Tarea

Usando Transfer Learning para predecir categorías en textos políticos

Fecha de entrega: Hasta el domingo 20 de agosto 10:00 pm

Formato de entrega: Un archivo comprimido con el nombre: "Nombre_Apellidos".zip, el archivo deberá contener un Notebook de Jupyter/Colab y/o archivos .py (si se desea). **Comentar los pasos que se van siguiendo**.

Utilizando como guía el Notebook del caso de aplicación de la sesión 1 (Ver Notebook) y tomando como referencia el flujo de solución de problemas en Natural Language Processing, elaborar **2 modelos predictivos** basados en Machine Learning que sean capaces de predecir una de las 7 categorías políticas del dataset "Manifesto Project" – utilizar solamente las columnas *phrase* y *label*.

Las arquitecturas de los modelos predictivos deberán contar con las siguientes características:

Arquitectura 1:

- Un vectorizador basado en Word Embeddings (Word2Vec o FastText) entrenado con el mismo corpus. (Embedding entrenado con el mismo vocabulario del dataset). **Nota: Especificar con que partición de datos están entrenando, e.g. "entrenamiento con el dataset completo (train+test)" o "entrenamiento con la partición de training".**
- Si fuera el caso, implementar una estrategia para solucionar el problema de OOV (palabras fuera de vocabulario).
- Un clasificador e.g. SVM, NaiveBayes, MLP.

Arquitectura 2:

- Un vectorizador preentrenado basado en Word Embeddings (Word2Vec o FastText). **Nota: Se recomienda usar los vistos en clase. Pueden tomar como referencia el Notebook de la sesión de Word Embeddings.**
- Si fuera el caso, implementar una estrategia para solucionar el problema de OOV (palabras fuera de vocabulario).
- Un clasificador e.g. SVM, NaiveBayes, MLP.

Evaluar la efectividad de predicción de los modelos en términos de Accuracy y F1-Score Macro y comparar los dos modelos.

Links del dataset: Manifesto(Cased), Manifesto(Uncased). También se encuentran colgados en el website del taller.

Notas adicionales:

Hay dos versiones del mismo dataset (Cased y Uncased):

- Cased: Las mayúsculas en el dataset se mantienen
- Uncased: Todo el texto del corpus está en minúsculas
- Utilizar el que crean conveniente de acuerdo al modelo preentrenado elegido.

Puede que surjan problemas de sobrecarga de memoria RAM al cargar los embeddings preentrenados mas realizar el proceso de vectorización de los datos, por lo que se recomienda reducir el dataset tal como se ha hecho en las sesiones pasadas e.g. utilizar solo los 10,000 primeros registros. **Especificar si se ha realizado esto.**