Trabajo: caracterización de textos

Objetivos

Con esta actividad, se tratará de que el alumno se familiarice con el uso de la librería Spacy para el procesamiento de lenguaje natural y que la utilice para la tarea concreta de caracterización de textos bajo un corpus de clasificación de datos aportado en el enunciado.

**Corpus de datos**

Se trabaja con el corpus de detección de odio, que está debidamente explicado en la descripción del proyecto transversal. Los campos relevantes en este ejercicio son «CONTENIDO A ANALIZAR» e «INTENSIDAD». El primero contiene el mensaje tal cual. El segundo indica cuál es la intensidad de odio recogida en el mensaje. A efectos de este ejercicio, el campo intensidad se debe interpretar como un cambio binario, atendiendo al siguiente criterio:

* Si INTENSIDAD == 0 -> El mensaje no contiene odio.
* Si INTENSIDAD > 0 -> el mensaje sí contiene odio.

Para descargar el corpus, sigue las instrucciones de la descripción del proyecto transversal.

Descripción

En esta actividad debes procesar, de forma automática, un texto en lenguaje natural para detectar características básicas en el mismo, y para identificar y etiquetar las ocurrencias de conceptos como localización, moneda, empresas, etc.

En la primera parte del ejercicio, se proporciona un código fuente a través del cual se lee el archivo con el *dataset* y se almacena en listas de Python. En esta parte, el alumno tan solo debe ejecutar y entender el código proporcionado.

En la segunda parte del ejercicio, se plantea una serie de preguntas que debe ser respondida por el alumno. Cada pregunta deberá responderse con un fragmento de código fuente que esté acompañado de la explicación correspondiente. Para elaborar el código solicitado, el alumno deberá visitar la documentación de la librería Spacy, cuyos enlaces se proporcionarán donde corresponda.

Para realizar la actividad, debes utilizar la herramienta Jupyter para abrir el Notebook (caracteristicasOdio.ipynb) que acompaña a este enunciado.

Preguntas que responder

Para responder cada una de las preguntas planteadas, deberás aportar tanto el código fuente con el cual puedes conseguir la respuesta como una explicación válida de la respuesta y de la forma de obtenerla. Las preguntas por responder están indicadas en el propio Notebook (caracteristicasOdio.ipynb).

Criterios de evaluación

* Debe resolver la actividad en el Jupyter Notebook proporcionado junto con este enunciado. Si no se usa el archivo proporcionado, la actividad será calificada con cero puntos.
* Para la evaluación de la actividad se debe entregar el Jupyter Notebook que contenga:
  + El código en Python que permita resolver cada una de las preguntas.
  + La respuesta a cada pregunta, así como una explicación de la forma en la que esta se ha obtenido.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Trabajo: caracterización de textos | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Código fuente | El código empleado es correcto y devuelve información que responde a la pregunta planteada. La puntuación está equidistribuída para cada pregunta. | 5 | 50% |
| Explicación | La explicación aportada explica debidamente el código fuente aportado y, también, aporta una respuesta a la pregunta planteada. La puntuación está equidistribuída para cada pregunta. | 5 | 50% |
|  |  | **10** | **100 %** |