2. Proyecto del curso

Descripción del proyecto

Film Junky Union, una nueva comunidad vanguardista para amantes de las películas clásicas, está desarrollando un sistema para filtrar y categorizar reseñas de películas. El objetivo es entrenar un modelo para detectar las críticas negativas de forma automática. Utilizarás un conjunto de datos de reseñas de películas de IMDB con leyendas de polaridad para construir un modelo para clasificar las reseñas positivas y negativas. Este deberá alcanzar un valor F1 de al menos 0.85.

Instrucciones del proyecto

1. Carga los datos.
2. Preprocesa los datos, si es necesario.
3. Realiza un análisis exploratorio de datos y haz tu conclusión sobre el desequilibrio de clases.
4. Realiza el preprocesamiento de datos para el modelado.
5. Entrena al menos tres modelos diferentes para el conjunto de datos de entrenamiento.
6. Prueba los modelos para el conjunto de datos de prueba.
7. Escribe algunas reseñas y clasifícalas con todos los modelos.
8. Busca las diferencias entre los resultados de las pruebas de los modelos en los dos puntos anteriores. Intenta explicarlas.
9. Muestra tus hallazgos.

¡Importante! Para tu comodidad, la plantilla del proyecto ya contiene algunos fragmentos de código, así que puedes usarlos si lo deseas. Si deseas hacer borrón y cuenta nueva, simplemente elimina todos esos fragmentos de código. Aquí está la lista de fragmentos de código:

- Un poco de análisis exploratorio de datos con algunos gráficos;
- `evaluate_model()` : una rutina para evaluar un modelo de clasificación que se ajusta a la interfaz de predicción de scikit-learn;
- `BERT_text_to_embeddings()` : una ruta para convertir lista de textos en insertados con BERT.

Tu trabajo principal es construir y evaluar modelos.

Como puedes ver en la plantilla del proyecto, te sugerimos probar modelos de clasificación basados en regresión logística y potenciación del gradiente, pero puedes probar otros métodos. Puedes jugar con la estructura de la plantilla del proyecto siempre y cuando sigas sus instrucciones.

No tienes que usar BERT para el proyecto porque requiere mucha potencia computacional y será muy lento en la CPU para el conjunto de datos completo. Debido a esto, BERT generalmente debe ejecutarse en GPU para tener un rendimiento adecuado. Sin embargo, puedes intentar incluir BERT en el proyecto para una parte del conjunto de datos. Si deseas hacer esto, te sugerimos hacerlo de manera local y solo tomar un par de cientos de objetos por cada parte del conjunto de datos (entrenamiento/prueba) para evitar esperar demasiado tiempo. Asegúrate de indicar que estás usando BERT en la primera celda (el encabezado de tu proyecto).

Descripción de los datos

Los datos se almacenan en el archivo `imdb_reviews.tsv`. Download the dataset.

Los datos fueron proporcionados por Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, y Christopher Potts. (2011). ***Learning Word Vectors for Sentiment Analysis.*** *La Reunión Anual 49 de la Asociación de Lingüística Computacional (ACL 2011).*

Aquí se describen los campos seleccionados:

- `review`: el texto de la reseña;
- `pos`: el objetivo, '0' para negativo y '1' para positivo;
- `ds_part`: 'entrenamiento'/'prueba' para la parte de entrenamiento/prueba del conjunto de datos, respectivamente;

Hay otros campos en el conjunto de datos, puedes explorarlos si lo deseas.

Evaluación de proyecto

Hemos recopilado los criterios de evaluación para el proyecto. Léelos con atención antes de pasar al ejercicio:

- Cargaste y preprocesaste los datos de texto para su vectorización.
- Transformaste los datos de texto en vectores.
- Definiste, entrenaste y probaste los modelos.
- Se alcanzó el umbral de la métrica.
- Colocaste todas las celdas de código en el orden de su ejecución.
- Puedes ejecutar sin errores todas las celdas de código.
- Sacaste conclusiones.

Nuestros revisores también observarán la calidad general de tu proyecto:

- ¿Mantuviste la estructura del proyecto?
- ¿Mantuviste limpio tu código?
- ¿Conseguiste evitar la duplicación del código?
- ¿Cuáles fueron tus hallazgos?

Tienes tus hojas informativas y los resúmenes de los capítulos, así que ya puedes continuar con el proyecto.

¡Buena suerte!