se realizan pruebas de carga y rendimiento para evaluar la capacidad del sistema de manejar un aumento en la demanda de usuarios y solicitudes?

Desarrollo sostenible y Equidad, Diversidad e Inclusión:

• ¿El desarrollo del trabajo contribuye al cumplimiento de los Objetivos de Desarrollo Sostenible (ODS) y promueve prácticas sostenibles? ¿A cuáles de los ODS contribuye?

• ¿Cómo va a evaluar el impacto diferenciado del sistema en diversos grupos demográficos, especialmente aquellos históricamente discriminados como las mujeres, la población racializada, étnica, la comunidad LGBTI+Q o personas con alguna discapacidad?

• ¿Qué medidas se han implementado para asegurar que el desarrollo del proyecto esté libre de sesgos que perpetúan las desigualdades en nuestra sociedad?

• ¿En el planteamiento del proyecto se ha tenido en cuenta particularidades relacionadas con la salud, seguridad o bienestar de mujeres y otros grupos, tanto en la fase de diseño como en el desarrollo del sistema?

• ¿De qué manera el desarrollo del proyecto contribuye a la reducción de las brechas de género y otras desigualdades en su campo de estudio?

## Avances hasta el momento

* Descarga de los pdfs:
* faltan 22. algunos son muy pesados. otros son un excel y otros no tiene archivo
* Investigación sobre los modelos SQAC:
  + tema

- context

- question

- answers[]

-text

Cada informe puede ser un tema.

-

* Script para generar preguntas.
  + hasta ahora he generado 20 manual
* Trabajo con logchain
* Crear un repositorio para el dataset - ya está hecho
* Evaluación de los modelos.

**Siguientes tareas**

* Seguir generando ejemplos
* Seguir buscando un asesor
* revisar los informes con los que no se puede tratar