

**Universidad de los Andes**  
**Ingeniería de Sistemas y Computación**  
**Inteligencia de negocios**

**Turismo de los alpes**

**Etapa 2. Automatización y uso de modelos de analítica de textos**

**Integrantes (GProy26)**

- Henry Santiago Antolinez - 202121785
- Juan David Orduz - 202123170
- Abel Arismendy - 202020625

**Grupo de estadística**

- Gabriela Coronado
- Alejandra Ramos

<b>Proceso de automatización del proceso de preparación de datos, construcción del modelo.....</b>	<b>1</b>
<b>Persistencia del modelo y acceso por medio de API.....</b>	<b>3</b>
<b>Desarrollo de la aplicación y justificación.....</b>	<b>3</b>
<b>trabajo transdisciplinar.....</b>	<b>3</b>
<b>Resultados:.....</b>	<b>6</b>
<b>Trabajo en equipo:.....</b>	<b>7</b>

**Proceso de automatización del proceso de preparación de datos, construcción del modelo**

Para poder automatizar el proceso de la preparación de los datos y la construcción del modelo fue necesario reestructurar todo el proceso del notebook para la etapa 1. Esto se debe a que no hubo implementación de pipelines en dicha entrega. El pipeline en este apartado nos permite secuencialmente aplicar transformaciones y en esencia poder automatizar los procesos mencionados anteriormente. Como fue necesario re estructurarlo, desde la propia aplicación se construyeron las siguientes clases:

**Tokenización:**

En este “step” se divide el texto en palabras individuales.

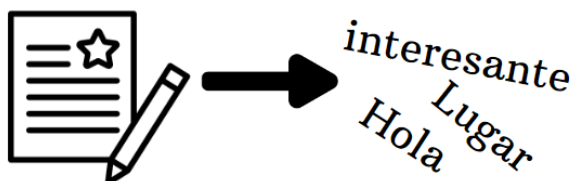


figura 1. representación básica de tokenización

### NoiseRemover:

En este apartado se eliminan todas las palabras consideradas ruido, como lo son en este caso las no ASCII o signos de puntuación

### Lemmatizer:

En este “step” se hace la lematización. La lematización es capaz de manejar palabras irregulares y excepciones gramaticales lo que nos permite normalizar los datos de una forma precisa y eficiente

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

figura 2. Ejemplo del funcionamiento de la lematización y comparación con Stemming

### stopwords\_remove:

En este “step” se eliminan todas las stopwords con determinantes, preposiciones, pronombres y conjunciones

### dataframe\_transformerSVM:

En este “step” es donde se desarrolla la transformación de los datos a modo que el modelo SVM funcione y de la misma forma logra hacer el modelo de forma correcta.

### model:

En este último “step” es donde se hace el modelamiento SVM que fue el que el que mejor funcionó en la etapa 1

Todos los pasos anteriores son necesarios para automatizar todos los pasos realizados sobre los datos en su construcción del modelo

La serialización de un modelo solo incluye la estructura y configuraciones realizadas sobre el pipeline, más no las instancias de los objetos que lo componen. Pues estos son provistos por la librería, por medio de la importación, en cualquiera que sea su ambiente de ejecución. Esto significa que si usted construye transformaciones personalizadas, debe incluir por separado estas en el ambiente donde cargará y ejecutará el modelo una vez sea exportado, ya que estas no están incluidas en la serialización

## **Persistencia del modelo y acceso por medio de API**

El equipo desarrolló una API utilizando FastAPI para acceder a un modelo de predicción de revisiones. El modelo se carga desde un archivo .pkl y se emplea para realizar predicciones sobre revisiones. El endpoint GET /predict/ está diseñado para manejar solicitudes GET, ofreciendo predicciones basadas en un dataframe previamente cargado. Por otro lado, el endpoint POST /predict/single/ espera solicitudes POST que contengan datos JSON con texto de revisión individual. Al recibir estos datos, realiza una predicción basada en el texto proporcionado y devuelve el resultado.

En el caso del endpoint GET /predict/, el resultado son todas las reseñas en el archivo cargado con la calificación dada por el modelo y las palabras más importantes que el modelo usó para darles esa calificación. Asimismo, el resultado del endpoint POST /predict/single/ retorna la calificación de la reseña que el usuario pasó como parámetro.

## **Desarrollo de la aplicación y justificación**

En este apartado se presentan los comentarios dados por el grupo de estadística, nuestras opiniones al respecto y las decisiones finales que se tomaron para el desarrollo de la aplicación.

### **trabajo transdisciplinar**

En este apartado se menciona la interpretación del proyecto por parte del grupo de estadística **(reunión 15 abril)**

### **¿Cómo entendemos el proyecto?**

El proyecto tiene como objetivo principal mejorar el sector turístico de Colombia a través de la implementación de un programa que permite la clasificación de palabras en reseñas de lugares turísticos colombianos con el ánimo de darle un puntaje en la escala de estrellas (1 a 5) e identificar los problemas específicos que generan una oportunidad de mejora en cada uno de ellos. Para la metodología, inicialmente se hace una preparación de datos con el fin de facilitar la interpretación de los mismos a partir de la implementación del modelo BAG OF WORDS (BOW), el cual interpreta las reseñas como un conjunto de palabras, dejando a un lado la gramática y centrándose en la repetición de las mismas. Después, a través del preprocesamiento se limpian y preparan los datos para un análisis exitoso. Esto incluye la tokenización y la normalización. Finalmente, se escogió el algoritmo de clasificación SVC (Support Vector Classifier) puesto que fue el que dio mejores resultados al ser comparado con otros modelos.

Usuarios consideramos que la creación de la aplicación sería de gran utilidad e interés para dos perfiles: los turistas y los dueños o socios de los lugares.

### Aplicación Web

Para la aplicación Web del programa, consideramos que es innovador e importante implementar dos tipos de sesiones para dos tipos de actores diferentes: los turistas y los dueños de los establecimientos, esto debido a que cada uno busca algo diferente a la hora de hacer uso del programa. Mientras que los visitantes buscan encontrar la mejor opción de establecimiento con respecto a las estrellas, los dueños y miembros del establecimiento buscan encontrar la raíz del problema para establecer una ruta clara de acción. Adicionalmente, también se debe aplicar un sistema que permita subir los archivos que contienen las reseñas.

### Para los turistas

Una vez se suban los archivos con las reseñas, sería ideal que se mostrará la clasificación de los establecimientos con respecto a las estrellas que se le atribuyen por medio del programa, respaldado por las reseñas y resaltando los problemas y las fortalezas con el ánimo de que los turistas tengan en cuenta el panorama completo a la hora de hacer su elección. También sería interesante aplicar filtros de búsqueda para que la información pueda ser digerida correctamente o analizada desde diferentes perspectivas.

1

Escoge tu establecimiento de interés

2



Sube las reseñas que los visitantes han hecho del establecimiento

El establecimiento en el que estas interesad@ tiene..



3 estrellas

Fortalezas:



Limpeza



Comida



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Debilidades:



Infraestructura



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

figura 3. idea de como se debería ver para turistas

### Para los miembros del establecimiento

Una vez se suban los archivos con las reseñas, sería ideal que se mostraran las debilidades del establecimiento (ej: limpieza, infraestructura, comida..) respaldado por las reseñas que las personas que lo han visitado. Adicionalmente, sería interesante disponer de propuestas de soluciones para el problema que se está atravesando y al cual se atribuyen el número de estrellas del lugar



figura 4. idea de como se debería ver para miembros de establecimientos

Como se puede observar, pese a haber establecido como negocio al ministerio en la etapa anterior, nuestras compañeras pensaron en hacer algo más personalizado para las experiencias de usuario y de dueños de establecimientos que se aleja mucho de la idea inicial. No obstante, de las ideas aportadas se conversó y en conjunto se llegó al acuerdo de desarrollar la aplicación de la siguiente manera:

El backend de la aplicación web inicia configurando las importaciones necesarias y variables globales. Luego, establece rutas para cargar archivos CSV, realizar predicciones sobre los datos, y predecir individualmente sobre una reseña. También incluye una ruta para obtener las reseñas clasificadas. Cada ruta maneja casos de error apropiadamente, como archivos incorrectos o datos faltantes.

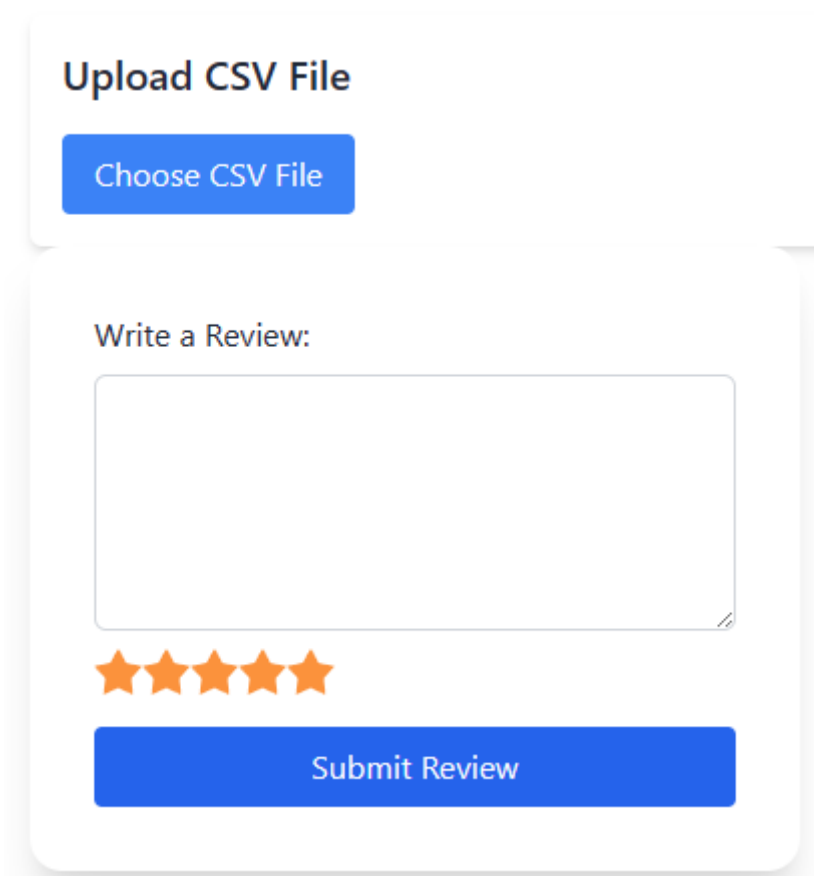
así mismo el backend de la aplicación web incluye la definición de dos clases y la configuración de herramientas de procesamiento de texto.

La clase DataModel utiliza Pydantic para definir un modelo que representa los datos esperados en formato JSON. Proporciona métodos para obtener los nombres de las columnas y convertir los datos a un diccionario.

La clase Model carga un modelo de machine learning desde un archivo y proporciona métodos para hacer predicciones tanto para un conjunto de datos como para una única entrada de texto.

Además, se definen clases para realizar transformaciones de texto, como tokenización, eliminación de ruido y lematización, utilizando librerías como NLTK. También se configuran listas de palabras vacías específicas del idioma español.

## Resultados:



The image shows a user interface with two main sections. The top section is titled "Upload CSV File" and contains a blue button labeled "Choose CSV File". The bottom section is titled "Write a Review:" and contains a large text input area. Below the text input area is a row of five orange stars, indicating a rating system. At the bottom of the review section is a blue button labeled "Submit Review".

figura 5. cuadro de texto para subir reseñas individuales y predecirlas

Write a Review:

Estuvimos en el hotel las navidades de 2016. El hotel tenía cucarachas (pequeñas) en el baño- Tuvimos problemas con el transfer y nadie del hotel se dignaba a ayudarnos, tuvimos

Predicted Rating: 1



Submit Review

figura 6. Reseña predecida

Write a Review:

Es un bonito lugar, las entradas son económicas, lástima que no abren los Lunes festivos. Caminar y degustar un café por la zona es un excelente plan.

Predicted Rating: 5



Submit Review

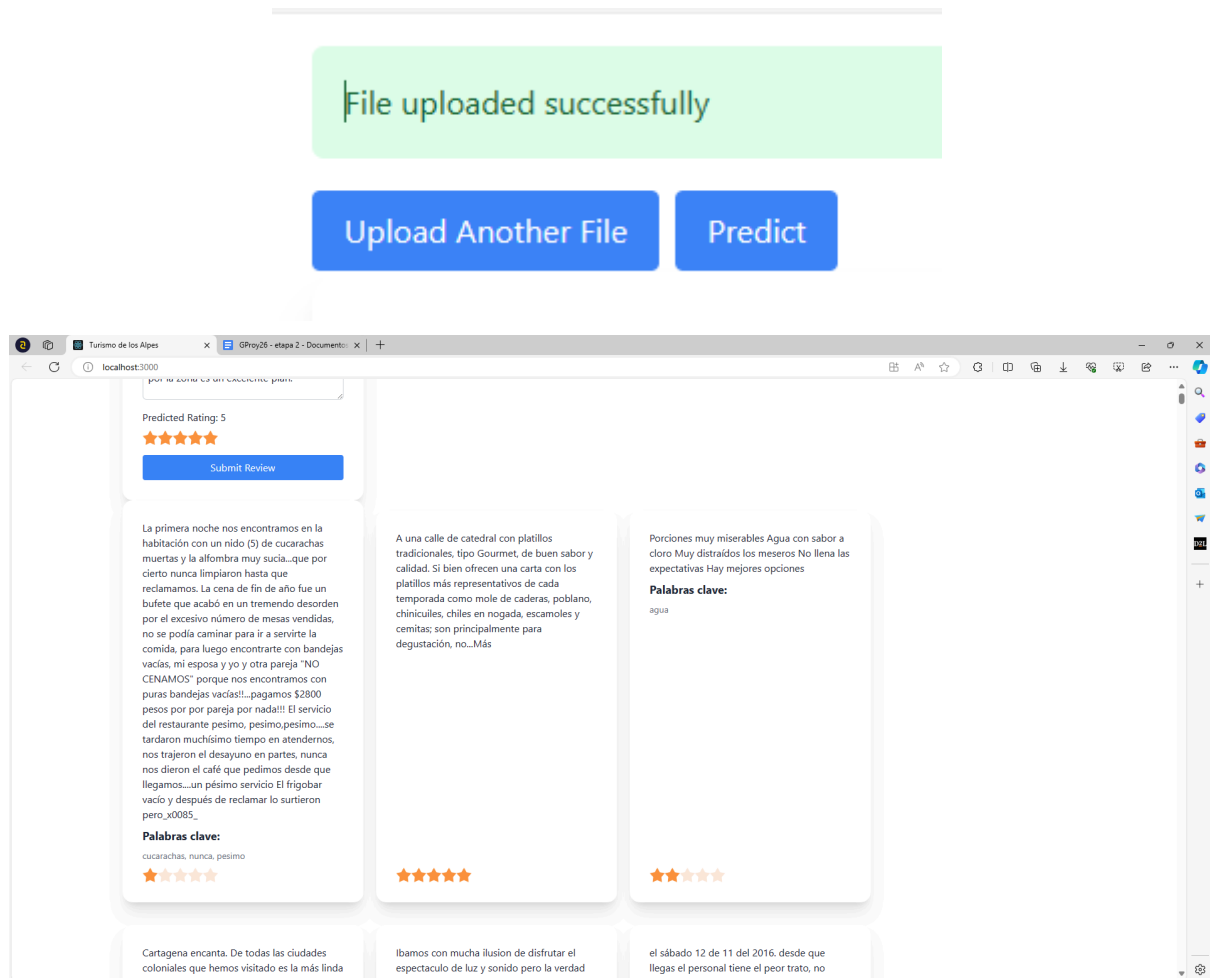


figura 7. Resultados de predicción

## Trabajo en equipo:

### Líder de Proyecto (Abel Arismendy) :

- Define fechas y horarios para reuniones (0.2 h)
- Establece pre-entregables y plazos para el grupo (0.2 h).
- Supervisa y verifica las asignaciones de tareas para garantizar una distribución equitativa de la carga de trabajo. (2 h)
- Sube la entrega final del grupo a la plataforma o lugar designado (0 h)
- Resuelve discrepancias y toma decisiones finales en caso de falta de consenso. (1h)

### Ingeniero de Datos (Juan David Orduz) :

- Garantiza la calidad del proceso de automatización relacionado con la construcción del modelo analítico. (3 h)
- Realiza la extracción, transformación y carga de datos según las necesidades del proyecto. (1 h)
- Limpia y preprocesa los datos para su posterior análisis. (1h)



**Ingeniero de Software Responsable del Diseño de la Aplicación y Resultados (Henry Santiago Antolinez) :**

- Lidera el diseño de la interfaz de usuario de la aplicación. (2 h)
- Trabaja en colaboración con otros miembros del equipo para definir los requisitos de la aplicación. (3 h)
- Crea el video que muestra los resultados obtenidos del modelo analítico. (2 h)

**Ingeniero de Software Responsable de Desarrollar la Aplicación Final (Abel Arismendy) :**

- Gestiona el desarrollo de la aplicación desde la fase de diseño hasta la implementación. (1 h)
- Escribe código limpio y bien documentado para la aplicación. (2 h)

El trabajo en equipo ha sido fundamental para el desarrollo de este proyecto. Cada miembro del equipo ha desempeñado un papel crucial y ha dedicado un tiempo considerable para garantizar que cumplimos con nuestros objetivos.

En cuanto a la distribución de los puntos, se repartieron de manera equitativa entre los miembros del equipo. Cada uno contribuyó de manera significativa al proyecto demostrando un alto nivel de compromiso y dedicación. Por lo tanto, proponemos que cada miembro del equipo reciba 33.3 puntos.

▪ **Reuniones:**

fecha	Reunión
15 abril	Basado en la etapa anterior se compartió con el equipo de estadística los requerimientos de la aplicación y se llegó a un acuerdo conjunto
20 abril	Exposición de resultado final