Universidad de los Andes
Ingeniería de Sistemas y Computación

Proyecto Etapa 2 - Automatización y uso de modelos de analítica de textos

Caso Fondo de Poblaciones de las Naciones Unidas (UNFPA)

Integrantes

- Matero Calderon
- Juan Ramirez
- Daniela Castrillón

Sección 1

Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API.

Definiciones de re-entrenamiento tenidas en cuenta

Para el desarrollo del segundo endpoint se tuvieron en cuenta distintos tipos de re-entrenamiento, a continuación, se van a mencionar los que se tuvieron en cuenta y se va a explicar cuál se implementó y por qué.

1. Re-entrenamiento con datos nuevos e históricos: El primer escenario que se tuvo en cuenta es usar los nuevos datos que el usuario provee en la aplicación y juntarlos con los datos históricos que anteriormente habían sido usados para el entrenamiento del modelo, pero en el caso de que los datos usados anteriormente no se quieran seguir usando por alguna razón, como que ya son demasiado antiguos y no representan el interés de la población sobre los pueden generar problemas, ya que es posible que con el paso de los años los conceptos usados para representar estos objetivos varíen y no solo ello, sino que usar datos históricos puede no ser la mejor opción para la organización en caso de que por alguna razón se hayan usado datos incorrectos o que no eran apropiados (como que tenían diversos errores) y que no quieran volver a ser usados para el re-entrenamiento.
2. Re-entrenamiento solo con datos nuevos: este escenario solo usa los datos nuevos que el usuario provee al momento de re-entrenar el modelo, es decir, que a diferencia del caso anterior no se tienen en cuenta los datos históricos para hacer el nuevo entrenamiento. La ventaja que identificamos es que de esta manera la organización puede controlar en mayor medida los datos que desea que sean usados para entrenar el modelo, manteniendo los datos que se ajustan a las nuevas tendencias e intereses de los ciudadanos en los comentarios sobre los ODS.
3. Entrenamiento Online: en este escenario el re-entrenamiento del modelo se hace actualizando en lotes el modelo, haciendo que este no se reentrene con todos los datos nuevos ni con todos los datos antiguos, de esta forma haciendo que el re-entrenamiento no se haga con todos los datos. Este escenario puede parecer a priori

muy bueno, pero consideramos que cuando al seleccionar lotes de información, puede llevar a que alguna información que resulta muy importante para la organización no sea tenida en cuenta en el re-entrenamiento del modelo.

Conclusión: se decidió usar el segundo escenario de re-entrenamiento (usar solo datos nuevos), ya que como el modelo se basa en comentarios que las personas dan sobre cualquiera de estos ODS, hace que las opiniones cambien con el paso del tiempo, haciendo que las preocupaciones que una persona pueda tener sobre un tema se desaparezcan con el paso del tiempo, o por el contrario estas se agudicen por alguna razón. Por ello consideramos que este escenario le permite a la organización poder usar información actualizada y que esta lo más cercana posible a la opinión de la población sobre los diversos temas, haciendo que el modelo de esta forma haga las predicciones lo más cercanas a la realidad.

Sección 2

Descripción del usuario/rol de la organización que va a utilizar la aplicación:

La aplicación está dirigida a analistas de datos, responsables de políticas públicas, y consultores dentro de organizaciones como el Fondo de Poblaciones de las Naciones Unidas (UNFPA) o instituciones gubernamentales y ONGs asociadas a los Objetivos de Desarrollo Sostenible (ODS). Su rol es recolectar, procesar y analizar datos cualitativos proporcionados por la ciudadanía para evaluar el impacto de iniciativas sociales y proponer nuevas estrategias.

La aplicación se integra en el proceso de recopilación de comentarios ciudadanos, permitiendo a la organización automatizar la clasificación de estos comentarios en función de los ODS a los que hacen referencia, como la salud, la educación o la igualdad de género. Esto optimiza el análisis de grandes volúmenes de información no estructurada, facilitando la identificación de áreas prioritarias para las políticas públicas.

Para los usuarios, la aplicación es crucial para obtener insights en tiempo real sobre las preocupaciones y sugerencias de los ciudadanos, lo que permite ajustar las estrategias organizativas en función de los ODS más relevantes. Esta herramienta reduce significativamente el tiempo y los recursos necesarios para categorizar los datos, ayudando a la toma de decisiones informadas y alineadas con los objetivos globales de desarrollo sostenible.

Sección 3

El video de los resultados del proyecto se encuentra en el Padlet correspondiente.

Sección 4

Trabajo en equipo

- **Líder de proyecto:** Este rol lo desempeño Daniela Castrillón, en donde definió las fechas de reuniones, pre-entregables del grupo, verificó que las asignaciones de tareas en cuanto a la carga fueran equitativas y fue la encargada de subir la entrega del grupo.

  Las reuniones realizadas fueron:
  - Reunión de lanzamiento y planeación: En esta reunión se definieron los roles y la forma de trabajo que el grupo iba a llevar a cabo; adicionalmente, se discutieron posibles soluciones para resolver el proyecto. Esta reunión tomó 30 minutos.
  - Reunión de seguimiento: Se realizaron dos reuniones de seguimiento de 30 minutos durante la semana 9 para llevar una bitácora de avances y posibles problemas durante el desarrollo que debían ser resultas en grupo.
  - Reunión de finalización: Finalmente, se consolidó el trabajo final, con el fin de validar los entregables con todo el grupo y verificar los puntos requeridos para obtener los mejores resultados en el proyecto. Esta reunión tomó 40 minutos.

- **Ingeniero de datos:** Este rol lo desempeño Mateo Calderon, el cual se encargó de velar por la calidad del proceso de automatización relacionado con la construcción del modelo analítico.

  Tiempo dedicado: 4.5 horas

- **Ingeniero de software responsable de los resultados:** Este rol lo desempeño Daniela Castrillón, quien se encargó de generar el video con los resultados obtenidos.

  Tiempo dedicado: 2 horas

- **Ingeniero de software responsable del diseño y desarrollo la aplicación final:** Este rol lo desempeño Juan Ramirez, quien estuvo encargado del diseño y gestión del proceso de construcción de la aplicación.

  Tiempo dedicado: 4.5 horas

Retos enfrentados en el proyecto

- Conocer cómo transformar el código del notebook de la entrega anterior en un pipeline consolidado.
- Encontrar documentación para construir el pipeline con Scikit Learn.

- Con el código del API hecho, existían problemas al ingresar el archivo que contenía el modelo, ya que la clase que contenía el código del pipeline generó muchos problemas para ejecutar con el código del API, lo que llevó a hacer varios cambios en el API.
- Manejo de los diferentes formatos de respuesta: El retorno de las predicciones y probabilidades y métricas requirió ajustes en cómo se presentaba la información en la interfaz.
- Validación y pruebas del flujo de trabajo: Comprobar que el modelo, la API y la interfaz funcionaran de manera fluida y sin errores, lo que requirió varias iteraciones y ajustes.

Formas planteadas para resolver el proyecto antes del desarrollo

- Especialmente a la hora de construir el 2do endpoint, ya que era necesario plantear tres definiciones de re-entrenamiento, dichas definiciones estarán en el documento explicadas.
- Usar librerías o frameworks adicionales para gestionar mejor las respuestas de la API, como axios o fetch con más control sobre los errores.
- Se considero desarrollar diferentes versiones del frontend, como React, para manejar mejor la interacción dinámica con el backend, pero se optó por una solución más ligera con JavaScript y Bootstrap para agilizar el desarrollo.
- Uso de bases de datos para almacenar comentarios históricos y las predicciones para análisis futuros, pero para esta versión decidimos centrarnos en el flujo directo de predicción.

Repartición de puntos entre integrantes

Los 100 puntos disponibles se asignarán de forma equitativa entre los tres integrantes: Mateo, Daniela y Juan. Cada uno recibirá 33.3 puntos como reconocimiento a su destacado desempeño, compromiso y la alta calidad de los entregables presentados.

Esta decisión refleja la colaboración y el esfuerzo conjunto que todos han demostrado a lo largo del proyecto, lo que ha llevado al éxito del mismo. Es importante resaltar que, al repartir los puntos de esta manera, se fomentan un ambiente de trabajo en equipo, donde cada contribución individual se valora y se considera esencial para el logro de los objetivos comunes.

Puntos de mejora para el siguiente proyecto

- Realizar el proyecto con mayor antelación permitirá revisar cada fase con más calma, resolver dudas y abordar problemas de manera efectiva. Esto también facilitará la implementación de ajustes necesarios sin presión del tiempo.
- Es importante recolectar retroalimentación tanto interna como externa durante todo el proceso. Esto ayudará a ajustar y mejorar el proyecto en tiempo real, asegurando que se satisfagan las necesidades de todos los involucrados.

- Fomentar una comunicación más efectiva entre los integrantes del equipo es esencial para asegurar que todos estén alineados y comprometidos con los objetivos del proyecto. Reuniones regulares y el uso de herramientas de colaboración son útiles en este sentido.
- Establecer indicadores calve de rendimiento para evaluar el impacto y el éxito del proyecto de manera objetiva. Aprender de estos resultados será fundamental para informar y mejorar en futuras iteraciones.
- Mantener una documentación detallada de cada etapa del proyecto facilitará la identificación de lecciones aprendidas y la replicación de buenas prácticas en futuros trabajos.

Link del repositorio

https://github.com/MateoCr816/Proy1BI