Proyecto 1 Etapa 2 Despliegue de Modelos de Lenguaje Automático

Miembros

Michael Dylan Blanquicett Carvajal

Juan Pablo Hernández

Juan Diego Ortega Medina

Universidad de los Andes

ISIS 3301 - Inteligencia de Negocios

Bogotá, Bogotá D.C., Colombia

20/04/2024

Objetivos

- Automatizar un proceso replicable para aplicar la metodología de analítica de textos en la construcción de modelos analíticos.
- Desarrollar una aplicación que utilice un modelo analítico basado en aprendizaje automático y sea de interés para una organización, empresa o institución y en particular para un rol existente en alguna de ellas.
- Interactuar con un grupo interdisciplinario para validar y mejorar la calidad de la solución analítica planteada y del producto de software construido.

Entendimiento del problema

El Ministerio de Comercio, Industria y Turismo de Colombia, la Asociación Hotelera y Turística de Colombia – COTELCO, cadenas hoteleras de la talla de Hilton, Hoteles Estelar, Holiday Inn y hoteles pequeños ubicados en diferentes municipios de Colombia están interesados en analizar las características de sitios turísticos quelos hacen atractivos para turistas locales o de otros países, ya sea para ir a conocerlos o recomendarlos. De igual manera, quieren comparar las características de dichos sitios, con aquellos que han obtenido bajas recomendaciones y que estánafectando el número de turistas que llegan a ellos. Además, quieren tener un mecanismo para determinar la calificación de un sitio por parte de los turistas y aplicar estrategias para identificar oportunidades de mejora que permitan aumentar la popularidad de los sitios y fomentar el turismo. Esos actores de turismo prepararon dos conjuntos de datos con reseñas de sitios turísticos. Cada reseña tiene una calificación según el sentimiento que tuvo el turistaal visitarlo. Estos actores quieren lograr un análisis independiente de los conjuntos de datos y al final del proyecto discutir sobre los grupos de científicos de datos e ingenieros de datos que acompañarán el desarrollo real de este proyecto.

Proceso de automatización

Para el desarrollo de esta etapa, inicialmente se realizaron descargas y importaciones de librerías externas para la limpieza inicial de los datos y la construcción del modelo. Debemos entender que los datos son reseñas, escritas en español, así que, para la preparación de estos datos, se pasó por un proceso de limpieza riguroso y extenso. Se eliminaron las reseñas duplicadas, se convirtieron todos los caracteres a minúsculas, se convirtieron todos los números escritos en su forma textual (3 ahora será “tres”), remover los caracteres que no sean del alfabeto, devolver las palabras a su forma base, es decir, lematizar, quitar signos de puntuación y acentuación y quitar palabras vacía. Una vez hecha esta limpieza, se procede a extraer una muestra de esos datos (exactamente el 80%) para poder entrenar el modelo en cuestión. Después de la primera etapa de este proyecto, pudimos como grupo concluir que el mejor modelo que se pudo implementar fue el algoritmo de regresión logística. Así que, para aplicar este modelo, se realizó la respectiva vectorización de las reseñas y se aplico el propio

algoritmo. Una vez hecho esto, se ejecutó este modelo entrenado con datos de prueba para poder obtener métricas que nos indicaban que tan bien se aproximaba el algoritmo a las calificaciones esperadas. Este algoritmo tuvo una precisión del 50,54% en su mejor iteración, por lo que podemos concluir que la mitad de las predicciones que hace son correctas. Una vez ya entrenado y evaluado este modelo, finalmente se persistió usando joblib. Este pipeline queda guardado para luego ser usado en un API. El pipeline guardado ya queda con toda la información necesaria, es decir, ya queda entrenado, para que cualquier persona que quiera, mediando una aplicación web, pueda insertar su archivo con reseñas y el modelo, que ya esta persistente, recibiendo el archivo por medio de peticiones HTTP, pueda evaluarlo y retornar las métricas con los resultados de la evaluación

Desarrollo de la aplicación

Para el desarrollo de la aplicación se usaron 3 tecnologías fundamentales: Jupyter notebook, para la creación del pipeline, Fast API, para la creación del API que conectará el back con el front, y React, para el desarrollo de la aplicación web con la que interactuara el usuario final. Esta aplicación esta pensada para que cualquier persona pueda entenderla. Claro que será una mejor experiencia si el usuario tiene entendimiento previo de las métricas que se muetran como conclusión del entrenamiento y evaluación del modelo.

Mejorar para la Siguiente Entrega

Manejo del Tiempo: Como grupo, consideramos que una buena mejora para la siguiente entrega es el manejo del tiempo. Esto es de suma importancia ya que eso define nuestra eficiencia en el desarrollo del proyecto

Planificación: Como grupo, consideramos que realizar una planificación exhaustiva del proyecto, estableciendo metas, plazos y recursos necesarios de manera anticipada puede contribuir a evitar retrasos y problemas inesperados.

Flexibilidad: Estar abiertos a adaptarse a cambios en el proyecto y ser capaces de ajustar la estrategia según sea necesario para enfrentar desafíos o aprovechar oportunidades.

Feedback constructivo: Fomentar una cultura de retroalimentación constructiva, donde se brinde y se reciba feedback de manera regular y respetuosa para mejorar el desempeño individual y del equipo.