Trabajo: Named Entity Recognition

Objetivos

Con esta actividad se tratará de que el alumno se familiarice con el uso de la librería Spacy para el procesamiento de lenguaje natura, y que lo utilice para la tarea concreta de NER, bajo un texto dado.

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 un archivo de texto y se realiza un preprocesado del mismo. En esta parte el alumno tan sólo debe ejecutar y entender el código proporcionado.

En la segunda parte del ejercicio se plantean una serie de preguntas que deben ser respondidas 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 (NER.ipynb) que acompaña a este enunciado.

Preguntas que responder

Para responder a 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 a responder son:

* Pregunta 1. ¿Cuántas palabras tiene el texto?
* Pregunta 2. ¿Cuántas oraciones tiene el texto?
* Pregunta 3. ¿Cuál el número de palabras de la oración más grande? ¿Cuál es dicha oración?
* Pregunta 4. ¿Cómo puedes acceder al lema, lexema y morfemas de cada token?
* Pregunta 5. ¿Cómo puedes identificar/eliminar las stop words?
* Pregunta 6. ¿Qué atributo del token contiene la etiqueta NER?
* Pregunta 7. ¿Qué entidades soporta Spacy?, ¿Qué significa cada una?
* Pregunta 8. ¿Qué entidades diferentes son reconocidas en el texto?, ¿cuántas hay de cada tipo?
* Pregunta 9. Explica con tus palabras qué es el código IOB para el reconocimiento de entidades. Pon un ejemplo, sacado del texto, de una etiqueta de un único token y una etiqueta compuesta por varios tokens.
* Pregunta 10. ¿Cuántas palabras tiene el texto?

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 ésta se ha obtenido

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

|  |  |  |  |
| --- | --- | --- | --- |
| Ítem | 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 %** |