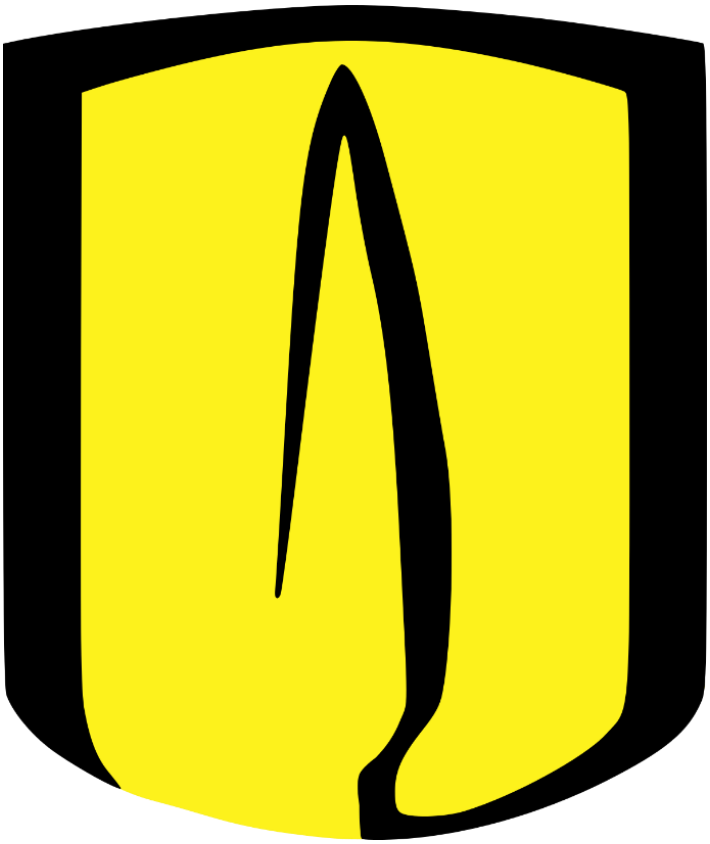
Etapa 2 - Automatización y uso de modelos de analítica de textos Turismo de los Alpes



Integrantes:

María Camila Luna Velasco - 201920993

Juan Manuel Jauregui Rozo - 201922481

Ana Sofía Castellanos Mosquera - 202114167

Universidad de Los Andes

2024-2

Tabla de contenido

1. Proceso automatización

2. 1 El entendimiento de datos se encuentra en NoteBook completo

2.1 Construcción Pipeline

A partir del proceso de preparación de datos llevado a cabo durante la etapa 1, se identificaron los procesos clave de preparación de datos, los cuales se mencionan a continuación:

1. Función softPreprocessing: Encargada de realizar un preprocesamiento ligero con el objetivo de preparar los datos para su posterior procesamiento por el Lemmatizer. Esta función incluye en primer lugar la eliminación de números, dado que estos no se transforman adecuada una vez pasan por tokenizador. En segundo lugar, cada una de las palabras de los textos se transforma a minúscula y finalmente se remueven los caracteres ASCII. Cabe resaltar que esta función tiene un paso intermedio llamado getTokens() que transforma el str en una lista con las palabras, esto se hace para poder aplicar la función de eliminación ASCII y llevar a minúsculas las palabras
2. Función applyLemmatizer: Encargada de generar la lematización de los textos. Esto es cambiar las conjugaciones de las palabras a su forma infinitiva y/o registrada en el diccionario.
3. Función Tokenizer: Implementa la funcionalidad WordPunctTokenizer() de la librería NLTK, la cual genera la tokenización considerando los caracteres especiales y la puntuación.
4. Función preProcessing: Aplica nuevamente el paso a minúsculas (en caso de que el lemmatizer devuelva palabras con mayúsculas), se aplica eliminación de puntuación y se remueven palabras de parada (conectoras y otras que no aportan información relevante).
5. Vectorizer: A partir de los textos preprocesados estos se vectorizan mediante el método tfidf para dejarlo en un formato que pueda ser luego procesado por el algoritmo.
6. Algoritmo SVC: De la etapa uno se encuentra que el mejor modelo es el de Support Vector Machines. Debido a esto, es el que se propone en el pipeline.

Tras la creación del pipeline este se ajusta al conjunto de entrenamiento (hallar los parámetros y crear vectorizer con el que se transformaran datos futuros) y se persiste el modelo mediante dump() de la librería joblib para mandarlo posteriormente a producción.

El archivo con el pipeline y su debida persistencia se encuentra en el repositorio con el nombre NotebookCompletoEtapa2.ipynb.

2.2 Construcción modelo y acceso API

Una vez se persiste el modelo el modelo se genera un acceso a este a través de una API mediante el framework FASTAPI. Para ello se crea un proyecto en este framework y dentro del folder aplicación se crean los archivos main.py.

En este archivo se crean el método get para verificar el funcionamiento de la aplicación y post (predict) con el objetivo de hacer la respectiva predicción de una reseña que ingresa por parámetro.

Una vez se llama al método post, en main.py este utiliza el pipeline previamente automatizado en modeloLemmatizer.joblib y con este se genera la clase predicha

que es devuelta finalmente en el método para su posterior integración en el API

3. Desarrollo de la aplicación

Descripción del usuario/rol de la organización:

El usuario principal de la aplicación desarrollada sería el personal del Ministerio de Comercio, Industria y Turismo de Colombia, así como miembros de la Asociación Hotelera y Turística de Colombia (COTELCO). También incluiría a gerentes y propietarios de cadenas hoteleras como Hilton, Hoteles Estelar, Holiday Inn, y hoteles más pequeños ubicados en varios municipios de Colombia.

Esta aplicación estaría directamente relacionada con el proceso de negocio de análisis y promoción del turismo en Colombia. Permitiría a los usuarios analizar las características de los sitios turísticos que influyen en su atractivo para los turistas, comparar estos sitios con otros que han recibido bajas recomendaciones, y determinar la calificación potencial de un sitio por parte de los turistas. La importancia de esta aplicación para estos roles radica en su capacidad para proporcionar información valiosa y análisis predictivos que pueden utilizarse para identificar oportunidades de mejora, aumentar la popularidad de los sitios turísticos y, en última instancia, fomentar el turismo en Colombia.

Trabajo transdisciplinar con el grupo de estadística:

Durante el desarrollo de la aplicación, el equipo de ingeniería de software trabajó estrechamente con el grupo de estadística para definir la aplicación a desarrollar y el tipo de usuario.

Definición de la aplicación y tipo de usuario:

Aplicación por desarrollar: La aplicación se definió como una plataforma web que permitirá a los usuarios ingresar reseñas de sitios turísticos y recibir predicciones sobre la calificación potencial de esos sitios por parte de los turistas.

Decisión y justificación:

La decisión de desarrollar una aplicación web se tomó debido a su accesibilidad y facilidad de uso para los usuarios. Además, al ser una plataforma en línea, permite un acceso rápido y conveniente desde diferentes ubicaciones. Respecto al tipo de usuario, se consideró crucial incluir a aquellos directamente involucrados en la industria del turismo en Colombia, ya que son quienes pueden beneficiarse directamente de la información y análisis proporcionados por la aplicación.

Validación y ajustes -Colaboración con equipo de estadística

Para el proceso de validación el equipo de estadística realizo una comprobación de la aplicación desde el rol de usuario. Así mismo, En este caso, en conjunto con el trabajo interdisciplinario llevado a cabo entre los dos grupos, se determinó que teniendo en cuenta el problema el grupo de sistemas consiguió junto con la aplicación crear un método de predicción estadísticamente significativo para entender y predecir las reseñas de los sitios turísticos en Colombia.

Lo anterior se puede afirmar dado que crearon una propuesta de valor enfocada en suplir las necesidades del target market que tiene una precisión del 48%. Adicionalmente, gracias a su interfaz intuitiva la UX de la aplicación consigue ser una herramienta no solo precisa, como fue demostrado con la prueba de hipótesis hecha en el primer ejercicio en conjunto, si no fácil de usar y una oportunidad de negocio que resuelve el problema de caso.

Por último, la propuesta de valor para el usuario es poder revisar y comparar

recomendaciones de sitios turísticos en Colombia. También predecir posibles reseñas de un sitio turístico en el país. Esto no solo es bueno para los turistas, si no también es una oportunidad excelente para el diseño de políticas públicas enfocadas en generar impacto turístico, así como para fomentar la inversión de grandes empresas hoteleras o pequeños emprendedores.

4. Resultados

El video con los resultados – aplicación desplegada - está en este enlace: https://github.com/acastem15/Inteligencia-de-Negocios/wiki/Video-E2

5. Trabajo en equipo

Los roles asignados a cada uno de los miembros, así como las tareas que hizo cadauno, se muestran a continuación.

Juan Manuel Jauregui Rozo

Rol: Ingeniero de software responsable desarrollar la aplicación final e Ingeniero de software responsable del diseño de la aplicación y resultados
Tareas: Responsable de desarrollar la aplicación final, se encargó de convertir los diseños y especificaciones en código funcional. Esto implica escribir el código, integrar los componentes necesarios, probar la aplicación. Además, fue responsable de implementar las funcionalidades requeridas, como la interacción con el modelo analítico, la presentación de resultados y la garantía de una experiencia de usuario fluida.

También fue responsable de diseñar cómo se presentan los resultados del modelo analítico al usuario, asegurándose de que sean comprensibles y útiles para el usuario final. En resumen, este ingeniero se centraría en la estética y la usabilidad de la aplicación, asegurando que sea atractiva y efectiva para los usuarios.

María Camila Luna Velasco

Rol: Líder del proyecto e Ingeniero de software responsable de desarrollar la aplicación final
Tareas: Se encargó de revisar el documento final entregado, el video entregado, el notebook entregado y la completitud de cada una de las secciones solicitadas para la entrega. Adicionalmente, se encargó de gestionar las reuniones que se realizaron para el desarrollo del proyecto. En cada reunión, se plantearon tareas e hitos que se debían cumplir antes de la siguiente reunión. Este integrante también trabajó el algoritmo de regresión logística y se encargó de redactar y acotar los resultados finales en el documento y en el video. Este proceso tomó aproximadamente 6 horas distribuidas en los días de desarrollo del proyecto. Por otro lado, se enfocó en la parte visual y de experiencia de usuario (UX). Esto incluiría diseñar la interfaz de usuario de la aplicación, crear prototipos, definir la arquitectura de la información y garantizar que la aplicación sea intuitiva y fácil de usar.

Ana Sofía Castellanos Mosquera

Roles: Líder de datos

Tareas: El ingeniero de datos se asegura de que todo el proceso de automatización relacionado con la construcción del modelo analítico se lleve a cabo con alta calidad. Esto incluye la preparación de datos, la construcción y entrenamiento del modelo, su almacenamiento y la creación de una API para acceder a él. Su papel es crucial para garantizar que el modelo sea eficiente, escalable y pueda utilizarse de manera efectiva por otros equipos. El proceso de construcción y validación del pipeline que genera la automatización junto con la construcción del API en el framework FASTAPI del proceso tomo un aproximado de 10h distribuidas a lo largo de la semana previa a la entrega del proyecto.

6. Entregables

Todos los entregables se encuentran en este enlace: https://github.com/acastem15/Inteligencia-de-Negocios

3. Trabajo en equipo

4. Entregables

El enlace al repositorio donde se encuentran los entregables es este: https://github.com/acastem15/Inteligencia-de-Negocios/wiki

Para ver en más detalle los tableros de control tanto de entendimiento como de resultados, se puede consultar en GitHub con el nombre: finalTablerosControl.pdf y finalTablerosControl.pbix

5. Referencias

[1] F. Murzone, “Procesamiento de Lenguaje Natural: Stemming y Lemmas,” EscuelaDeInteligenciaArtificial. Accessed: Apr. 06, 2024. [Online]. Available: https://medium.com/escueladeinteligenciaartificial/procesamiento-de-lenguaje-natural-stemming-y-lemmas-f5efd90dca8

[2] A. Jha, “Vectorization Techniques in NLP [Guide],” neptune.ai. Accessed: Apr. 06, 2024. [Online]. Available: https://neptune.ai/blog/vectorization-techniques-in-nlp-guide

[3] Gitlab, ISIS3301, Procesamiento de Textos, Accessed: Apr. 06, 2024. [Online]. Available: https://gitlab.virtual.uniandes.edu.co/ISIS3301/practicas/blob/master/ProcesamientoTextos/Preparaci%C3%B3n_de_textos_estudiante.ipynb

[4] T., Joachims. Text categorization with Support Vector Machines.https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf