

Proyecto Final de Bootcamp: Big Data, Machine Learning y AI

Equipo : Data Divers

Amado Martínez | Jorge Arrojo | Juan David Pardo | Rolando Rodriguez | Santiago Bedoya



KEEPCODING
Tech School



Bienvenidos y Bienvenidas

Primero un poco de contexto...



Presentación

Base de Datos:

Utilizamos el dataset de Harvard University Dataverse,
Específicamente, United Nations General Assembly Voting Data.

Resumen de Investigación:

Enfocamos nuestro estudio en el análisis de votaciones nominales de la Asamblea General de la ONU desde 1946 hasta 2024, destacando temas globales críticos:

- Conflicto palestino (19% de los votos)
- Armas nucleares (13%)
- Control de armas y desarme (16%)
- Colonialismo (18%)
- Derechos humanos (17%)
- Desarrollo económico (9%)



Presentación

Enfoque Específico:

Nos centramos en las votaciones sobre el conflicto palestino y las opiniones consultivas relacionadas, **en particular la opinión de 2004 (A/ES-10/273)** sobre las consecuencias jurídicas de la construcción del muro israelí en territorio palestino.

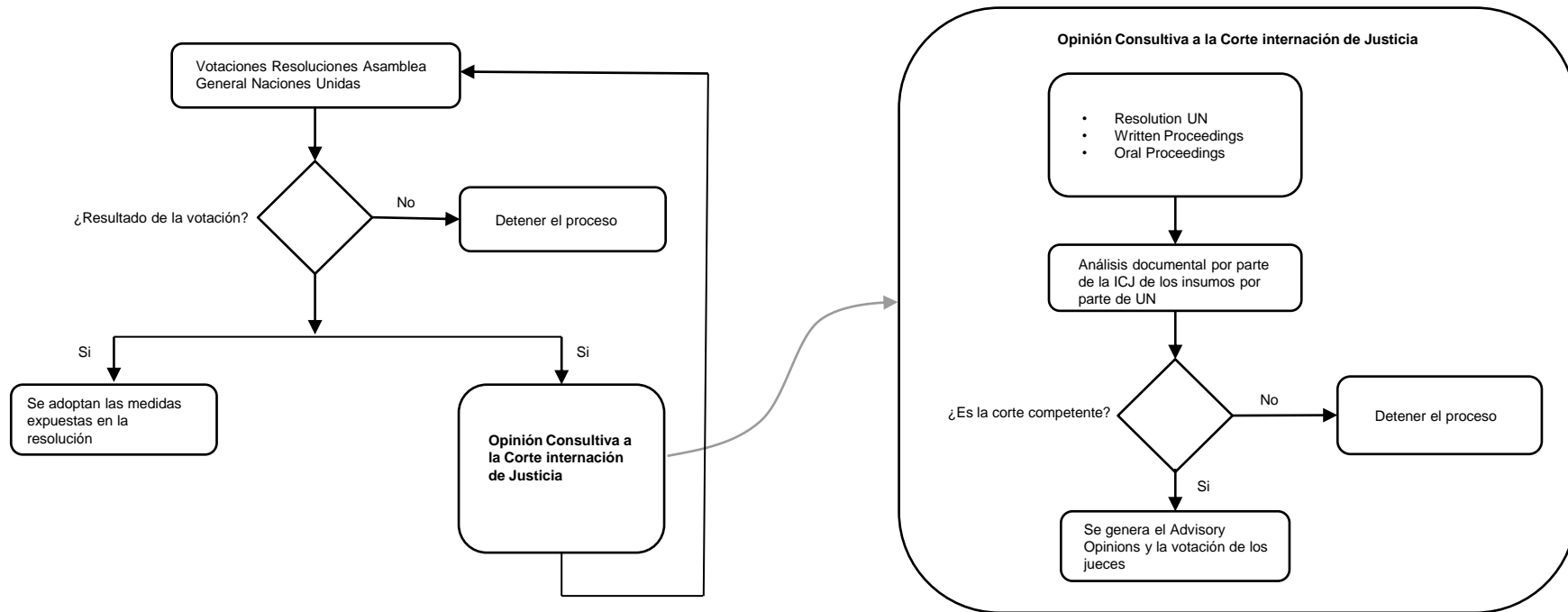
Metodología de Análisis:

Aplicamos NLP a documentos legales y opiniones consultivas, eligiendo el modelo BERT para el análisis de sentimientos.

Además de un modelo de Preguntas y Respuestas para inferir resultados de votaciones futuras, como la **opinión consultiva en curso (A/Res/77/247)**.

Presentación

Diagrama del proceso de opinión consultiva desde la solicitud de UN





Análisis exploratorio

Descripción general:

Utilizamos Tableau para realizar un análisis exploratorio de las votaciones de la Asamblea General de la ONU, partiendo de datos almacenados en la carpeta Dataverse Files.

¿Qué procesamiento pudimos hacer gracias al análisis?

Unificación de nombres de países, unificación de formatos de votaciones y generación de registros para la resolución de 2024.

Caso de Estudio:

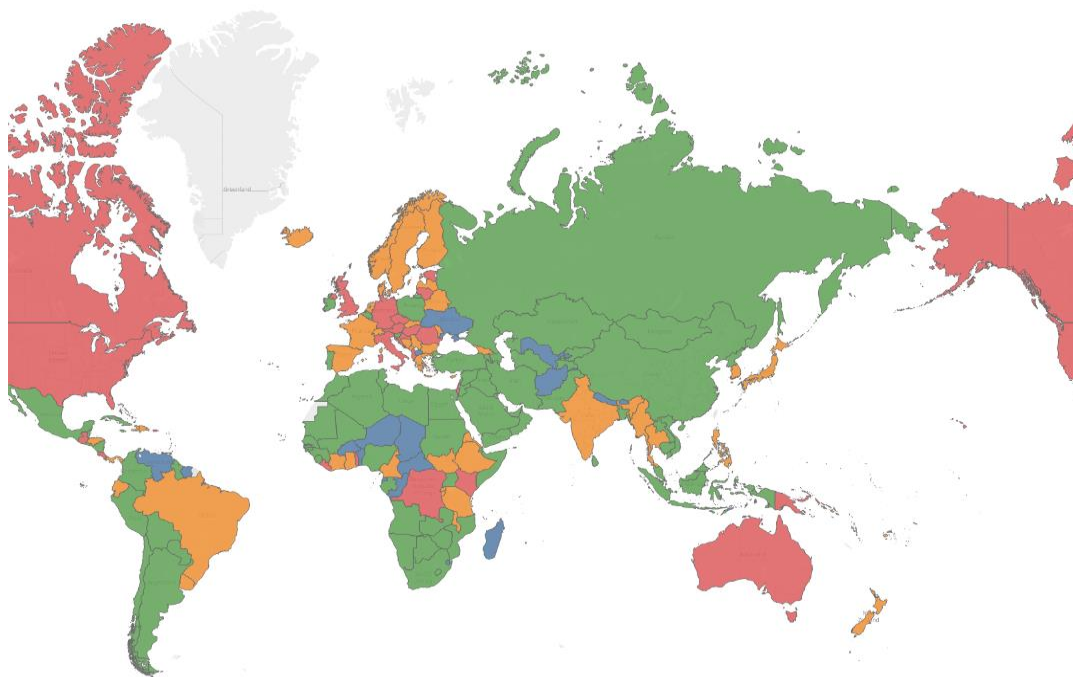
Análisis discriminativo de votaciones globales frente a votaciones de países miembro de CIJ

Análisis de votaciones para las opiniones consultivas del conflicto palestino

Este análisis facilita una comprensión detallada de las dinámicas de votación y su impacto en decisiones internacionales.

Este contenido proporciona una visión integral de las votaciones de la ONU.

Análisis exploratorio | Visualizaciones



Votaciones
miembros
permanentes
consejo de
seguridad.

CountryNameM..	
China	●
France	●
Russia	●
United Kingdom ..	●
United States of..	●

Votaciones
miembros no
permanentes del
consejo de
seguridad 2023.

CountryNameM..	
Albania	●
Brazil	●
Ecuador	●
Gabon	●
Ghana	●
Japan	●
Malta	●
Mozambique	●
Switzerland	●
United Arab Emi..	●

Votaciones
miembros
permanentes
consejo de
seguridad.

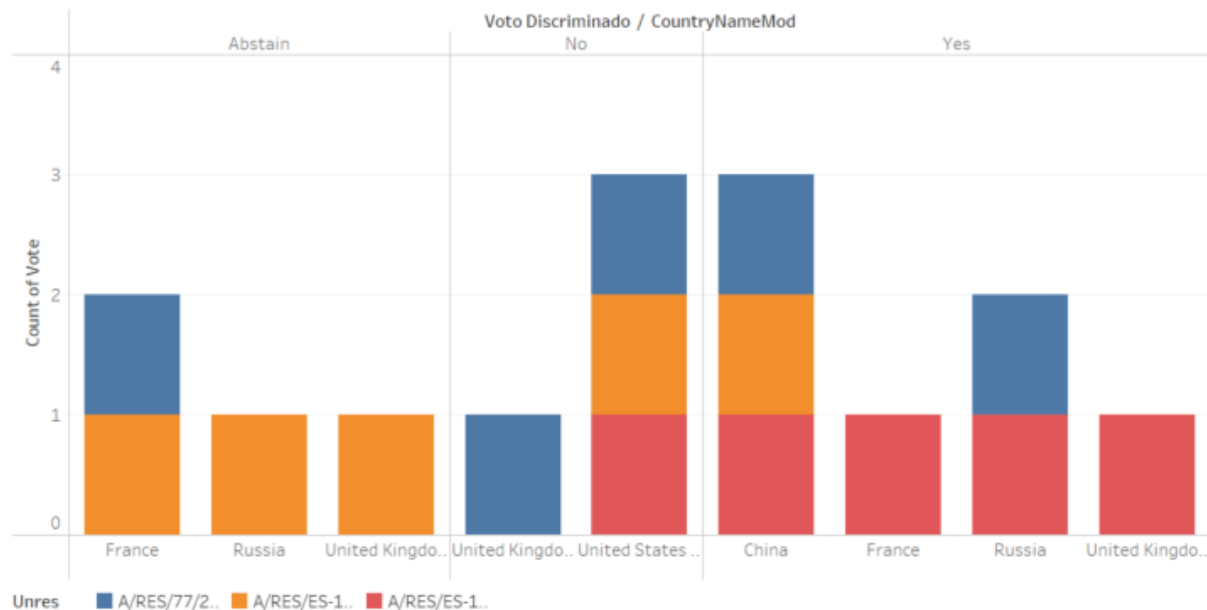
CountryNameM..	
China	●
France	●
Russia	●
United Kingdom ..	●
United States of..	●



Votaciones
miembros no
permanentes del
consejo de
seguridad 2023.

CountryNameM..	
Albania	●
Brazil	●
Ecuador	●
Gabon	●
Ghana	●
Japan	●
Malta	●
Mozambique	●
Switzerland	●
United Arab Emi..	●

Análisis exploratorio | Visualizaciones

La distribución de votaciones de los miembros permanentes del Consejo de Seguridad de las Naciones Unidas, en relación con las opiniones consultivas dirigidas a la Corte Internacional de Justicia sobre el conflicto palestino.



- 
- 
1. Partimos de bases de datos
 2. Se hizo un análisis exploratorio con esta base
 3. Se tomaron decisiones (decisión de hacer el NLP dirigido a las opiniones consultivas)
 4. Arquitectura
 5. Preprocesamiento (crawler, orals, advisory)
 6. Montó el modelo, primero un Regresión, Gradient, Regresión LSTM, DL (RNN), TF-IDF, BERT
 7. Evaluación de modelo (métricas y predicción positiva/negativa)
 8. Q & A

Presentación

Objetivo del Modelo:

Predecir el resultado de futuras votaciones relacionadas con el conflicto palestino, basándonos en análisis de sentimiento y respuestas a preguntas legales específicas.



**Naciones
Unidas**



**Corte Internacional
Justicia**



Documentación de Archivos del Proyecto

Datos y Ficheros:

Dividimos los datos y ficheros en dos categorías principales:

1) Votaciones de la Asamblea General de la ONU:

- Datos desde 1946 hasta 2014 provienen del Harvard University Dataverse y el repositorio United Nations General Assembly Voting Data.
- Datos desde 2014 obtenidos mediante técnicas de Crawling y Scrapping de la Biblioteca Digital de la ONU (United Nations Digital Library).

2) Modelo de NLP para la opinión consultiva de la CIJ:

- Datos de la Corte Internacional de Justicia, principalmente documentos en PDF de las opiniones consultivas de 2004 y 2023.



Documentación de Archivos del Proyecto

Datos y Ficheros:

Procesamiento y Estructuración de Datos:

- Implementamos un sistema de Crawling con Selenium y WebDriver para extraer URLs y un sistema de Scrapping con lxml para procesar y almacenar los datos para prepararlos en formatos JSON y CSV.
- Segmentamos y transformamos documentos PDF usando PyPDF2, permitiendo la organización de textos en formatos más manejables (texto plano y JSON).

Dificultades:

- Encontramos desafíos con los documentos de Oral Proceedings debido a la calidad de digitalización, lo que limitó el alcance inicial del dataset.



Arquitectura del DAaaS



Diseño y Estrategia del DAaaS:

- Creamos una arquitectura pensando en escalar el proyecto a un sistema que sea capaz de evolucionar en una solución que permita universalizar la informa que se trata en el ámbito de las Naciones Unidas y por eso diseñamos una arquitectura escalable.
- Para eso propusimos crear una aplicación web para mejorar el acceso y la difusión de información sobre asuntos tratados por la ONU, con capacidades de búsqueda, visualización, consulta y predicción de resoluciones basadas en documentos y recomendaciones. Incluir notificaciones en tiempo real sobre nuevas resoluciones, opiniones y votaciones. Además de un chat para realizar consultas sobre temas específicos que serán atendidos por IA Generativa



Arquitectura del DAaaS (Data as a Service)

Infraestructura y Almacenamiento:

- Google Cloud Storage para almacenamiento de datos y documentos
- Google Cloud SQL y Chroma DB para gestión de bases de datos
- Load Balancers y Firewalls para gestión de tráfico y seguridad

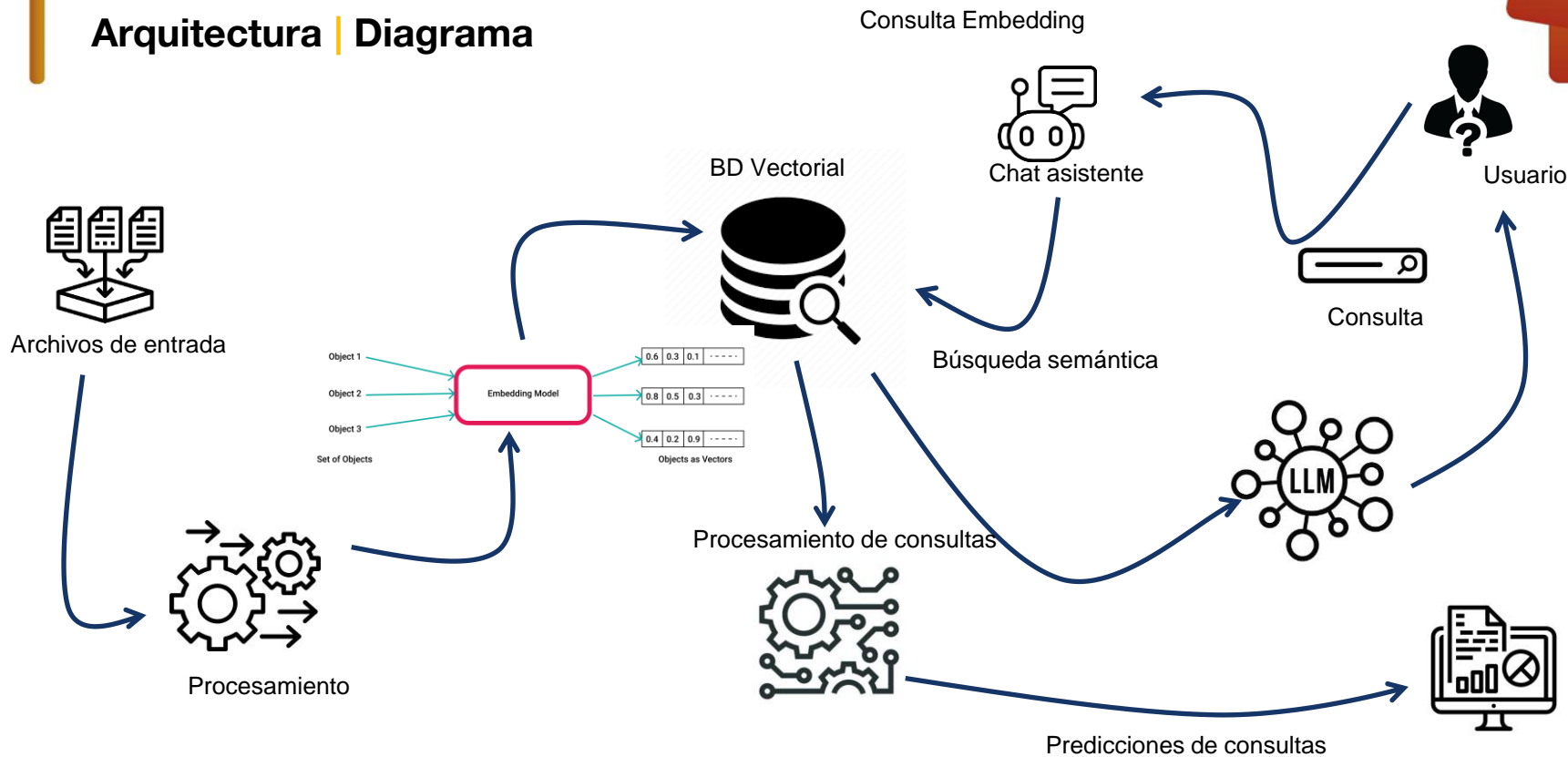
Procesamiento de Datos:

- Cloud Functions para ejecutar crawlers y scrappers
- Cloud Scheduler para tareas programadas
- DataPrep y BigQuery para Datawarehousing
- ElasticSearch y Kibana para análisis de logs

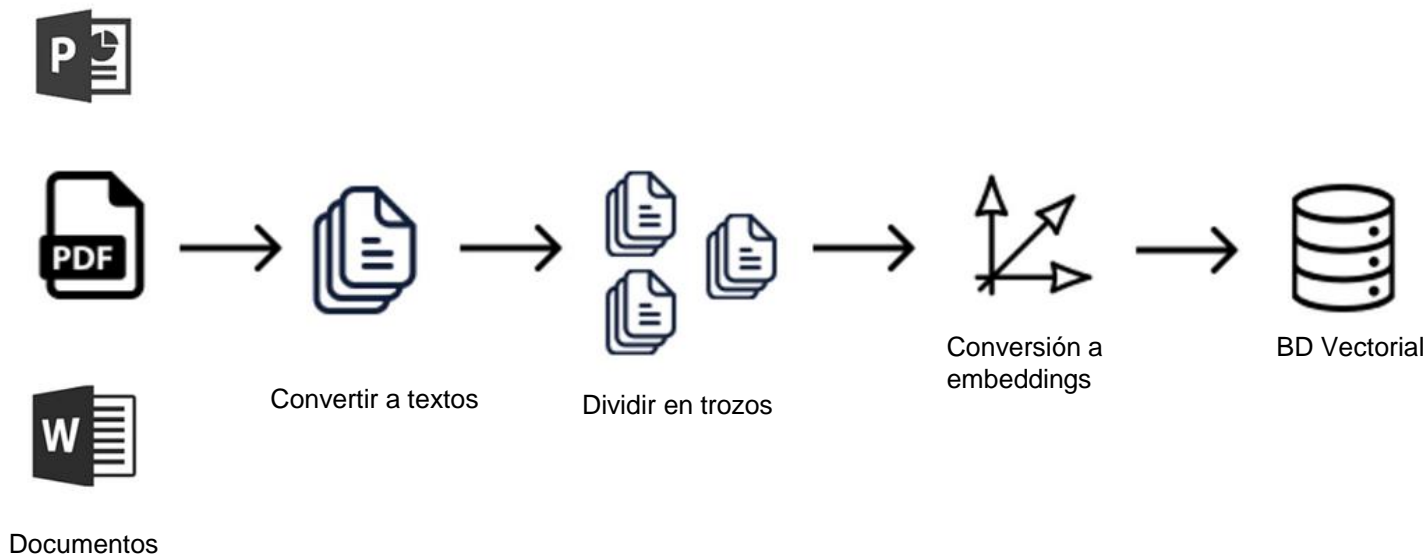
Notificaciones y Monitoreos:

- Firebase Cloud Messaging para notificaciones push
- Configuración de retención de logs para análisis detallado

Arquitectura | Diagrama



Arquitectura | Almacenamiento de documentos





Modelado

Análisis de Cardinalidad y Balanceo de Datos:

- Utilización de funciones y gráficos para evaluar la diversidad y distribución de las clases en el dataset, asegurando un balance adecuado para un análisis efectivo.

Ngramas y Nube de Palabras:

- Herramientas: Ngrams y Wordclouds.
- Objetivos: Identificar patrones de palabras y frases clave, verificar la calidad de la data para iteraciones de preprocesamiento, y facilitar la comprensión de los temas predominantes.

Word Embeddings con Word2Vec:

1. Inicialización y Entrenamiento:

- Definición de Parámetros: Tamaño de vector, contexto de ventana, frecuencia mínima de palabras.
- Construcción del Vocabulario: A partir de un corpus preprocesado y tokenizado.
- Entrenamiento del Modelo: Realizado durante 10 épocas para optimizar la representación vectorial del texto.



Modelado

Word Embeddings con Word2Vec:

2. Operaciones Post-Entrenamiento:

- Guardado y Carga: Modelo guardado como `w2v_sg_model.pkl` para reutilización futura sin reentrenamiento.
- Validación: Función `print_sim_words` para mostrar palabras semánticamente similares y validar la efectividad del modelo.

3. Visualización de Embeddings:

- Representación en 2D: Gráfica que muestra la agrupación semántica de palabras clave y sus similares para una interpretación visual clara de la semántica.

Modelado

Implementación de diferentes modelos de NLP.

A. **Modelo de Regresión Logística (Logistic Regression)** El modelo de Regresión Logística se utiliza para la clasificación binaria. Calcula la probabilidad de que una observación pertenezca a una clase utilizando una función logística.

B. **Modelo de Clasificación Gradient Boosting** El modelo de Clasificación Gradient Boosting es un algoritmo de conjunto que combina múltiples árboles de decisión más débiles para construir un modelo predictivo más fuerte.

C. **Modelo de Deep Learning RNN** El modelo utiliza una red neuronal recurrente LSTM para procesar secuencias de texto, seguido de una capa densa de salida con activación sigmoide para clasificación binaria, y se compila con la función de pérdida de entropía cruzada binaria y el optimizador Adam.

D. **Modelo de Regresión Logística TF-IDF:** En este caso, implementaremos la técnica de TF-IDF (Term Frequency-Inverse Document Frequency) en el modelo de regresión logística. TF-IDF asigna pesos a palabras según su frecuencia en un documento y su rareza en el conjunto de documentos, destacando términos importantes en el análisis de texto.

E. **Modelo Basado en BERT** Bidirectional Encoder Representations from Transformers (BERT) es un modelo de lenguaje basado en transformadores que preentrena representaciones de texto bidireccionales para comprender mejor el contexto de las palabras en una oración. Este modelo es especialmente interesante para nosotros, dado el alto nivel técnico de las sentencias de la corte, ya que su tipo de análisis podría generar una representación más fiable del caso de estudio.



Informe

Conjunto al modelo BERT que usamos para clasificación de secuencias en un conjunto de datos judiciales, pudimos hacer el montaje de un modelo de Question & Answering (Q&A) para responder preguntas específicas relacionadas con las opiniones consultivas de la Corte Internacional de Justicia (CIJ). Los resultados fueron satisfactorios para todos nosotros puesto que cumplimos con nuestro objetivo principal

Informe

Pregunta: What are the legal consequences arising from the construction of the wall being built by Israel, the occupying Power, in the Occupied Palestinian Territory, including in and around East Jerusalem, as described in the report of the Secretary-General, considering the rules and principles of international law, including the Fourth Geneva Convention of 1949, and relevant Security Council and General Assembly resolutions?

Respuesta: [CLS] what are the legal consequences arising from the construction of the wall being built by israel , the occupying power , in the occupied palestinian territory , including in and around east jerusalem , as described in the report of the secretary - general , considering the rules and principles of international law , including the fourth geneva convention of 1949 , and relevant security council and general assembly resolutions ? [SEP] have raised the further argument that the court should decline to exercise its jurisdiction because it does not have at its disposal the requisite facts and evidence to enable it to reach its conclusions . in particular , israel has contended , referring to the advisory opinion on the interpretation of peace treaties with bulgaria , hungary and romania , that



Conclusiones

En términos generales, el modelo cumplió con nuestras expectativas para el análisis de los datos judiciales. Sin embargo, reconocemos la necesidad de enriquecerlo con una mayor cantidad de datos para asegurar que pueda generalizar de manera más precisa los casos negativos.



Materiales de interés

Biblioteca Digital de Naciones Unidas <https://digitallibrary.un.org/?ln=en>

Corte Internacional de Justicia <https://www.icj-cij.org/advisory-proceedings>

<https://www.icj-cij.org/case/131/written-proceedings>

<https://www.icj-cij.org/case/131/oral-proceedings>

<https://www.icj-cij.org/case/131/advisory-opinions>



Agradecimientos

Gracias !!



KEEPCODING

Tech School

Madrid | Barcelona | Bogotá