Extracción de entidades nombradas

Índice

1. Use Cases

La extracción de entidades de textos puede ser el primer paso para análisis posteriores. A continuación, se presentan varias de las aplicaciones en la industria real:

- **Agencias de noticias y publicidad**: se hace necesario explotar al máximo el elevado volumen de contenido online que se genera diariamente. El reconocimiento de entidades puede extraer las organizaciones, personajes, y lugares que se discuten en los artículos, y categorizarlos jerárquicamente mediante sistemas de etiquetas, de forma que la navegación sea lo más amigable posible.
- **Algoritmos de búsqueda eficientes**: la extracción de entidades en forma de etiquetas de entre una serie de documentos digitales, puede ayudar a los algoritmos de búsqueda que hayan sido diseñados para encontrar información en estos documentos. De esta forma, se pueden asociar determinados términos de búsqueda con etiquetas, y así retornar información que esté relacionada.
- **Recomendación de contenidos**: los sistemas de recomendación de contenidos

actuales pueden suponer el éxito masivo de una plataforma, como pueda ser el caso de Netflix. Mediante la extracción de entidades es posible (para el caso de publicadores de artículos, por ejemplo), recomendar a los usuarios artículos relacionados con los que han visitado, mediante la comparación de entidades entre textos.

- **Soporte a usuarios**: la extracción de entidades ayuda a extraer partes clave en un texto enviado por un cliente, sobre un determinado problema o sugerencia, y el análisis de esta extracción puede categorizar y encauzar la petición hacia el servicio correcto, de forma eficiente.
- **Tareas de investigación**: muchos tipos de investigaciones implican el procesado de un gran volumen de documentación, y la extracción de entidades y su categorización puede ayudar a la búsqueda y organización de los datos más relevantes, como sucedía de forma similar en casos anteriores.

2. ENUNCIADO

En esta práctica vamos a implementar una aplicación en Flask mediante la cual introduciremos un texto y podremos extraer las entidades nombradas de dicho texto.

Para ello, se proporciona el archivo index.html adjunto a este enunciado, que se utilizará como template para interactuar con la lógica de la aplicación:

Se pide:

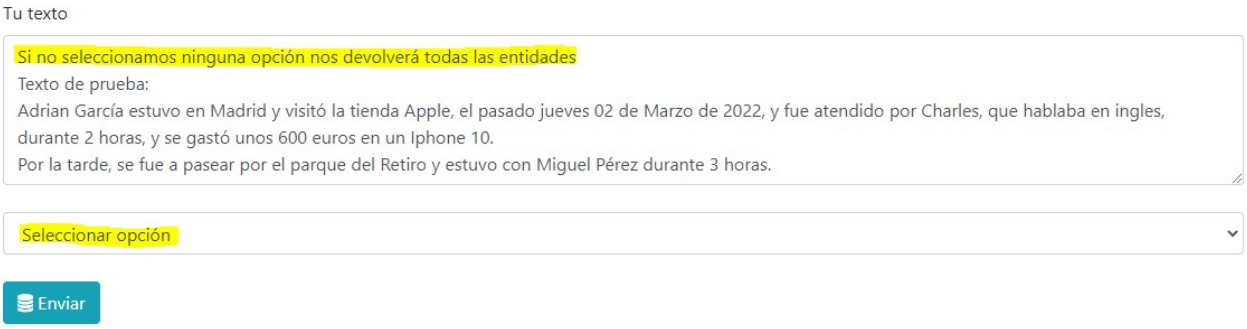
1. Realiza la versión básica de la aplicación en Flask, de forma que se introduzca el texto en el campo "Tu texto", y al pulsar sobre el botón "Enviar" se han de mostrar los resultados de las entidades, categorizados (**5 puntos**).

2. La lógica actual (en el HTML) solo contempla el reconocimiento de dos tipos de entidades. Añade al código (tanto al HTML como al código Python) lo necesario para que admita 3 entidades más de tu elección (**1 punto**).

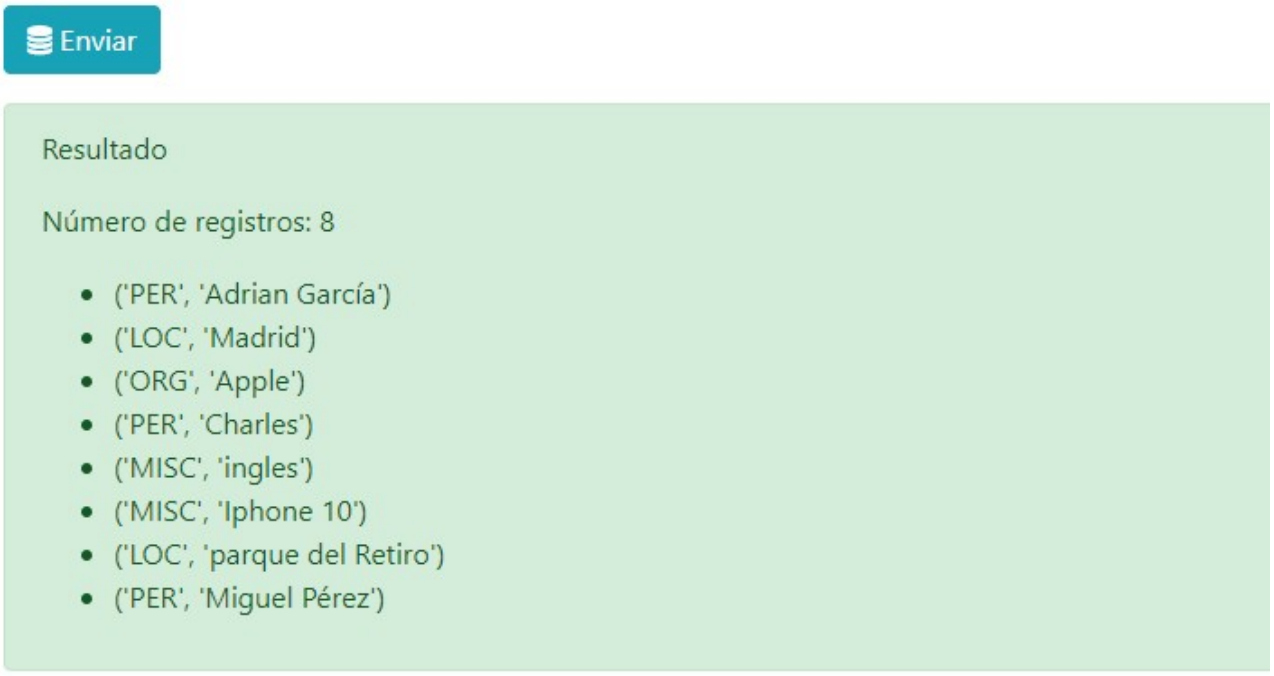
3. La aplicación ha de detectar el idioma en el que está escrito el texto, para cargar el modelo correspondiente. Prepáralo para que detecte inglés y castellano. Aquí tienes un enlace que te puede ayudar (**2 puntos**).

4. Demuestra con capturas de pantalla y explicaciones que el sistema está funcionando según los requisitos. Sube el código de la aplicación a Github y proporciona el enlace de tu repositorio (**2 puntos**).

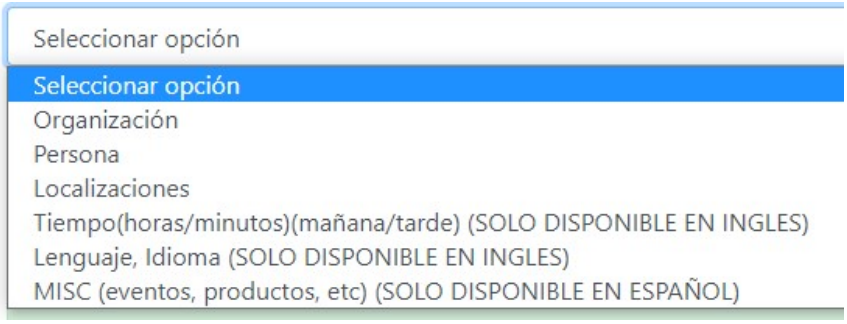
Demuestra con capturas de pantalla y explicaciones:



Al pulsar Enviar nos devuelve el resultado:

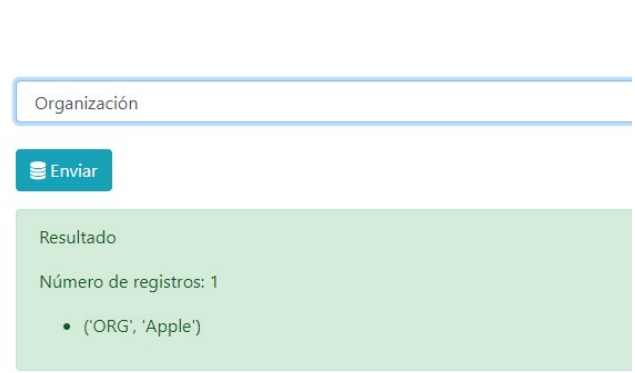


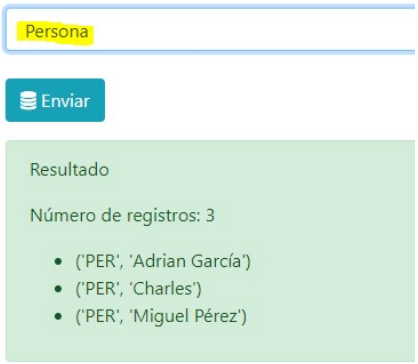
Estas son las opciones disponibles:

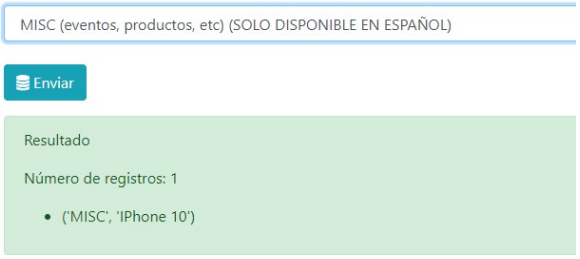


Con el mismo texto realizamos varias pruebas:

Texto de prueba:
Adrián García estuvo en Madrid y visitó la tienda Apple, el pasado jueves 02 de marzo de 2022, y fue atendido por Charles, que hablaba en inglés, durante 2 horas, y se gastó unos 600 euros en un IPhone 10.
Por la tarde, se fue a pasear por el parque del Retiro y estuvo con Miguel Pérez durante 3 horas.





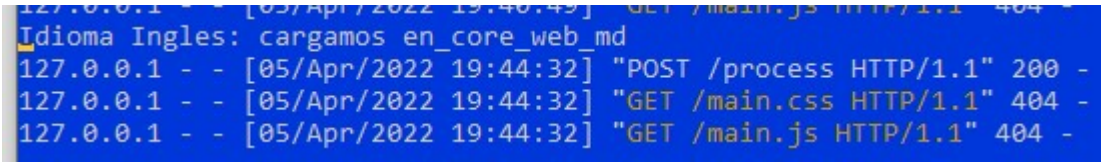


Vemos que esta detectando el idioma correctamente y nos carga español:

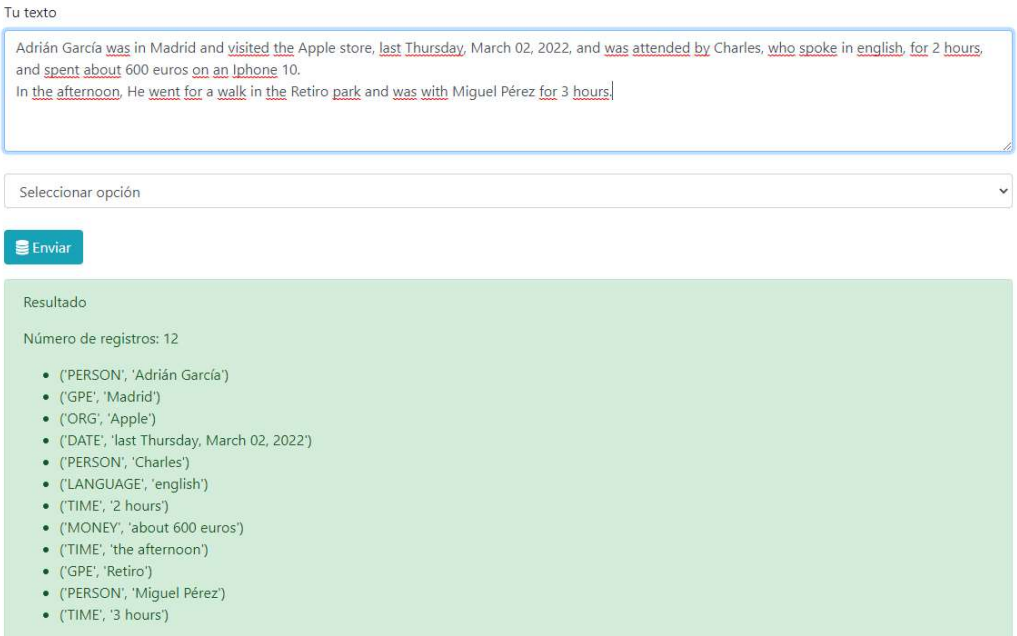
Ahora realizamos otra prueba con el mismo texto en inglés:

Adrián García was in Madrid and visited the Apple store, last Thursday, March 02, 2022, and was attended by Charles, who spoke in english, for 2 hours, and spent about 600 euros on an Iphone 10. In the afternoon, He went for a walk in the Retiro park and was with Miguel Pérez for 3 hours.

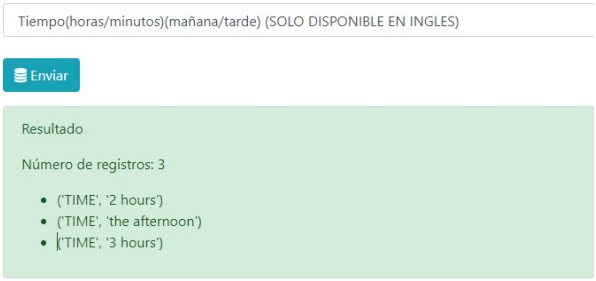
Detecta el idioma y lo carga.

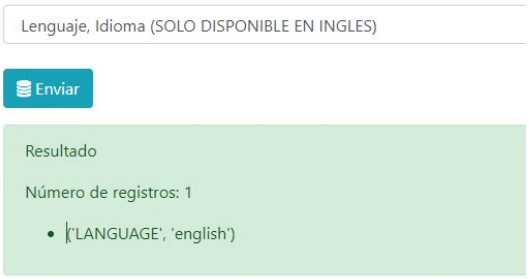


Comprobamos que en inglés detecta muchas más entidades:



Probamos los resultados con las 2 entidades que solo están disponibles en Ingles:





Para la deteccion de idioma hemos utilizado la librería:
from langdetect import detect

Y simplemente comprobamos el texto entrante, antes de extraer las entidades:

```
idioma = detect(rawtext)
    if idioma=='es':
        nlp = spacy.load("es_core_news_lg")
        print("Idioma Español: cargamos es_core_news_lg")
    elif idioma=='en':
        nlp = spacy.load("en_core_web_md")
        print("Idioma Ingles: cargamos en_core_web_md")
    else:
        print("No reconozco el idioma, se queda cargado Español: es_core_news_sm, por defecto ")
```

El resto del codigo podemos obtnerlo en GitHub(directorio code) .
Enlace Repositorio GitHub:
https://github.com/Zurichk/PracticaIA-NLP-spaCy.git