Práctica final NLP

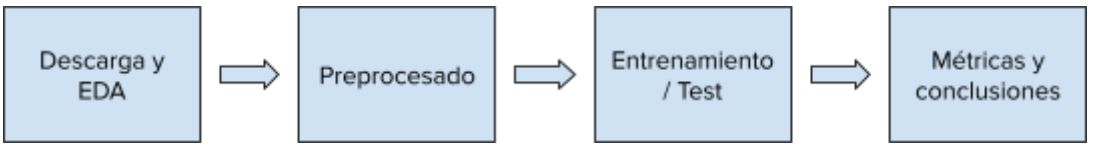
Introducción

El objetivo principal de la presente actividad es el de evaluar los conocimientos del alumno planteando algunos retos de NLP lo más prácticos y realistas posibles. De hecho, los ejercicios que se proponen a continuación son retos a los que se enfrentan muchas empresas y organizaciones hoy en día. Durante el desarrollo de la práctica el alumno tendrá ocasión de desarrollar las principales tareas y módulos necesarios en un proyecto de análisis de sentimiento.

No se busca que los modelos tengan un performance (precision, recall, f1-score, ...) **excelente, sino que los pasos que se propongan para resolver los ejercicios sean razonables** (acorde a lo visto en la asignatura) **y estén justificados. Es importante también comentar los resultados obtenidos.**

Ejercicios

Los siguientes ejercicios que se proponen cubren todas las etapas en un proyecto de NLP clásico. Cada ejercicio se corresponde con una etapa y todos ellos están relacionados.



Cada ejercicio será resuelto en un notebook. **El alumno deberá resolver todos los ejercicios**.

1. Descarga y exploración del corpus

El alumno **descargará el/los corpus** que desee (detalle sobre los datos a utilizar en el siguiente punto) y realizará un **análisis exploratorio** de los datos.

Este ejercicio deberá contener:

- Cardinalidad del vocabulario
- Distribución de reviews por número de estrellas
- Nº de reviews positivas y negativas
- N-grams más frecuentes
- Nubes de palabras
- Visualización en 2 dimensiones de algunos word embeddings calculados con Word2Vec (elegir 4-5 palabras y pintar las top 10 más similares)
- Conclusiones de la exploración
- Cualquier otra métrica / exploración / cálculo que el alumno considere

2. Etapa de preprocesado de texto

El alumno preparará una **etapa de preprocesado de reviews** que permita adecuar el formato de las mismas a uno más adecuado. Será la etapa previa al entrenamiento del modelo de sentimiento.

Todo el preprocesado deberá incluirse en una función de Python que contenga todo el procesado de texto. Esta función puede (es recomendable) contener otras funciones que realicen tareas más concretas (eliminar stopwords, eliminar signos de puntuación, etc.).

3. Etapa de entrenamiento y testeo de un modelo de análisis de sentimiento

El alumno, con los datos preprocesados del ejercicio 2, **deberá entrenar dos modelos distintos** de los que, tras comparar sus resultados, elegirá uno como el mejor. Para tomar esta decisión se basará en las métricas que calcule (precision, recall, f1-score, ...). **El enfoque será el de un problema de clasificación binaria supervisada**.

Los modelos deberán tomar a su entrada los datos codificados con un modelo de bolsa de palabras (**bag-of-words)**. Se deberán justificar los parámetros del *vectorizer*, así como tener en cuenta aspectos como el balanceo de clases.

La elección de los modelos es libre.

4. Reporte de métricas y conclusiones

El alumno, tomando como referencia los resultados del modelo escogido en el ejercicio 3, **calculará las métricas** que permitan validar la bondad del modelo. También incluirá **comentarios** y las **conclusiones finales**.

Datasets

Reviews de Amazon: Reviews de productos de Amazon clasificadas por categorías (libros, electrónica, automoción, ...). Las reviews contienen la calificación (número de estrellas) por lo que es perfecto para problemas de clasificación supervisada del sentimiento.

El alumno deberá escoger el dataset que desee (puede trabajar con varios si así lo prefiere) **de los que aparecen en la columna "5-core"**. Se recomienda trabajar con subsets de dichos datasets para evitar problemas de memoria.

Enlace: http://jmcauley.ucsd.edu/data/amazon/

Formato de entrega

La práctica deberá resolverse en **Python** con las librerías descritas durante la asignatura en **un notebook de Jupyter por para cada dataset** . En cada notebook se incluirá **todo el código** necesario y los **comentarios**.