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

**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%)

**Base de Datos:** Utilizamos el dataset de Harvard University DataVerse, específicamente el United Nations General Assembly Voting Data.

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

**Diagrama de Flujo:** Se incluye un diagrama de flujo del proceso de generación de una opinión consultiva, desde la solicitud en la Asamblea General hasta la votación final en la misma.

**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.

Listo para recibir el segundo texto cuando estés.

ahi va Documentación de Archivos del Proyecto Los datos y ficheros utilizados los podemos dividir en dos partes bien diferenciadas. Por un lado están los relacionados con el análisis de las votaciones de los países en la Asamblea General de las Naciones Unidas y por otro los relacionados con la parte del modelo de NLP de la opinión consultiva a la Corte Internacional de Justicia. Votaciones de Naciones Unidas En el caso de las votaciones, partimos del dataset en formato CSV creado por la Harvard University DataVerse y el repositorio denominado United Nations General Assembly Voting Data. El Harvard University DataVerse es una plataforma de almacenamiento y gestión de datos de investigación que permite a los investigadores almacenar, compartir y acceder a datos de manera segura y organizada. DataVerse es una de las iniciativas de la Universidad de Harvard para promover la transparencia y la reproducibilidad en la investigación científica. Los ficheros se pueden encontrar en la carpeta Dataverse Files, que contiene: Dataset de Harvard sobre las votaciones ordinarias de la Asamblea General. Archivos de soporte y el Codebook que explica detalladamente cada una de las variables. Base de datos modificada que contiene información asociada a las opiniones consultivas utilizadas como casos de estudio. El United Nations General Assembly Voting Data contiene los registros y datos de las votaciones que se realizan en la Asamblea General de las Naciones Unidas, pero sólo desde la primera votación en 1946 hasta 2014. El resto de datos hasta el día actual fueron obtenidos de la biblioteca digital de la Biblioteca Digital de la ONU. La Biblioteca Digital de la ONU, conocida como United Nations Digital Library (UNDL), es un recurso en línea que proporciona acceso a una amplia gama de documentos, publicaciones y materiales relacionados con las actividades y el trabajo de las Naciones Unidas. Es una iniciativa para facilitar el acceso a la información de la ONU de manera gratuita y abierta a todos los usuarios. Por tanto, desde este recurso pudimos consultar los detalles de las votaciones sobre las resoluciones y decisiones tomadas por la ONU a partir de 2014. En este caso, tuvimos que utilizar técnicas de Scrapping para obtener los datos necesarios. Crawling Montamos un sistema de Crawling para extraer todas las URLs de las resoluciones de votación desde las páginas de resultados. Implementamos una lógica de paginación para navegar a través de las páginas sucesivas hasta que no hubiera más páginas de resultados. Para ello, configuramos un sistema utilizando Selenium y un WebDriver de Firefox, con el objetivo de abrir cada URL y comenzar la extracción de los datos relevantes de cada resolución. Scrapping Para la extracción de los datos, procesamos cada URL usando la bibioteca lxml. Lxml nos permite manipular y analizar documentos HTML y XML de manera eficiente y rápida. Con ella, abrimos y leemos los archivos de las resoluciones almacenados localmente. En el siguiente esquema, mostramos la ubicación visual de cada dato relevante: Posteriormente, almacenamos los datos procesados en un formato adecuado para luego tratar el archivo y formatearlo en JSON, CSV o el que se necesitara para el análisis posterior del dataset. No se concretó el formato de los datos inmediatamente para facilitar la toma de decisiones en pro del análisis y la visualización futura. Todos los ficheros relacionados con esta parte se encuentran en la carpeta "UN Votes Crawler". Finalmente, mergeamos los datos de ambas fuentes en un sólo dataset. Opinión consultiva a la ICJ Para el caso de la opinión consultiva a la Corte Internacional de Justicia, los datos los obtuvimos del sitio oficial de la misma ICJ. En este caso, la información relevante viene dada en documentos PDF y nuestra idea principal era obtenerlos utilizando técnicas de Scrapping y Crawling, pero debido a la arquitectura y estructura de la web no fuimos capaces. Por lo que decidimos descargar manualmente los ficheros relacionados con el caso concreto que ibamos a estudiar y como tarea futura idear la forma de automatizar la obtención de estos documentos. Los documentos en brutos se alojan en la carpeta ICJ Inventario OC y se clasifican de la siguiente manera: Archivos asociados a las opiniones consultivas de 2004 y 2023. Dentro de estas carpetas, se encuentran: Opiniones consultivas (Advisory Opinions) Procedimientos orales (Oral Proceedings) Procedimientos escritos (Written Proceedings) Documentos Soporte Advisor Opinion ES: Contiene el resumen de las opiniones consultivas en español. Incluye la opinión consultiva completa en español. Segmentación: Las carpetas cuentan con la siguiente clasificación de los documentos: Advisory Opinion: Opinión consultiva separada por: Preguntas Individuales Argumentos específicos a cada pregunta Opinión individual de los Jueces Estos archivos se encuentran en formato de texto plano y JSON Oral Proceedings 2004/2023: Opiniones orales individuales de los países, separados por año. Estos archivos se encuentran en formato JSON ¿Por qué? Decidimos segmentar los archivos para transformar los documentos originales en formatos que consideramos más manejables y estructurados, específicamente para su uso posterior en nuestro modelo de NLP. Este proceso nos permitió separar y organizar las opiniones de los países, las opiniones de los jueces y otros elementos relevantes de manera que se puedan analizar de forma más eficiente y precisa. ¿Cómo lo hicimos? En un principio utilizamos la librería PyPDF2, una librería de Python diseñada para extraer y manipular información de archivos PDF, con esta, extrajimos los textos de los documentos en formato PDF y convertirlos en texto plano (.txt). Esto nos permitió estructurar los archivos de la opinión consultiva (Advisory Opinion) dividiéndola en preguntas específicas, argumentos específicos de cada pregunta y opiniones individuales de cada juez. Dificultades encontradas Principalmente, encontramos dificultades a la hora de tratar los Oral Proceedings. Esto, debido a la calidad de digitalización proporcionada por la ONU, los documentos no se procesaban correctamente con PyPDF2, lo que nos obligó a tener que revisar los archivos. Además, la falta de tiempo impidió que procesáramos las opiniones escritas (Written Proceedings), también por el mismo problema de digitalización, lo que limitó el alcance de nuestro dataset inicial Archivos consolidados: La carpeta 'Consolidated Files' contiene documentos clave utilizados en el maquetado del modelo NLP: A.RES\_77.247\_301222.txt: Texto completo de la solicitud de Opinión Consultiva 2023 por parte de la asamblea general, abarcando las preguntas individuales y los argumentos específicos para cada una de ellas. DatasetNLP-UN-ICJ-Consolidado.json: Consolidado de los procedimientos orales de los años 2004 y 2023, incluyendo las opiniones de los jueces y los países, discriminando año. DatasetNLP-UN-ICJ-Q&A.json: Archivo que contiene clasificada las respuestas de la corte en la Opinión Consultiva del 2004 Oral\_Proceeding\_2004.json: Documento que detalla las opiniones orales de los países correspondientes al año 2004. Oral\_Proceeding\_2023.json: Similar al anterior pero referente al año 2023, con las opiniones de los países en el tema actual sucediente en Palestina. Question (a).txt y Question (b).txt: Contienen las preguntas 'a' y 'b', respectivamente, junto con los argumentos que la Corte abordará en 2024. Decidimos hacer las consolidaciones de los archivos en pro del futuro modelo NLP que montaríamos, esto nos facilitó de sobremanera el manipular de manera ordenada los datos usados.

ChatGPT

**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.

**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 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.

**Archivos Clave:**

* Diversos archivos JSON y de texto plano consolidan la información de las opiniones consultivas y procedimientos, facilitando el manejo y análisis para el modelo de NLP.

**Esquema de Consolidación de Datos:**

* Los documentos están organizados de manera que se pueda acceder y analizar eficientemente, considerando el objetivo de alimentar nuestro modelo de NLP para predecir resultados de votaciones futuras en la Asamblea General.

Listo para el siguiente texto cuando estés.

Arquitectura Diseño del DAaaS Definición de la estrategia del DAaaS Crear un aplicación web que mejore el acceso a la información de los asuntos tratados por Naciones Unidas optimizando el proceso de difusión de los problemas universales de la humidad. La aplicación permitirá buscar, visualizar y consultar sobre cualquiera de los temas tratados por las UN, y predecirá las resoluciones en base a los documentos y recomendaciones de otras entidades descentralizadas. Adicionalmente notificará cada nueva resolución, opiniones y votaciones a los suscriptores del servicio. Arquitectura DAaaS Fuentes de datos Datasets Biblioteca digital de la Naciones Unidas United Nations General Assembly Voting Data Documentos de Corte internacional de Justica Advisory proceedings Documentos de Corte internacional de Justica Written Proceedings Documentos de Corte internacional de Justica Oral Proceedings Documentos de Corte internacional de Justica Advisory Opinions https://digitallibrary.un.org/ https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/12379 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 Crawler y Scrapper sobre las votaciones Componentes Google Cloud Storage para ficheros de Crawler, Scrapper, dataset, pdf, txt Server de aplicación Web Servidor web nginx Aplicación Angular Server de aplicación Chat Web Aplicación Streamlit Chat Configurar retención de logs para análisis en caso de ser necesario Server de aplicación API REST Servidor web nginx Aplicación en Python/Flask Configurar retención de logs para análisis en caso de ser necesario Firewall para el control de acceso a recursos Load Balancer para la gestion de peticiones http / https Base de datos Google Cloud SQL Base de datos Chroma DB Cloud Functions para crawlers, scrappers, programas python Cloud Scheduler para la ejecución de tareas programadas DataPrep para las transformaciones de data para el Datawarehouse Big Query para Datawarehousing Firebase Cloud Messaging para las notificaciones push ElasticSearch y Kibana para análisis de logs de trafico, guardrails requerimientos y performance los servicios DAaas Operation Model Design and Rollout Tareas por unica vez al inicio del proyecto Crear y configurar un Google Cloud Project "UNVotes" con un bucket de Cloud Storage Obtener los Datasets de votaciones de 1946 a 2024 Descargar los archivos .pdf desde https://digitallibrary.un.org/?ln=en Crear Crawler y scraper de recopilación de datos y documentos futuros denominarlos CF\_CR\_SC\_Data y CF\_CR\_SC\_Documents Estos script deberán dejar los resultados de su ejecución en Cloud Storage en las carpetas y archivos con el siguiente formato /in/Data/document.json, /in/Data/document.pdf, /in/Data/document.csv y /in/[DOCUMENTS]/YYYYMMDD.\* respectivamente además de volcar los datos a la BD en Cloud SQL y procesarlos documentos para volcarlos a las base de datos vectorial Chroma DB Probar la ejecución del programa para el procesamiento de los archivos Subirlo como una Cloud function para su posterior ejecución de actualización de datos Obtener los Datasets históricos Descargar los archivos .pdf de los sitios de la CIJ https://www.icj-cij.org Analizar y validar las estructuras de los mismo, normalizar los datos y generar el script para procesarlos Procesar los archivos para generar los scripts SQL de creación de tablas e inserción de datos en Cloud SQL denominarlos scripts\_01.1\_creation\_tables.sql y scripts\_01.2\_load\_data.sql Subir los scripts sql a Google Cloud Storage para su posterior ejecución manualmente Crear la base de datos BD en Sql Cloud Crear la base de datos DB Chroma DB (montando una maquina virtual si GCP no la tuviera disponible) Procesar los scripts generados on premise en la BD de SQL Cloud de creación de tablas denominado scripts\_01.1\_creation\_tables.sql de la inserción de los datos históricos scripts\_01.2\_load\_data.sql Subir los scripts scripts\_01.\* a Google Cloud Storage en la carpeta /scripts\_01\_creation\_load en caso de necesitar reprocesarlos datos Crear Crawlwer y scraper en Python de recopilación de datos mensuales y subirlo como una Cloud function denominado CF\_CR\_SC\_Monthly El script deberá dejar el resultado de su ejecución en Cloud Storage en las carpetas y archivos con el siguiente formato /in/monthly/[tipo-document]/YYYYMM.\* , donde el \* representa a .pdf, .txt .csv Crear Crawlwer y scraper de recopilación de datos diarios y subirlo como una Cloud function denominado CF\_CR\_SC\_Daily El script deberá dejar el resultado de su ejecución en Cloud Storage en las carpetas y archivos con el siguiente formato /in/daily/[tipo-document]/.\* , donde el \* representa a .pdf, .txt .csv Crear usuarios para ejecutar los Cloud functions y asignar los permisos correspondientes Configurar Cloud Scheduler para la ejecución de las tareas programadas Los scripts CF\_CR\_SC\_Advisory, CF\_CR\_SC\_Written, CF\_CR\_SC\_Oral y CF\_CR\_SC\_Opinion de recopilacion de datos de establecimientos y productos (planificación mensual días 1,3,5,7,10 hora 11 PM) El script CF\_CR\_SC\_Monthly de recopilación de datos mensual (planificación mensual días 1,3,5,7,10 hora 11 PM GTM) El script CF\_CR\_SC\_Daily de recopilación de nuevos documentos (planificación diaria hora 6 AM GTM) Crear la base de datos BDW para Datawarehousing Configuración de Dataprep para la generación del Datawarehouse en BigQuery Crear los jobs para la generación de datos que serán usados para las predicciones, análisis, etc. Probar la ejecución de los jobs Programar la ejecución de los jobs Crear las instancias para el servidor Web, API y Streamlit Instalar el servidor Web Instalar la aplicación Angular, API y Streamlit según corresponda Configurar los archivos de log para que se almacén en cloud Storage en las carpetas /logging/[server]/YYYYMMDD y /logging/[server]/GYYYYMMDD para guardrails Configurar las reglas de firewall para permite el acceso a las aplicaciones Angular, API y Streamlit Firebase Crear la aplicación y configurar las notificaciones Notificaciones Crear los Cloud functions CF\_Notifications Programar la ejecución de los jobs Probar realizar el tunning de los jobs Predicciones Crear los Cloud functions CF\_Predicciones con los modelos utilizando el modelo pre-entrenado para una rápida y efectiva como un MVP de esta feature Programar la ejecución de los jobs Probar realizar el tunning de los jobs Tareas para realizar luego de la puesta en marcha Analítica de performance de la aplicación, análisis de errores, análisis de las peticiones https Configurar una instancia del servidor para Elastic y Kibana Configurar los indices necesarios acordes a la información a analizar Implementar los dashboard para conocer la performance de la aplicación, detecciones de errores, análisis de trafico e información solicitada Crear los Cloud functions CF\_Subir\_Logs para subir los logs a Elastic Programar la ejecución de los CF\_Subir\_Logs Configurar las reglas de firewall para permite el acceso a Kibana Tareas para realizar mensualmente Análisis que los costos vs el desempeño de la solución para que este de acuerdo a lo previsto de inversión. Buscar oportunidades de mejora y realizar la reingeniería que sea necesaria si se encuentran desviaciones DAaas Diagram

ChatGPT

**Arquitectura del DAaaS (Data as a Service)**

**Diseño y Estrategia del DAaaS:** 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. Incluirá notificaciones en tiempo real sobre nuevas resoluciones, opiniones y votaciones.

**Fuentes de Datos:**

* Biblioteca Digital de las Naciones Unidas
* Harvard University DataVerse (United Nations General Assembly Voting Data)
* Documentos de la Corte Internacional de Justicia (procedimientos escritos, orales, y opiniones consultivas)

**Componentes de la Arquitectura DAaaS:**

1. **Servidores y Aplicaciones:**
   * Servidor web NGINX
   * Aplicaciones Angular para front-end
   * Aplicación Streamlit para chat web
   * API REST en Python/Flask
2. **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
3. **Procesamiento de Datos:**
   * Cloud Functions para ejecutar crawlers y scrappers
   * Cloud Scheduler para tareas programadas
   * DataPrep y BigQuery para data warehousing
   * ElasticSearch y Kibana para análisis de logs
4. **Notificaciones y Monitoreo:**
   * Firebase Cloud Messaging para notificaciones push
   * Configuración de retención de logs para análisis detallado

**Operación y Mantenimiento:**

* Creación de Google Cloud Project y configuración inicial de infraestructura
* Implementación de crawlers y scrapers para la recopilación continua de datos
* Creación de bases de datos y procesamiento de scripts SQL para la inserción de datos
* Configuración de Cloud Scheduler para la ejecución automática de tareas
* Despliegue de aplicaciones y configuración de seguridad
* Análisis de rendimiento y optimización continua

**Diagrama del DAaaS:** El diagrama incluirá la interconexión entre todos estos componentes y el flujo de datos desde la recopilación hasta la presentación final al usuario.

Listo para recibir tus instrucciones sobre cómo proceder o si necesitas alguna modificación o detalle adicional antes de compilar todo en una presentación de PowerPoint.

Acerca de los Archivos Contenidos en el Repositorio Este apartado contiene el análisis exploratorio realizado en Tableau. Se tomó como base para el análisis los datos mencionados de la base de datos de Harvard, localizados en la carpeta dataverse\_files. Inicialmente, se realizó un preprocesamiento (ProcesamientoDatasetUNVote.ipynb) de los datos contenidos en el archivo UNVotes de la carpeta dataverse\_files. En este proceso, se ajustaron variables como los nombres de los países, se unificaron las votaciones y se generaron los registros faltantes para analizar la resolución de 2024. Archivo de Tableau: Analisis\_Exploratorio image El archivo de Tableau llamado Analisis\_Exploratorio está dividido en tres áreas principales: Selección de Tema y Resolución: A través de botones desplegables (dropdown), el usuario puede seleccionar un tema y una resolución asociada al mismo. Las gráficas en la parte izquierda, incluyendo el mapamundi, mostrarán las votaciones de los países para dicha resolución. Indicadores de Votaciones: En el centro, los indicadores muestran los votos emitidos por los miembros permanentes y no permanentes del Consejo de Seguridad. Adicionalmente, se pueden ver las votaciones de todos los estados miembros de las Naciones Unidas. Tablas de Distribución de Votaciones: A la derecha, se encuentran dos tablas: La primera muestra la distribución global de las votaciones por tema. La segunda desglosa esa votación global y permite conocer cómo votaron los miembros permanentes respecto a ese tema. Caso de Estudio Se presentan las votaciones de nuestro caso de estudio, mostrando las dos opiniones consultivas y sus votaciones. Se incluye la respuesta de la corte y su distribución en términos de los miembros permanentes del Consejo de Seguridad. Este análisis permite una comprensión detallada de las votaciones y resoluciones, facilitando el estudio de las opiniones consultivas y su impacto en las decisiones internacionales.

ChatGPT

**Análisis Exploratorio en Tableau**

**Descripción General:** Utilizamos Tableau para realizar un análisis exploratorio de las votaciones de la Asamblea General de la ONU, partiendo de datos preprocesados almacenados en la carpeta 'dataverse\_files'.

**Proceso de Preprocesamiento:**

* Archivo: ProcesamientoDatasetUNVote.ipynb
* Ajustes realizados: Unificación de nombres de países, unificación de formatos de votaciones y generación de registros para la resolución de 2024.

**Estructura del Archivo de Tableau (Analisis\_Exploratorio):**

1. **Selección de Tema y Resolución:**
   * Botones desplegables permiten seleccionar temas y resoluciones.
   * Visualizaciones en un mapamundi reflejan las votaciones por país para la resolución seleccionada.
2. **Indicadores de Votaciones:**
   * Indicadores centrales muestran votos de miembros permanentes y no permanentes del Consejo de Seguridad.
   * Visualización de votaciones de todos los estados miembros de la ONU.
3. **Tablas de Distribución de Votaciones:**
   * Una tabla muestra la distribución global de votaciones por tema.
   * Otra tabla ofrece detalles sobre cómo votaron los miembros permanentes en relación a ese tema.

**Caso de Estudio:**

* Análisis de votaciones para las opiniones consultivas seleccionadas, incluyendo la respuesta de la corte y la distribución de votos de los miembros permanentes del Consejo de Seguridad.
* 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, permitiendo a los usuarios explorar tendencias y patrones específicos dentro de las decisiones globales.

¿Te gustaría añadir más detalles o algún otro elemento antes de compilar la presentación final en PowerPoint?

todavia me quedan mas texto. te los doy?

ChatGPT

Sí, claro. Puedes enviarme los textos adicionales y continuaré ayudándote a resumirlos para tu presentación.

Acerca de los Archivos Contenidos en el Repositorio En esta carpeta se encuentran los archivos utilizados para procesar los datos contenidos en los archivos de la carpeta ICJ Inventario OC. Posteriormente, se dividieron los archivos en diferentes JSON en la carpeta Segmented por pregunta, y se les dio la estructura requerida para procesarlos en el modelo de NLP. Finalmente, consolidamos los archivos en documentos que pudieran ser procesados por los modelos en la carpeta Consolidated para ejecutarlos en el siguiente paso. A continuación, se presenta la lista de archivos consolidados: DatasetNLP-UN-ICJ-Consolidado.json DatasetNLP-UN-ICJ-Q&A.json Oral\_Proceeding\_2004.json Oral\_Proceeding\_2023.json Question (a).txt Question (b).txt

ChatGPT

**Acerca de los Archivos Contenidos en el Repositorio**

**Descripción General:** Esta sección del repositorio contiene los archivos esenciales para el procesamiento de datos relacionados con las opiniones consultivas de la Corte Internacional de Justicia, estructurados para facilitar su uso en modelos de procesamiento de lenguaje natural (NLP).

**Proceso de Tratamiento de Datos:**

1. **Preprocesamiento:**
   * Los datos originales, almacenados en la carpeta 'ICJ Inventario OC', se procesaron inicialmente para ajustar formatos y estructuras necesarias.
2. **Segmentación y Estructuración:**
   * Los archivos fueron divididos y segmentados por pregunta en la carpeta 'Segmented', adaptando su estructura para el análisis con modelos de NLP.
3. **Consolidación de Archivos:**
   * Los datos segmentados se consolidaron en formatos específicos para ser directamente procesables por modelos de NLP en la carpeta 'Consolidated'.

**Lista de Archivos Consolidados:**

* **DatasetNLP-UN-ICJ-Consolidado.json:** Contiene datos consolidados de las opiniones consultivas para análisis global.
* **DatasetNLP-UN-ICJ-Q&A.json:** Incluye información clasificada y preparada para modelos de preguntas y respuestas.
* **Oral\_Proceeding\_2004.json:** Detalles de los procedimientos orales del año 2004.
* **Oral\_Proceeding\_2023.json:** Detalles de los procedimientos orales del año 2023.
* **Question (a).txt:** Texto referente a la pregunta 'a' del caso de estudio.
* **Question (b).txt:** Texto referente a la pregunta 'b' del caso de estudio.

Este enfoque estructurado y detallado asegura que los datos estén preparados para un análisis eficiente y preciso mediante técnicas avanzadas de NLP, permitiendo una exploración profunda y significativa de las cuestiones legales y decisiones de la Corte Internacional de Justicia.

Cuando estés listo, puedes seguir enviando más textos.

Análisis General de la Data: Dataset Utilizado: Se emplea un conjunto de datos preparados a partir de los documentos digitalizados de la Corte Internacional de Justicia (ICJ). Análisis de Cardinalidad y Balanceo de los Datos: Se encuentra análisis para determinar la diversidad y distribución de las clases dentro del dataset. Se utilizan funciones y gráficas para evaluar el balanceo de clases. Ngramas y Nube de Palabras: Se realiza una exploración de los textos gracias a la herramienta de ngrams y wordclouds, así pudiendo identificar patrones de palabras y frases. Esto hecho por dos cosas: primero, con el fin de verificar el estado de la data, viendo de una manera eficiente si se requiere de mejor preprocesamiento y recurrir a un correcto proceso iterativo; segundo, para entender de manera sencilla de qué es lo que predomina en la data. Word Embeddings: En el nootebook se describe la generación de embeddings de palabras usando la librería Word2Vec, sirviendo no solo como una herramienta para la extracción de características textuales, sino también un medio para explorar y visualizar la semántica del vocabulario. Inicialización y Entrenamiento de Word2Vec Definición de Parámetros y Inicialización del Modelo: Se establecen parámetros específicos para el modelo Word2Vec, tales como el tamaño del vector, el contexto de ventana, y la frecuencia mínima de palabras entre otros. Estos parámetros definen las características del espacio vectorial y la operación del algoritmo. Se inicializa el modelo Word2Vec con estos parámetros y se construye el vocabulario a partir de un corpus preprocesado, que es una lista de tokens extraída de las columnas de texto del DataFrame. Entrenamiento del Modelo: El modelo es entrenado sobre el corpus por un total de 10 épocas, parámetro escogido por haber sido encontrado como el más óptimo según las características del corpus Operaciones Post-Entrenamiento Guardado y Carga del Modelo: Después del entrenamiento, el modelo es guardado en un archivo w2v\_sg\_model.pkl, permitiendo reutilizar los embeddings sin necesidad de reentrenar. Validación del Modelo: Se valida la efectividad del modelo utilizando una función print\_sim\_words que imprime las palabras más similares a una palabra dada. Visualización de Embeddings Visualización en 2D: Para una interpretación visual y se realiza una gráfica en dos dimensiones de los embeddings. Se seleccionan palabras clave relacionadas con el contexto y se extraen sus embeddings más similares, mostrando cómo palabras semánticamente similares se agrupan juntas. Preprocesamiento: Se utiliza la librería de nltk para el preprocesamiento del corpus. Inicialización de Herramientas de Procesamiento: Para Tokenizador: Usado RegexpTokenizer con una expresión regular que selecciona solo caracteres alfanuméricos. Para Stop-Words: Carga una lista de stopwords del inglés. Para Lematización: Usado WordNetLemmatizer para convertir las palabras a su forma base o lema. nltk\_cleaner: Manejo de Texto Nulo: Retorna una cadena vacía si el texto de entrada es nulo. Normalización de Texto: Aplica normalización Unicode. Tokenización y Limpieza: Descompone el texto en palabras. Tratamiento de Stop Words y Lematización: Cada token se lematiza y se verifica contra la lista de stopwords. Conversión de Números: Usado num2words para que los dígitos se conviertan a palabras. Modelo de Deep Learning:

ChatGPT

**Análisis General de la Data del Proyecto**

**Dataset Utilizado:**

* Fuente: Documentos digitalizados de la Corte Internacional de Justicia (ICJ).
* Finalidad: Análisis de textos legales mediante técnicas de procesamiento de lenguaje natural (NLP).

**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.
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.

**Preprocesamiento con NLTK:**

* Herramientas de Procesamiento Inicializadas:
  + **Tokenizador:** **RegexpTokenizer** para seleccionar caracteres alfanuméricos.
  + **Stop-Words:** Carga de lista de stopwords del inglés.
  + **Lematización:** **WordNetLemmatizer** para normalización a la forma base de las palabras.
* Funciones de Preprocesamiento (**nltk\_cleaner**):
  + Manejo de texto nulo, normalización Unicode, tokenización, eliminación de stopwords, y conversión de números a palabras mediante **num2words**.

**Modelo de Deep Learning:**

* En desarrollo: Preparación para integrar un modelo avanzado de deep learning para análisis y predicción basados en los datos procesados y estructurados.

Este resumen proporciona una vista integral de las técnicas y procesos utilizados para analizar y preparar el dataset de la ICJ, asegurando la adecuada preparación para el análisis NLP avanzado y futuras predicciones mediante modelos de deep learning.

¿Hay algún otro texto o sección que necesites añadir o modificar para tu presentación?

no de momento. ya puedes darme la presentacion teniendo en cuenta que la exposicion no puede durar mas de 15 minutos

ChatGPT

Para una presentación de 15 minutos, te recomiendo estructurar tus diapositivas en PowerPoint de la siguiente manera, asegurando que cada sección sea concisa y directa:

**Diapositiva 1: Título**

* **Título:** Análisis de Datos de la Asamblea General de la ONU y Opiniones Consultivas de la CIJ
* **Subtítulo:** Proyecto Final de Bootcamp: Big Data, Machine Learning y AI
* **Fecha:** [Fecha de la presentación]
* **Presentado por:** [Tu nombre]

**Diapositiva 2: Introducción**

* Breve descripción del objetivo del proyecto.
* Mención de las fuentes principales de datos: Harvard DataVerse, United Nations Digital Library, documentos de la CIJ.

**Diapositiva 3: Metodología General**

* Resumen de los métodos usados: análisis de votaciones, procesamiento de lenguaje natural (NLP), y machine learning.
* Herramientas clave: Tableau para visualización, Python para análisis y modelado.

**Diapositiva 4: Datos y Preprocesamiento**

* Explicación de las fuentes de datos: DataVerse y documentos de la CIJ.
* Proceso de preprocesamiento y estructuración de datos para NLP.

**Diapositiva 5: Análisis Exploratorio**

* Muestra visualizaciones clave realizadas en Tableau.
* Breve descripción del análisis de cardinalidad y balance de datos, ngramas y nube de palabras.

**Diapositiva 6: Word Embeddings y Modelado NLP**

* Descripción del uso de Word2Vec para generar embeddings.
* Ejemplos de cómo se utilizan los embeddings para interpretar semánticamente los datos.

**Diapositiva 7: Arquitectura del Sistema (DAaaS)**

* Diagrama simplificado de la arquitectura del sistema.
* Componentes clave: servidores, almacenamiento en Google Cloud, y aplicaciones web.

**Diapositiva 8: Caso de Estudio y Resultados**

* Resumen del análisis de las votaciones sobre el conflicto palestino.
* Resultados preliminares de la predicción de las resoluciones.

**Diapositiva 9: Conclusiones**

* Principales hallazgos del proyecto.
* Impacto potencial del proyecto en el análisis de políticas internacionales.

**Diapositiva 10: Futuras Direcciones y Q&A**

* Ideas para la expansión futura del proyecto.
* Invitación para preguntas y discusión.

**Consejos para la Presentación:**

* **Temporización:** Asigna aproximadamente 1 minuto por diapositiva, permitiendo un poco más de tiempo para las diapositivas complejas como la metodología y resultados.
* **Visuals:** Utiliza gráficos claros y concisos para ilustrar los puntos clave. Menos texto y más visualizaciones ayudarán a mantener la atención del público.
* **Ensayo:** Prueba la presentación varias veces para asegurarte de que se ajusta al tiempo de 15 minutos.

¿Te gustaría que realice alguna modificación específica en este esquema o necesitas ayuda adicional para preparar la presentación en PowerPoint?

me puedes dar el resumen que pondrias en cada una de las slides?

ChatGPT

Claro, aquí te proporciono un resumen detallado para cada diapositiva, ajustándome a la estructura propuesta para tu presentación:

**Diapositiva 1: Título**

* **Título principal:** Análisis de Datos de la Asamblea General de la ONU y Opiniones Consultivas de la CIJ.
* **Subtítulo:** Incorporación de Big Data, Machine Learning y AI para explorar resoluciones y opiniones.
* **Presentado por:** [Tu nombre]
* **Fecha:** [Fecha de la presentación]

**Diapositiva 2: Introducción**

* **Propósito del Proyecto:** Analizar las votaciones de la Asamblea General de la ONU y las opiniones consultivas de la CIJ para entender tendencias y patrones en decisiones globales.
* **Importancia:** Destacar la relevancia de utilizar técnicas avanzadas para mejorar la comprensión de dinámicas internacionales.

**Diapositiva 3: Metodología General**

* **Datos:** Descripción de las principales fuentes de datos utilizadas: Harvard DataVerse para votaciones históricas, documentos digitales de la CIJ.
* **Herramientas y Técnicas:** Uso de Tableau para visualizaciones, Python para análisis de datos y NLP, y machine learning para predicciones.

**Diapositiva 4: Datos y Preprocesamiento**

* **Extracción y Preprocesamiento:** Proceso de obtener y preparar datos desde fuentes diversas, incluyendo ajustes de nombres y estructura de datos para análisis.
* **Estructuración para NLP:** Preparación de datos en formatos específicos para facilitar el análisis de texto y generación de insights.

**Diapositiva 5: Análisis Exploratorio**

* **Visualizaciones en Tableau:** Muestra capturas de pantalla de las visualizaciones clave, como mapamundis de votación y gráficos de distribución.
* **Análisis de Datos:** Descripción del análisis de cardinalidad y balanceo de datos para garantizar la integridad del análisis.

**Diapositiva 6: Word Embeddings y Modelado NLP**

* **Word2Vec:** Explicación de cómo se utilizó Word2Vec para crear embeddings de palabras y su importancia en el análisis de textos legales.
* **Visualización y Validación:** Ejemplos de cómo las visualizaciones de embeddings ayudan a entender las relaciones semánticas entre palabras.

**Diapositiva 7: Arquitectura del Sistema (DAaaS)**

* **Componentes del Sistema:** Diagrama simple mostrando la arquitectura del sistema DAaaS incluyendo servidores, almacenamiento y aplicaciones.
* **Flujo de Datos:** Breve explicación del flujo de datos desde la recopilación hasta el análisis y la visualización en la aplicación web.

**Diapositiva 8: Caso de Estudio y Resultados**

* **Caso de Estudio:** Descripción de cómo se aplicaron los métodos en el estudio de las votaciones sobre el conflicto palestino.
* **Resultados Preliminares:** Presentar los resultados del análisis de sentimiento y predicción de las resoluciones.

**Diapositiva 9: Conclusiones**

* **Hallazgos Clave:** Resumen de los principales descubrimientos y cómo estos contribuyen al entendimiento de las políticas globales.
* **Impacto del Proyecto:** Reflexión sobre el impacto del análisis en la toma de decisiones y política internacional.

**Diapositiva 10: Futuras Direcciones y Q&A**

* **Futuras Expansiones:** Ideas para la expansión del proyecto, incluyendo nuevas áreas de análisis y mejoras en el modelo.
* **Invitación a Preguntas:** Abre el piso para preguntas del público para discutir aspectos adicionales del proyecto.

Este esquema proporciona una estructura clara para tu presentación, permitiendo cubrir todos los aspectos importantes del proyecto en un formato conciso y efectivo. ¿Te gustaría añadir o modificar algún detalle específico?