Proyecto interdisciplinario analítica de textos
Parte 2

Inteligencia de negocios
Estadística

Turismo de los Alpes

Felipe Rueda
Santiago Pardo
Luis Felipe Plazas
Isabella Nova
Ana Sánchez

07 de abril de 2024
Bogotá D.C

Contenido

Contexto:

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

El proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso a través de una API se realizó siguiendo un enfoque basado en pipelines. Aquí se describe cada paso del proceso junto con el código correspondiente:

Para la Preparación de Datos se recopilaron las reseñas de sitios turísticos junto con sus calificaciones correspondientes, después se realizó la limpieza de datos para eliminar caracteres especiales, convertir texto a minúsculas, eliminar palabras vacías y tokenizar el texto. Para finalizar, se utilizó la librería NLTK para realizar el procesamiento de lenguaje natural.

Para automatizar el proceso de la preparación de los datos utilizamos el trabajo hecho anteriormente en la fase 1 de este mismo proyecto. Para cada reseña, se aplicó cierto proceso de limpieza en donde implementamos las siguientes estrategias:

Se aseguró de que todas las opiniones estuvieran escritas de la misma manera. Se eliminaron los caracteres que no pertenecían al código ASCII (como varios puntos o símbolos), se convirtieron todas las palabras a minúsculas, se quitaron los signos de puntuación (incluyendo tildes) y las palabras que no aportan mucha información (como “el”, “la”, “y”, “o”), y se hizo el reemplazo de los números que estaban escritos como dígitos en letras, es decir, por ejemplo, los 1, 2 y 3 pasaron a ser uno, dos y tres.

Así mismo, se dividió cada opinión en palabras individuales y se simplificó a su forma básica. Por ejemplo, las palabras “corriendo”, “correr” y “corrió” las simplificamos todas a “correr”. Esto ayuda a entender mejor el significado general de cada opinión.

Este proceso se aplica a cada reseña agregada dentro de la aplicación pues, como vimos en la fase 1, nos ayudó a tener la base para implementar diferentes modelos de procesamiento de lenguaje natural que ayuden posteriormente a entender mejor las opiniones de los turistas y a descubrir qué características hacen que un lugar turístico sea atractivo. Con esta información, se puede ayudar a mejorar los lugares que no son tan populares y a promover el turismo en Colombia.

El modelo utilizado fue un modelo de Naive Bayes, el cual es uno de los algoritmos de Machine Learning más utilizados para clasificar y predecir una clase de un conjunto de datos, basado en seguir las características de una distribución de probabilidad, y siendo la más popular para problemas de clasificación de texto (como este problema). Existen varios tipos de modelo como lo son el Gaussiano, el multinomial y el de bernoulli, los cuales, al igual que con Support Vector Machines, serán probados empíricamente cuál modelo es mejor con las pruebas. Por lo tanto, para esta selección de modelo, se intentará probar 2 modelos, los cuales son el multinomial y la distribución de Bernoulli y para el mejor modelo se escogerá aquel que tenga mejores métricas en prueba:

	Naive Bayes
Exactitud	0,49
Precisión	0,5
Recall	0,45
F1-Score	0,46

Se guardó el modelo entrenado para su uso posterior en la predicción, esto se hace exportando el pipeline, para así poderlo usar dentro de la aplicación web con el siguiente código:

```
joblib.dump(model_pipeline, 'model.pkl')
```

Para permitir a los usuarios enviar reseñas y obtener predicciones de calificación de forma sencilla, se creó una API utilizando Flask, un marco de desarrollo web ligero para Python. Dentro de esta API, se integró el modelo de aprendizaje automático previamente entrenado para realizar las predicciones de calificación en tiempo real.

La API consta de un endpoint /predict que espera recibir una solicitud POST con un cuerpo JSON que contenga la reseña del sitio turístico. Cuando se

recibe una solicitud, la API preprocesa la reseña utilizando las mismas técnicas de procesamiento de texto que se utilizaron durante el entrenamiento del modelo. Luego, utiliza el modelo entrenado para realizar la predicción de calificación basada en la reseña proporcionada.

Una vez que se realiza la predicción, la API devuelve la calificación predicha en formato JSON como respuesta a la solicitud del usuario. Este enfoque proporciona una forma eficiente y conveniente para que los usuarios interactúen con el modelo de predicción sin necesidad de comprender los detalles internos del proceso de modelado o preprocesamiento de datos.

Al proporcionar esta API, los usuarios finales pueden integrar fácilmente la funcionalidad de predicción de calificación en sus propias aplicaciones, lo que les permite obtener rápidamente estimaciones de calificación para las reseñas de sitios turísticos que ingresen.

Desarrollo de la aplicación y justificación.

Para implementar una forma de saber en dónde el hotel tenía oportunidades de mejora o no, se creó un asistente virtual que ayuda a las diferentes entidades del sector turístico a mejorar sus servicios. Funciona analizando las opiniones que los visitantes dejan en varios sitios turísticos.

Se pueden cargar de forma manual las reseñas de diferentes clientes por medio de un archivo .csv o cargarlas directamente una por una. Este una caja de sugerencias en la que los clientes dejan sus comentarios sobre su estancia. Pero en lugar de tener que leer cada comentario uno por uno y sacar conclusiones, este asistente virtual lo hace de forma automática.

Primero, el asistente recoge todas las reseñas de los visitantes. Luego, las analiza para entender qué es lo que más les gustó a los visitantes y qué áreas necesitan mejoras.

Para interactuar con este asistente, los usuarios utilizan una interfaz de usuario amigable. Esta interfaz permite a los usuarios cargar las reseñas y visualizar los resultados del análisis.

En el lado del servidor, tenemos un motor de análisis que procesa las reseñas. Este motor es capaz de entender el lenguaje natural y extraer información útil de las reseñas.

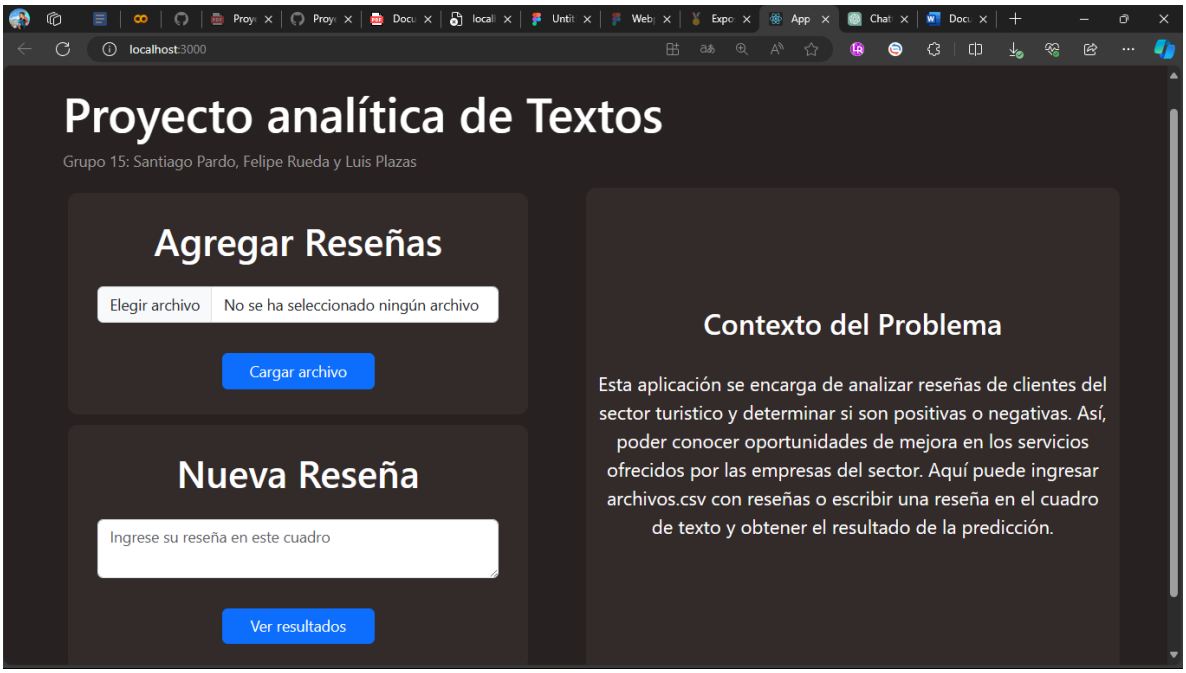
En resumen, este proyecto es como un asistente personal para los hoteles, ayudándoles a entender mejor las necesidades de sus clientes y a mejorar sus servicios en consecuencia.

Para conectar la lógica de nuestra modelo se usó Fast-API para conectar nuestro modelo del Back-end con el Front-end. Este último fue desarrollado en React.

Dentro de la aplicación, desarrollamos principalmente 3 funcionalidades:

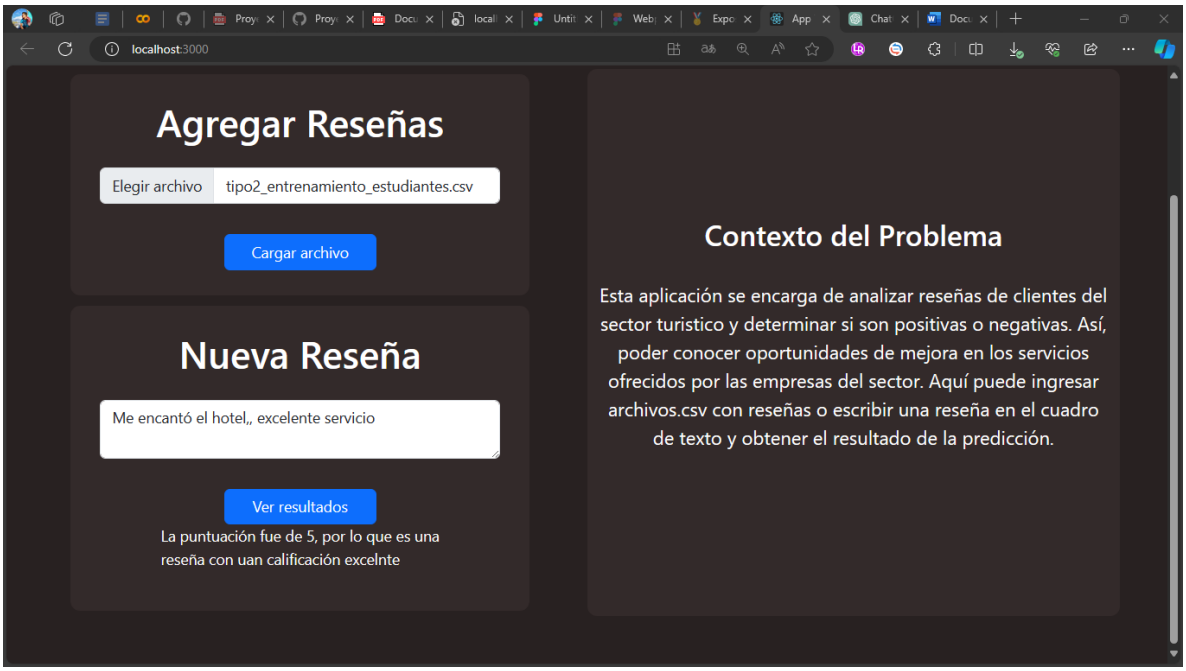
- Subir archivos .csv para analizar reseñas
- Agregar nuevas reseñas
- Predecir el resultado numérico de estas nuevas reseñas.

A continuación, podemos ver la interfaz web de la aplicación.



Resultados.

Como podemos ver a continuación, la aplicación es capaz de crear un modelo de NLP predictivo a través de un archivo .csv con todas las reseñas, entrenarlo y recibir nuevas reseñas para predecir su puntuación esperada como se ve en la imagen a continuación:



Así mismo, podemos ver que funciona correctamente la aplicación y le dice al usuario a qué hace referencia su puntuación numérica. Algo increíblemente valido para diferentes entidades turísticas, pues, gracias a las reseñas de cierto servicio en especial, se puede conocer la opinión promedio de los clientes.

(10%) Trabajo en equipo.

Resumen Ejecutivo: Este documento resume las actividades y responsabilidades del equipo encargado del proyecto de análisis turístico. El objetivo es evaluar las características de los sitios turísticos y su impacto en la popularidad y recomendaciones de los turistas.

Roles y Responsabilidades:

● Líder del Proyecto (Santiago Pardo): Responsable de la gestión general del proyecto, incluyendo la planificación de reuniones, distribución equitativa de tareas y la subida de entregables. Tiene la autoridad final en decisiones sin consenso.

● Líder de Negocio (Luis Felipe Plazas y Felipe Rueda): Asegura que el proyecto resuelva el problema identificado y esté alineado con la estrategia del negocio. Coordina con expertos estadísticos y comunica efectivamente el producto.

● Líder de Datos (Santiago Pardo): Gestiona los datos utilizados en el proyecto y asigna tareas relacionadas con los mismos, asegurando su disponibilidad para el equipo.
● Líder de Analítica (Santiago Pardo): Supervisa las tareas analíticas y verifica que los entregables cumplan con los estándares de análisis y restricciones del proyecto.

Cronograma de Reuniones:

Fecha	Tipo reunión	Descripción	Medio
8/04/2024	*Reunión Lanzamiento y planeación*	Definición de roles y metodologías de trabajo, además de generar ideas.	Zoom
13/04/2024	*Reunión de seguimiento*	Monitoreo de el progreso y ajuste de estrategias.	Zoom
20/04/2024	*Reunión de finalización*	Revisión de resultados finales, y discusión de pasos a seguir.	Zoom

Repartición de los 100 puntos:

- *Felipe Rueda:* 33 puntos
- *Santiago Pardo:* 34 puntos
- *Luis Felipe Plazas:* 33 puntos

Como aspecto a mejorar, en próximas entregas se hará una implementación un código de más calidad y se hará una mejor comunicación para realizar el trabajo conjunto, tanto con los expertos en datos como los analistas de datos. Conclusiones: El equipo ha demostrado una colaboración efectiva, cumpliendo con los roles asignados y manteniendo una comunicación constante. Los resultados obtenidos serán fundamentales para mejorar la experiencia turística y fomentar el turismo en Colombia.