Ejercicio realizado por: Alberto Rodríguez Álvarez

Procesamiento y análisis de información textual

1. Elección y preparación del dataset (10%)

A partir de la selección del dataset y de las tareas realizadas en el apartado 1 de la práctica, titulada “Preparación del dataset”, contesta a las siguientes preguntas:

1. Describe el dataset escogido. Indica el origen del mismo, la cantidad de datos que contiene, la tipología de los mismos y justifica su elección.
2. Explica los problemas que has encontrado en los datos y las actividades de limpieza y preprocesamiento que has realizado en la etapa de “preparación del dataset”.

El dataset seleccionado es *"Rotten Tomatoes movies and critic reviews dataset"*, obtenido de Kaggle (Enlace). En su descripción vemos que este conjunto de datos fue extraído de la web pública Rotten Tomatoes hasta el 31 de octubre de 2020. Contiene información sobre películas, incluidas las reseñas de críticos y las calificaciones de usuarios y críticos.
El dataset original contiene 1.130.017 de registros y 8 variables. Tras la limpieza el resultado es un dataset con 485.392 registros y las variables 'reviewText', 'sentimiento' y 'text' descritas en el enunciado. Dado que este dataset fue generado mediante web scraping presenta diversas incongruencias y múltiples casos especiales. Se han intentado corregir los más relevantes como caracteres especiales (acentos especiales, repetición de signos, etc), emojis, múltiples idiomas, etc. La mayoría solucionado mediante *regex.* Otro factor a valorar es la diversidad de formatos para puntuar, que ha requerido analizar todos los casos errores manualmente.
La elección de este dataset se justifica por su pertinencia al objetivo de la práctica: analizar, modelar y clasificar opiniones. Además, su estructura permita crear la variable puntuación y así cumplir los requisitos.

2. Obtención de datos (30%)

A partir de las tareas realizadas en el apartado 2 de la práctica, titulada “Obtención de datos”, contesta a las siguientes preguntas:

1. Comenta y compara los n-gramas y las colocaciones obtenidos mediante los distintos métodos en este ejercicio.
2. Analiza los términos obtenidos utilizando el modelo Word2Vec, escoge los 5 términos (aspectos) más relevantes y comenta los criterios tomados en cuenta para la selección.

Se analizaron n-gramas (2 y 3) y colocaciones mediante **PMI** y **Likelihood Ratio**, destacando frases como *"great movie"* y *"bad script"* (PMI) o *"award winning"* y *"slow pacing"* (Likelihood Ratio). Mientras PMI detecta asociaciones raras, Likelihood Ratio prioriza términos frecuentes.
Por otro lado, el modelo **Phraser** (Gensim) preservó el contexto semántico al generar tokens como *"top_notch_acting"* o *"poor_visual_effects"*. Comparativamente, Phraser facilita tareas como clasificación al combinar términos en unidades significativas, mientras PMI y Likelihood Ratio identifican patrones de co-ocurrencia.

Tras aplicar Word2Vec, se generaron vectores que capturan relaciones semánticas entre términos. Se evaluaron términos mediante similitud de coseno con aspectos como "director", "actors" y "plot". Los cinco aspectos más relevantes son:

Director: vision, work, filmmaker. Representa la calidad percibida de las películas, reflejando estilo y dirección.
Actors: performance, cast, role. Es fundamental para la recepción de una película, destacando las actuaciones individuales.
Plot: story, narrative, twist. Es clave para evaluar si la trama resulta atractiva o predecible.
Visuals: effects, cinematography, scenes. Las críticas sobre efectos especiales y fotografía son frecuentes en géneros como ciencia ficción.
Soundtrack: music, score, composer. La banda sonora es central en dramas y thrillers, evaluándose su impacto en la experiencia.

Los aspectos fueron seleccionados por su frecuencia y relevancia, representando las principales dimensiones críticas en reseñas de películas.

3. Detección de temas (30%)

A partir de las tareas realizadas en el apartado 3 de la práctica, titulada “Detección de temas”, contesta a las siguientes preguntas:

1. Analiza los distintos hallazgos encontrados en la primera parte del ejercicio, la exploración de los temas con WordNet.
2. Compara los distintos modelos LDA utilizados, elige el más adecuado de forma justificada.

Al explorar términos relevantes como *director* y *actors* con WordNet, se observó que Word2Vec captura relaciones específicas del dominio, como *filmmaker* y *vision*, mientras que WordNet ofrece sinónimos más generales, como *manager*, que no siempre son útiles en contextos cinematográficos. Para términos como *plot*, ambos modelos coincidieron más estrechamente. Esto evidencia que Word2Vec es más efectivo para capturar la semántica contextual de dominios especializados.
Se compararon tres modelos LDA con 5, 8 y 10 tópicos, evaluados mediante coherencia y perplejidad:

LDA con 5 tópicos: Alta coherencia (0.5378), pero poca diferenciación temática.

LDA con 8 tópicos: Mayor especificidad temática, aunque con coherencia más baja (0.4690).

LDA con 10 tópicos: Mejor equilibrio (coherencia 0.5283), logrando tópicos bien definidos como *visuals* (e.g., *cinematography*, *effects*) y *performance*.

Modelo seleccionado: LDA con 10 tópicos por su balance entre granularidad y coherencia, lo que permite identificar temas claros y específicos del dominio sin comprometer interpretabilidad.

4. Clasificación automática de opiniones positivas y negativas (20%)

A partir de las tareas realizadas en el apartado 4 de la práctica, titulada “Crear un clasificador automático de opiniones positivas y negativas”, comentar los algoritmos utilizados, los resultados obtenidos y la coherencia de los resultados con el contenido de los comentarios.

En este apartado, se implementó un clasificador de opiniones utilizando **Logistic Regression** y **SVM**. Ambos clasificadores fueron entrenados con vectores tf-idf de las opiniones procesadas para identificar patrones que determinaran la polaridad positiva o negativa de las reseñas.
El clasificador basado en **Logistic Regression** destacó las palabras más representativas para ambas clases. Para las opiniones negativas, términos como *worst*, *bad*, *fails* y *dull* fueron las más informativas, reflejando críticas relacionadas con guiones aburridos o problemas en la trama.
Por otro lado, para las opiniones positivas, términos como *best*, *entertaining*, *great* y *performance* evidenciaron elogios hacia elementos destacados, como actuaciones memorables y calidad general de la película. Este clasificador alcanzó un F1-score promedio de 0.85, mostrando resultados consistentes con las características esperadas del contenido de las opiniones.
El modelo **SVM**, aunque más lento en entrenamiento, presentó un rendimiento similar, con un F1-score promedio de 0.83. Sin embargo, Logistic Regression fue preferido por su capacidad para identificar términos informativos y su eficiencia en el entrenamiento.
En general, los resultados obtenidos son coherentes con el contenido de las opiniones, ya que los términos más informativos coinciden con los aspectos positivos y negativos que los usuarios suelen destacar en sus comentarios. Esto demuestra que los modelos capturan adecuadamente la semántica y polaridad del texto.

5. Evaluación (10%)

A partir de las métricas calculadas en los clasificadores del apartado 5 de la práctica, titulada “Evaluación”, deberás comparar y evaluar los dos modelos propuestos en función de las métricas de evaluación vistas en clase (precision, recall y f1) y del tiempo de ejecución. Concluye finalmente cuál de los dos modelos elegirías para predecir nuevas reseñas sobre productos y el porqué.

En el apartado de evaluación, se compararon Logistic Regression y SVM utilizando las métricas de precisión, recall y F1-score, así como el tiempo de ejecución. Logistic Regression destacó por su eficiencia, completando el entrenamiento en segundos y alcanzando un F1-score promedio de 0.85, reflejando un buen equilibrio entre precisión y recall. Este modelo identificó con claridad palabras clave como best y worst, ofreciendo interpretaciones claras de los resultados. Por su parte, SVM mostró robustez en datasets complejos, pero tuvo un F1-score ligeramente inferior (0.83) y requirió más tiempo de entrenamiento.

El F1-score es crucial para medir el equilibrio entre esperados y encontrados en clases desbalanceadas, siendo útil para evaluar el rendimiento general del modelo.

Bajo mi criterio personal elegiría la Logistic Regression por su precisión, rapidez y capacidad interpretativa, lo cual permite predecir de manera eficiente en aplicaciones prácticas. Aunque ambos son algoritmos excelentes, dependerá de los datos su idoneidad.

Criterios de valoración

Los ejercicios tienen un peso de 10%, 30%, 30%, 20% y 10% respectivamente. Se valorará, para cada apartado, la validez de la solución y la claridad de la argumentación de acuerdo con la rúbrica facilitada en cada ejercicio. Muchos de los ejercicios planteados no tienen una única respuesta, por lo que es importante justificar la respuesta propuesta adecuadamente. Cualquier solución no justificada se considerará incompleta.