Sentiment Analysis sobre reviews

A través de este trabajo se busca poder integrar nociones y conocimientos sobre NLP vistas en el módulo, así como en los previos para generar un modelo de machine learning.
El objetivo que van a tener es construir un clasificador el cual pueda predecir si una revisión realizada por un usuario es positiva o negativa (buena o mala).
Para ello, utilizaremos un conjunto de datos que pertenece a la plataforma Yelp. Esta, posee una red de usuarios, los cuales realizan opiniones sobre lugares nocturnos, espacios culturales, locales comerciales, entre otros.

El dataset a trabajar se encuentra en el siguiente link. Deberán realizar un análisis de features, así como su preparación necesaria antes de iniciar el desarrollo del modelo.

Objetivos

Deberán generar un modelo de machine learning el cual pueda clasificar review en inglés para la plataforma Yelp. Es decir, nuestro modelo recibirá una review de un usuario, y deberá ser capaz de determinar si esta es positiva o negativa.

Dataset

Las features que contiene este dataset son las siguientes:

- business_id: identificador del negocio al que se está realizando la review.
- cool: cantidad de votos por haber sido una review “cool”.
- date: fecha de realización de la revisión
- funny: cantidad de votos para una revisión “divertida”.
- review_id: identificador único de revisión (ofuscado).
- stars: cantidad de estrellas otorgadas por el usuario en referencia a la review.
- text: revisión realizada por el usuario sobre un determinado negocio.
- useful: cantidad de votos recibido por los usuarios a los cuales le resultó útil la revisión.
- user_id: id del usuario en la plataforma (ofuscado)

Cuento con los datos del business (negocio) sobre el que se realizó la review por si consideran que es necesario para sumar features, datos, etc.

Consideraciones

- No contamos con una variable target como pasa en problemas de la vida real. Por ello, un desafío extra que se presenta es cómo definir un target, basado en las features del dataset.
- Muchas veces cuando importamos un dataset pandas infiere que valor podría ser, de no encontrar un valor conocido pone uno por defecto.Validar que los tipos de datos de las features después de importarse correspondan con su valor intrínseco es una buena práctica.

- Haga una rápida exploración de valores atípicos (outliers) del conjunto de datos. Realice los gráficos que considere pertinente para entender la naturaleza del problema.
- Como aplicación opcional, sería interesante investigar, y evaluar si sobre nuestro problema sirve implementar Recursive Feature Elimination, y como y donde lo harían.

Evaluación

Para la evaluación de los modelos vamos a utilizar las siguientes métricas:

- Precision
- Recall
- F1-score
- Análisis de AUC ROC

Entrega

Se deberá entregar un notebook en el que se puede reproducir paso a paso todo lo necesario para la generación del modelo, así como al momento de realizar predicciones.
Todos los pasos donde se hayan tomado decisiones deben estar documentados, así como cualquier librería o link a una fuente de datos externa.
La fecha de entrega puede ser el mismo día de la presentación (ver apartado presentación).

En cuanto las correcciones irán sucediendo a lo largo del mes de Noviembre.

Presentación

Se realizará una presentación en no más de 30 minutos por grupo, donde explicarán su proceso de desarrollo, ideas, y decisiones tomadas para llegar a su modelo.
Esto tendrá lugar en la última clase el 19 de Noviembre. Se destinarán las últimas horas de la clase para la presentación.

Dependiendo de las tecnologías utilizadas, un extra para la presentación del trabajo sería poder implementar un pipeline de scikit-learn. Más información acá.