Construcción de modelos de analítica de textos

Tabla de contenido

Entendimiento del negocio y enfoque analítico

Descripción de la oportunidad del negocio

La oportunidad o el problema del negocio se centra en el análisis de características de sitios turísticos para comprender qué los hace atractivos para los turistas locales y extranjeros. Además, se busca comparar estos sitios con aquellos que han recibido bajas recomendaciones y afectan la afluencia de turistas. El objetivo es determinar la calificación de los sitios según las reseñas de los turistas y aplicar estrategias para mejorar su popularidad y fomentar el turismo.

Para lograrlo, se dispone de dos conjuntos de datos con reseñas y calificaciones de los sitios turísticos. El análisis independiente de estos datos permitirá tomar decisiones informadas y colaborar con científicos e ingenieros de datos en el desarrollo del proyecto.

Enfoque analítico

Para poder cumplir satisfactoriamente con los objetivos planteados, se deciden emplear modelos de aprendizaje supervisados. Esto debido a que ya se cuentan con dos sets de datos (reseñas) que cuentan con las calificaciones. La tarea de los modelos es de clasificación, dado que debe poder terminar la calificación de un sitio dadas sus reseñas clasificando estas en distintas calificaciones. En cuanto a las técnicas utilizadas, se realizan 3 modelos con técnicas diferentes. En primer lugar, se utiliza una regresión logística como primer modelo. Para el segundo modelo se utiliza un Random Forest. Y para el último modelo, se decide utilizar SVC (Support Vector Classifier). Se utilizan tres diferentes técnicas y algoritmos con el fin de determinar de mejor forma que modelo se logra acoplar mejor al objetivo planteado.

Organización beneficiada

Comprender las características clave que influyen en las calificaciones de los sitios turísticos permitirá a los científicos de datos identificar patrones y tendencias relevantes. Esto les ayudará a determinar la relación entre la calidad del servicio y las calificaciones o la influencia de las atracciones turísticas en la satisfacción del cliente. En cuanto a las organizaciones involucradas en este caso están El Ministerio de Comercio, Industria y Turismo de Colombia, así como la Asociación Hotelera y Turística de Colombia ya que esta información les ayudará a mejorar la competitividad del sector turístico colombiano y a desarrollar estrategias para atraer más visitantes.

El ingeniero de datos será fundamental en la implementación práctica del modelo, trabajando en conjunto con las organizaciones mencionadas anteriormente. Se asegurará de que el modelo sea escalable, eficiente y persistente, lo que significa que puede manejar grandes cantidades de datos de manera eficiente y confiable. De igual forma, el ingeniero de software construirá la aplicación web o móvil que permitirá a los usuarios finales interactuar con el modelo, en colaboración con las organizaciones que se beneficiarán del proyecto. Esto facilitará la toma de decisiones basada en las recomendaciones del modelo, como la elección del mejor hotel para un viaje o la planificación de un itinerario turístico.

En este contexto, las organizaciones que más se verán involucradas son las cadenas hoteleras como Hilton, Hoteles Estelar y Holiday Inn además de los hoteles locales.

Contacto con experto externo

Para poder cumplir satisfactoriamente con los objetivos y asegurarnos de comprender bien cómo funciona la organización, se planea una reunión de validación del enfoque con estudiantes de estadística el día 7 de abril del 2024. Las estudiantes encargadas de esta validación son:

- Eivy Miyirelly Torres Quiñones
- Catalina Zamora Gonzalez

Entendimiento y preparación de los datos

Perfilamiento de los datos

Para el perfilamiento de los datos, lo primero que se realizó fue extraerlos de los archivos csv para una mejor comprensión de estos. Ambos sets de datos traían la misma forma, una matriz de 7875 filas representando cada uno de los registros con tan solo 2 columnas. Cuando se inspeccionaron de qué forma venían los datos, nos dimos cuenta de que cada registro cuenta con una variable Review que es una cadena de caracteres que representa la reseña completa realizada por un usuario de un sitio turístico, la segunda variable es Class que es una variable numérica que representa la calificación dada por el usuario que escribió dicha reseña. También se pudo ver que, para ambos conjuntos de datos, la mínima calificación es 1, la máxima calificación es 5 y el promedio de las calificaciones es de 3,5.

Análisis de calidad de los datos

Completitud.

Para validar la completitud de los datos, lo primero que se realizó fue un cálculo del porcentaje de los valores nulos por columna. Tanto para el primer set de datos como para el segundo, este porcentaje fue 0%. Indicando así que ningún set de datos contaba con datos nulos, debido a esto, no fue necesario tomar ninguna acción al respecto.

Unicidad.

Al probar la unicidad de los conjuntos de datos, se imprimió la cantidad de datos duplicados por set. En este caso, el primer set tenía 109 registros duplicados, mientras que el segundo solo contaba con 102 registros duplicados.

Dada la naturaleza de los datos y que el modelo predictivo no se beneficia de datos duplicados, se tomó la decisión de eliminar todos los datos duplicados. Esto se realizó de esta forma dado que los registros representan la calificación de una reseña en específico, si se cuenta de nuevo con la misma reseña y calificación, esto podría llegar a perjudicar los modelos predictivos.

Consistencia.

En cuento a la consistencia de los datos, se revisó la columna categórica, en este caso la variable Review. Como esta variable son las reseñas escritas por usuarios, no se identificaron datos inconsistentes. Sin embargo, es prudente la limpieza de estas reseñas de palabras sin sentido, espacios, símbolos especiales y demás. Esto se realiza en el tratamiento de los datos.

Validez.

Por último, para asegurar la calidad de los datos obtenidos, se revisó la validez de estos. Para eso se revisó que las calificaciones no tuvieran valores negativos y todos se encontraran entre 1 y 5. En cuanto a las reseñas, al ser escritas por usuarios, estas eran validas. De esta forma, no fue necesario realizar ninguna acción al respecto.

Tratamiento de los datos

Modelo de regresión logística.

Para el modelo de regresión logística, dados los objetivos y la naturaleza del problema, lo primero que se identifico es que antes de construir el modelo, es necesario realizar un preprocesamiento de las reseñas para limpiar y normalizar el texto. Esto implica eliminar caracteres especiales, convertir el texto a minúsculas,

eliminar palabras vacías (stopwords) y realizar stemming para reducir las palabras a su forma base. Esto se realizó descargando una lista de stopwords en español y utilizando un SnowballStemmer en español. Al final del proceso las reseñas estaban todas en minúsculas, sin caracteres especiales, sin palabras vacías y con stemming. El stemming es el proceso de eliminar tanto prefijos como sufijos en las palabras, intentando llevarlas hasta la raíz. Un ejemplo sería si en diferentes reseñas se encuentran palabras como “bueno”, “buenisimo”, “buenisisimo”, “buenito” o “buenardo”, el stemming reduce todas estas palabras a “buen”. Esto se realiza para poder controlar de una mucho mejor forma las reseñas.

Una vez que las reseñas están preprocesadas, fue necesario representarlas en forma numérica para que puedan ser utilizadas por el modelo. La técnica utilizada en este caso fue utilizar la matriz TF-IDF (Term Frequency-Inverse Document Frequency) para asignar un peso a cada palabra en función de su frecuencia en la reseña y en el conjunto de reseñas.

Cuando ya se cuentan con las matrices, lo siguiente fue concatenar las matrices TF-IDF y las calificaciones para cada una de las reseñas. Como resultado se obtiene una matriz TF-IDF grande y una lista que contiene las reseñas y las calificaciones.

Modelo Random Forest.

Al igual que con el modelo de regresión logística, el primer paso es preprocesar las reseñas para limpiar y normalizar el texto. Esto implica eliminar caracteres especiales, convertir el texto a minúsculas, eliminar palabras vacías (stopwords) y realizar stemming para reducir las palabras a su forma base.

Una vez que las reseñas están preprocesadas, se representan en forma numérica utilizando la matriz TF-IDF (Term Frequency-Inverse Document Frequency) para asignar un peso a cada palabra en función de su frecuencia en la reseña y en el conjunto de reseñas. Se procede a entrenar un modelo de Random Forest con los datos preprocesados. Este modelo es un método de aprendizaje supervisado que funciona construyendo múltiples árboles de decisión y combinando sus resultados.

El modelo se evalúa utilizando varias métricas, incluyendo la precisión, el recall y el F1-score. Estas métricas proporcionan una visión completa del rendimiento del modelo. Para entender mejor cómo el modelo hace sus predicciones, se calcula la importancia de cada característica (en este caso, cada palabra). Esto se hace utilizando el atributo feature_importances_ del modelo, que proporciona un valor de importancia para cada característica. Las características se ordenan por importancia y se imprimen las 10 más importantes.

Modelo SVC (Support Vector Classifier).

Este modelo tiene los mismos pasos previos que los dos anteriores (limpiar, normalizar, vectorizar palabras, etc.). Para este caso, el único atributo que se le otorgó al modelo fue *kernel* = "*linear*". Este modelo busca encontrar un hiperplano que mejor separe las diferentes clases en el espacio de características.

Con el parámetro kernel='linear', busca un hiperplano lineal para realizar la separación. Este enfoque es útil cuando las clases son linealmente separables en el espacio de características. El objetivo es maximizar el margen entre las clases y minimizar la clasificación errónea, lo que resulta en un modelo de clasificación linealmente separable.

Modelado y evaluación

Modelo de regresión logística por Santiago Ramírez

En la construcción del modelo por regresión logística lo primero que se realizó fue dividir los datos en un conjunto de entrenamiento y un conjunto de prueba. El conjunto de entrenamiento se utilizará para entrenar el modelo y el conjunto de prueba se utilizará para evaluar su rendimiento.

Una vez se tienen divididos los datos, se crea un modelo de regresión logística limitado a un máximo de 1000 iteraciones. Para posteriormente, entrenar este modelo con los datos antes separados.

La regresión logística es un algoritmo de aprendizaje supervisado que se utiliza para la clasificación. Aunque se llama "regresión", es más adecuado para problemas de clasificación binaria (donde la salida puede ser una de las dos clases) y multiclase (donde la salida puede ser una de varias clases). Sacado de: ¿Qué es la regresión logística? - Explicación del modelo de regresión logística - AWS (amazon.com)

La regresión logística toma las características y las multiplica por unos coeficientes que se determinan en el entrenamiento, luego aplica la función logística que toma un número real y lo transforma en un valor entre 0 y 1, que podría ser la "probabilidad" de que esa reseña pertenezca a la calificación correspondiente. De esta forma, aquella calificación con mayor "probabilidad" sería donde se clasifica la reseña.

Como resultado de todo esto, el modelo de regresión logística tiene una precisión del 49.52% y una sensibilidad del 45.92%. Esto significa que de todas las predicciones que hizo el modelo, aproximadamente el 49.52% fueron correctas. En otras palabras, si el modelo predice 100 reseñas, podemos esperar que alrededor de 50 de esas predicciones sean correctas. De igual forma, el modelo fue capaz de identificar correctamente el 45.92% de las reseñas positivas reales.

El modelo es capaz de describir las palabras con mayor peso a la hora de clasificar una reseña en una calificación. Cabe recordar que las palabras fueron reducidas por un proceso de stemming a su raíz o parte principal. Para las calificaciones de 1 a 5, las palabras más relevantes fueron:

Calificación 1: ['pag', 'ningun', 'terribl', 'cucarach', 'rob', 'horribl', 'suci', 'mal', 'peor', 'pesim']

Calificación 2: ['car', 'sabor', 'des', 'desgraci', 'siqu', 'suci', 'pobr', 'parec', 'decepcion', 'mal']

Calificación 3: ['bien', 'buen', 'men', 'bastant', 'falt', 'pareci', 'demasi', 'regul', 'embarg', 'normal']

Calificación 4: ['histori', 'limpi', 'agrad', 'unic', 'disfrut', 'estupend', 'bien', 'excelent', 'comod', 'buen']

Calificación 5: ['gran', 'sup', 'hermos', 'perfect', 'maravill', 'recomend', 'encant', 'increibl', 'delici', 'excelent']

Podemos ver que en las calificaciones más bajas se encuentran palabras como sucio, horrible, mal o decepción. Mientras en las calificaciones más altas hay palabras como excelente, perfecto, hermoso o estupendo. Esto quiere decir que el modelo es capaz de identificar aquellas palabras en una reseña que hacen que esta tenga una buena o mala calificación.

Modelo de Random Forest por Mario Ruíz

En la construcción del modelo de Random Forest, lo primero que se realizó fue dividir los datos en un conjunto de entrenamiento y un conjunto de prueba. El conjunto de entrenamiento se utilizará para entrenar el modelo y el conjunto de prueba se utilizará para evaluar su rendimiento.

Una vez se tienen divididos los datos, se crea un modelo de Random Forest con 100 árboles de decisión. Posteriormente, se entrena este modelo con los datos antes separados.

Random Forest es un algoritmo de aprendizaje supervisado que se utiliza tanto para la clasificación como para la regresión. Es un conjunto de árboles de decisión, lo que significa que genera un bosque de una manera aleatoria. Cada árbol en el bosque genera una predicción y la clase que obtiene más votos se convierte en la predicción del modelo.

El algortimo toma las características y las utiliza para hacer divisiones en los nodos de los árboles de decisión. Cada árbol se construye de manera que minimice la impureza en sus nodos finales o hojas. La importancia de una característica se mide por cuánto reduce la impureza.

Como resultado de todo esto, el modelo de Random Forest tiene una precisión del 0.46 y una sensibilidad del 0.41. Esto significa que de todas las predicciones que hizo el modelo, aproximadamente el 46% fueron correctas. En otras palabras, si el modelo predice 100 reseñas, podemos esperar que alrededor de 0.46 de esas predicciones sean correctas. De igual forma, el modelo fue capaz de identificar correctamente el 41% de las reseñas positivas reales.

El modelo es capaz de describir las características con mayor importancia a la hora de clasificar una reseña en una calificación. Las 10 características más relevantes y sus respectivos weights fueron:

1. feature mal (0.010801210067890683)

2. feature excelent (0.010552403298136243)

3. feature buen (0.008426643092562643)

4. feature com (0.005987460641987274)

5. feature habit (0.00563642694640854)

6. feature servici (0.00547302310953717)

7. feature hotel (0.005406683890929452)

8. feature lug (0.0053825597703205195)

9. feature bien (0.0046811648599627284)

10. feature si (0.004559028918364147)

Modelo SVC por Santiago Gélvez

Estos fueron los valores de precisión, recall, y f1-score obtenidos para el modelo:

	precision	recall	f1-score	support
1	0.53	0.44	0.48	160
2	0.44	0.40	0.42	240
3	0.40	0.37	0.38	312
4	0.39	0.42	0.40	397
5	0.61	0.67	0.64	452
accuracy			0.48	1561
macro avg	0.47	0.46	0.47	1561
weighted avg	0.48	0.48	0.48	1561

Con un promedio ponderado de precisión, recall y f1-score de 0,48. Es decir, el 48% de las reseñas las acertó correctamente (recall), y el 48% de las reseñas que dijo que eran de una determinada calificación, realmente eran de esa calificación (precisión).

Modelo seleccionado: Regresión logística

Se seleccionó el modelo de Regresión logística ya que obtuvo el f1-score más alto.

Resultados

Luego de seleccionar el mejor modelo, se analizaron las 10 palabras más significativas para la clasificación de cada calificación.

Para las calificaciones de 1 y 2: palabras como “cucarach” y “suci” son indicadores de las precarias condiciones en las que se encontraba el lugar turístico en cuanto a higiene. Por otro lado, palabras como “rob”, “pag” indican que las personas que calificaron no estuvieron de acuerdo con los precios vs la calidad del servicio, o fueron directamente “robadas” en los sitios turísticos a los que fueron. Por último, palabras como “sabor” sugieren que la comida en aquellos lugares no fue de agrado para los reseñadores, y “decepción” sugiere que hay un descontento en aquellas personas que probablemente consideraron el sitio turístico como atractivo según lo que vieron en línea, pero que al ir físicamente no cumplió con sus expectativas.

Para la calificación de 3: existen adjetivos con connotación positiva como “bien”, “buen”, neutra como “regul”, “normal”, y negativa como “falt”, “men”, “embarg”. No hay sustantivos que den pistas sobre la higiene, la comida, el hospedaje, la atención, etc.

Para las calificaciones 4 y 5: el único adjetivo dentro de las 10 palabras más importantes que sugiere algo respecto a la comida es “delici”. “Hermos” sugiere que estos lugares son estéticamente agradables para los visitantes. Por su parte, palabras como “limpi”, “comod” sugieren higiene y comfort en aquellos lugares turísticos con calificaciones altas. El resto de las palabras como “gran”, “sup”, “perfect”, “increibl”, “excelent” son calificadores positivos, pero que no dan pistas de factores específicos como atención, precios, comida, hospedaje, etc.

Mapa de actores relacionado con el producto de datos creado

Rol dentro de la organización	Tipo de actor	Beneficio	Riesgo
Departamento de Análisis de Datos	Usuario-cliente	Puede utilizar el modelo para identificar oportunidades de mejora en los sitios turísticos	Si el modelo no es preciso, puede llevar a la toma de decisiones incorrectas
Dirección de Finanzas	Financiador	Puede utilizar el modelo para asignar de manera más eficiente los recursos a los sitios turísticos	Si el modelo no es preciso, puede llevar a la asignación ineficiente de recursos
Departamento de IT	Proveedor	Puede garantizar que el modelo cumple con los estándares de calidad y seguridad de datos	Si se manejan incorrectamente los datos, puede llevar a la violación de la privacidad
Turistas	Beneficiado	Pueden beneficiarse de las mejoras en los sitios turísticos basadas en el modelo	Si el modelo no es preciso, puede llevar a experiencias turísticas insatisfactorias

Trabajo en equipo

Santiago Ramírez

Los roles asociados son Líder de datos y Líder de negocio, esto porque fue responsable de velar por resolver el problema o la oportunidad identificada y estar alineado con la estrategia del negocio para el cual se plantea el proyecto. Además, se encargó de gestionar los datos que se usaron en el proyecto y de las asignaciones de tareas sobre datos.

El número de horas dedicadas al proyecto fueron 6 horas. El modelo asignado fue el modelo de regresión logística. Algunos de los retos encontrados al realizar el proyecto fueron el cómo tratar los datos de la mejor manera, el que hacer para poder identificar relaciones entre las diferentes palabras en las reseñas. La mejor estrategia encontrada para cumplir el objetivo fue el tener un muy buen tratamiento de los datos, así se podían identificar mucho más fácil las relaciones.

Mario Ruíz

El rol asociado fue el de líder de proyecto, esto porque estuvo a cargo de la gestión general del proyecto, definiendo las fechas de reuniones, pre-entregables del grupo y verificando las asignaciones de tareas para asegurar una distribución equitativa de la carga de trabajo. Además, se encargó de subir la entrega del grupo y tomó decisiones finales en caso de falta de consenso.

Se dedicaron 6 horas al proyecto, enfocándose en coordinar y gestionar el desarrollo de este en su totalidad.

Durante el proyecto, enfrentó el reto de garantizar la alineación y colaboración eficiente entre todas las partes involucradas, así como cumplir con los plazos establecidos y mantener la calidad del trabajo en equipo.

Santiago Gelvez

El rol asociado fue Líder de analítica, ya que fue responsable de interpretar los resultados del mejor modelo y la utilidad de los resultados para sugerir recomendaciones a los actores de turismo.

Se dedicaron 6 horas al proyecto. Que incluyen creación del modelo individual e interpretación de resultados del mejor modelo para proponer oportunidades de mejora.

En cuanto a los retos, el modelo SVC inicialmente tuvo problemas de ejecución y errores al obtener las palabras más significativas. Asimismo, la interpretación cualitativa de los resultados, ya que va más allá de identificar el f1-score más alto.

Puntos y cosas por mejorar

Como equipo decidimos repartir los puntos equitativamente, así cada integrante del grupo tendría 33.3 puntos, esto debido a que sentimos que todos trabajamos de la misma forma.

En cuanto a puntos a mejorar para las próximas entregas, queremos crear un tiempo establecido para ciertos entregables. Así facilitaríamos la unión y consolidación del proyecto. También queremos mejorar en la comunicación con los miembros del equipo, tanto en el equipo interno como en el interdisciplinario, dado que encontramos algunas fallas en la comunicación que podemos corregir para futuras entregas.

Repositorio:

El código, conjuntos de datos usados, modelos y predicciones se encuentran en el siguiente repositorio del proyecto:

https://github.com/ma-r-s/Proyecto_1-Inteligencia_de_Negocios-G12