Laboratorio: Word Embedings y Transformers para clasificación de texto

**Objetivos**

Con este laboratorio el alumno comparará diferentes modelos de clasificación de texto, todos mediante el uso de técnicas basadas en Word Embedings y Transformers. El alumno por tanto adquirirá dos competencias: primero, la capacidad de aplicar un modelo neuronal para la clasificación de texto. Segundo, la capacidad de comparar diferentes modelos entre sí.

**Descripción**

En esta actividad vamos a trabajar en clasificar textos. Se recorrerá todo el proceso desde traer el dataset hasta proceder a dicha clasificación. Durante la actividad se llevarán a cabo muchos procesos como la creación de un vocabulario, el uso de embeddings y la creación de modelos.

Las cuestiones presentes en esta actividad están basadas en un Notebook creado por François Chollet, uno de los creadores de Keras y autor del libro "Deep Learning with Python".

**Nota**: Esta actividad está generada en coordinación con la asignatura “Sistemas Cognitivos”. En concreto, en dicha asignatura se te propondrá realizar un ejercicio similar, pero con una aproximación tecnológica diferente.

Para realizar la actividad, se partirá del fichero “mia7\_lab.ipynb”. Dicho fichero contiene el código fuente en Python necesario para realizar las siguientes tareas:

1. Importar las librerías necesarias
2. Descargar los datos del dataset
3. Mezclar y separar los datos en catálogos de entrenamiento y validación
4. Tokenización del texto
5. (…)
6. Evaluación del clasificador

En este laboratorio deberás completar diferentes versiones del apartado que falta (punto 5), que es la creación y entrenamiento del modelo clasificador.

Para ello, **en la semana del laboratorio se publicarán nuevas versiones del Notebook** que incluirán diferentes variantes del modelo a implementar, además de material de ayuda.

Criterios de evaluación

* Debe resolver la actividad en los Jupyter Notebook proporcionados (habrá uno por cada variante del ejercicio). Si no se usan los archivo proporcionados la actividad será calificada con cero puntos.
* Para la evaluación de la actividad se deben entregar, para cada variante del ejercicio, el **Jupyter Notebook** que contenga:
  + El código en Python que permita resolver la actividad, debidamente comentado.
  + Las respuestas a las preguntas planteadas en el propio Notebook.
* Se valorarán positivamente, los comentarios del código claros y oportunos, que permitan entender las decisiones de implementación realizadas. Si el código no está comentado de forma suficiente, se penalizará en la calificación de la actividad.
* Se valorará positivamente, la respuesta clara, breve y bien argumentada a las preguntas.
* Si se detecta **copia** entre alumnos en el código, en los comentarios o en las respuestas, todos los alumnos involucrados obtendrán una calificación para la actividad de cero puntos.

**Extensión máxima:** no hay restricciones en la extensión.

|  |  |  |  |
| --- | --- | --- | --- |
| Laboratorio: Word Embedings y Transformers para clasificación de texto | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Código Python | El código es correcto, está bien estructurado y debidamente comentado con comentarios claros y oportunos. | 5 | 50% |
| Respuesta a las preguntas | Las preguntas están debidamente respondidas de forma clara, breve y bien argumentada | 5 | 50% |
|  |  | **10** | **100 %** |