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

Además de dicho fichero, dispones de dos ficheros adicionales como son:

* “mia7\_lab\_red\_clasica.ipynb”.
* “mia7\_lab\_red\_transformers.ipynb”.

Estos ficheros completan el apartado (5) con una aproximación basada en redes clásicas y otra en transformers, respectivamente. En este laboratorio deberás cargar y ejecutar paso a paso las diferentes versiones del código.

Una vez ejecutado el código, responde a las siguientes preguntas:

1. Utilizando el tokenizador de spacy, que ya conoces, calcula el número promedio de tokens de una muestra de 15 ficheros de la categoría ‘com.graphics’. Indica el código utilizado y el resultado obtenido. **(1 punto)**
2. El código proporcionado lee los ficheros uno a uno y, antes de generar el catálogo de datos de entrenamiento y validación, descarta las 10 primeras líneas de cada fichero. ¿Cuál es el trozo de código en el que se realiza dicho descarte?, ¿por qué crees que se descartan dichas líneas?, ¿por qué 10 y no otro número? **(1 punto)**
3. ¿Qué se controla con el parámetro 'validation\_split'?, ¿por qué se ha elegido ese valor?, ¿qué ocurre si lo modificas? **(1 punto)**
4. Imprime por pantalla un ejemplo (es decir, un elemento del array) de ‘train\_samples’, ‘val\_samples’, ‘train\_labels’ y ‘val\_labels’. A tenor de las etiquetas que se utilizan, ¿qué tarea crees que se está intentando entrenar? **(1 punto)**
5. Con 'output\_sequence\_length' se establece un tamaño fijo para la salida de Vectorizer. ¿Por qué se necesita un tamaño fijo, y por qué se ha elegido el valor ‘200’? **(1 punto)**
6. Indica cuál es la precisión del modelo en el conjunto de datos de entrenamiento y en el conjunto de datos de validación. ¿Qué interpretación puedes dar? Haz en este punto un análisis comparativo de los dos modelos ejecutados. **(1.5 puntos)**
7. En la parte final del código se hace un análisis cualitativo de la salida. Explica el funcionamiento de este análisis e interpreta los resultados. Haz también en este punto un análisis comparativo de los dos modelos ejecutados. **(1 punto)**
8. Explica algunas de las limitaciones que puedes encontrar al modelo entrenado. **(1.5 puntos)**
9. ¿Qué sería necesario para que este modelo pueda interpretar textos en español? **(1 punto)**

Entrega y evaluación

* Debe resolver la actividad en los Jupyter Notebook proporcionados (habrá uno por cada variante del ejercicio).
* Debe entregar un informe con la respuesta a las preguntas planteadas en el presente documento.
* 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  % |
| Pregunta 1 | La respuesta es válida y está bien argumentada. | 1 | 10% |
| Pregunta 2 | La respuesta es válida y está bien argumentada. | 1 | 10% |
| Pregunta 3 | La respuesta es válida y está bien argumentada. | 1 | 10% |
| Pregunta 4 | La respuesta es válida y está bien argumentada. | 1 | 10% |
| Pregunta 5 | La respuesta es válida y está bien argumentada. | 1 | 10% |
| Pregunta 6 | La respuesta es válida y está bien argumentada. | 1.5 | 15% |
| Pregunta 7 | La respuesta es válida y está bien argumentada. | 1 | 10% |
| Pregunta 8 | La respuesta es válida y está bien argumentada. | 1.5 | 15% |
| Pregunta 9 | La respuesta es válida y está bien argumentada. | 1 | 10% |
|  |  | **10** | **100 %** |