Universidad Rey Juan Carlos

TEMA 2 – HOJA DE EJERCICIOS IV

Se proponen diferentes ejercicios, relacionados con alguna fase de preprocesamiento de textos, o con tareas concretas de PLN utilizando exclusivamente expresiones regulares en Python.

1. Se pide hacer tokenización utilizando expresiones regulares. Dado un texto, hay que obtener todos los tokens dividiendo el texto separando por espacios en blanco.

   Texto de entrada= "*La AEMET lanza una alerta por nieve y una nueva ola de frío, será mejor que nos preparemos para lo peor en cuestión de horas. Habrá llegado el momento de empezar a pensar en lo que está por llegar, por lo que, habrá llegado el momento de prepararnos para los últimos coletazos del invierno y hacerlo de tal manera que tendremos que afrontar una recta final de la semana cargada de acción. Tenemos que empezar a pensar en lo que llegue a toda velocidad.*"

2. Se pide normalizar un texto a partir del uso de expresiones regulares. Dado el siguiente texto, hay que normalizarlo haciendo varias tareas:

   Texto de entrada = "*Francia aspira a jugar un papel protagonista en el auge de los algoritmos. Esta semana, París fue el epicentro de una cumbre mundial sobre IA, donde expertos de todo el mundo se adentraron en las amenazas y promesas de esta tecnología. En el marco de este evento, el presidente Emmanuel Macron y el jeque Mohamed bin Zayed Al Nahyan, líder de los Emiratos Árabes Unidos, presenciaron la firma de un acuerdo de cooperación entre sus países, un pacto que promete potenciar el desarrollo de proyectos conjuntos.*

   Como recoge la Agencia de Noticias de los Emiratos, la alianza incluye una inversión por parte de la nación rica en petróleo en Francia, así como "la adquisición de chips de vanguardia, la infraestructura de centros de datos y el desarrollo de talento, y mediante el establecimiento de Embajadas de Datos Virtuales para permitir la IA soberana y la infraestructura en la nube en ambos países". El Gobierno francés, por su parte, ha señalado que la iniciativa contempla la construcción de un enorme centro de datos."

   a. Eliminación de los signos de puntuación.

   b. Eliminación de las palabras vacías (stopwords) del texto. Se puede obtener la lista de palabras vacías de español con NLTK.

   c. Eliminación de todas las palabras con más de 5 caracteres.

3. Dado el texto "*Este es un texto que tiene ejemplos de fechas. Hoy es 09/02/2025, esta es una fecha posterior al 13 de enero de 2025. ¿Nos gustaría estar ya a 5 de julio y empezar las vacaciones? Casi que mejor no, que el tiempo avance a su ritmo. El primer día de clase fue el 30-01-2025, y el último día será el 8 de mayo del 2025.*" Hay que construir expresiones regulares para que sean capaces de reconocer los diferentes formatos de fechas que aparecen (DD/MM/AAAA, DD-MM-AAAA, DD de MM de AAAA y DD de MM).

4. Dado un texto que contenga números de teléfono de España, fijos (puede comenzar el prefijo por 8 o 9) o móviles (pueden comenzar con 6 o 7) y formados por 9 dígitos en total.

5. Se tiene que diseñar una expresión regular que haga NER para nombres de personas, es decir, que dado un texto sea capaz de reconocer nombres de personas en él.

   Texto de entrada= "*Un premio cantado, el de Eduard Fernández (mejor actor por Marco), y otro que fue una sorpresa, el de Carolina Yuste por La infiltrada, dejaron la puerta abierta a que pudiese suceder cualquier cosa en la recta final de la noche. Incluso lo mejor que podía pasar sucedió. Subió a recoger el Goya a la mejor dirección Pol Rodríguez, uno de los responsables junto a Isaki Lacuesta de la película mejor dirigida del año: Segundo premio.*"

   ¿Has podido reconocer todos los nombres de personas que aparecían en el texto? ¿Has reconocido cosas que no son nombres de personas?

6. Mediante expresiones regulares, reconocer los acrónimos presentes en un texto. Pueden ser acrónimos como URJC o acrónimos como U.S.A.

Después de solucionar los ejercicios anteriores, habrás podido comprobar que las expresiones regulares representan una forma fácil de procesar textos, permitiendo filtrar texto, sustituirlo, encontrar información, etc. Sin embargo, presentan desventajas claras en algunas tareas, por la dificultad para realizarlas correctamente. Es el caso del Reconocimiento de Entidades Nombradas (NER, *Named Entity Recognition*). Aunque esta tarea la veremos más adelante en detalle, ya se ha podido ver cómo utilizar expresiones regulares puede presentar ventajas y desventajas (ver Tabla 1).

Ventajas	**Desventajas**
Las ER son fáciles de aplicar y se pueden obtener resultados rápidamente.	Con ER sólo se puede tener en cuenta información contextual limitada y se suelen cometer errores al identificar patrones de texto (pero no lo que se buscaba).
Las ER se pueden utilizar para procesar grandes cantidades de datos.	Las ER requieren ajuste y supervisión manual al identificar patrones de texto erróneos.
Las ER también puede utilizarse en datos no estructurados que tengan texto.	Con ER no se pueden identificar patrones de texto complejos y su capacidad para comprender las relaciones semánticas entre las palabras es limitada.
Utilizar ER para NER suele ser más rápido y eficaz que un sistema NER cuando se trata de localizar patrones de texto sencillos.	Con ER no se pueden reconocer sinónimos o variaciones de patrones de texto.
Las ER se pueden utilizarse en sistemas o entornos en los que no se pueda aplicar un NER.	Las ER requieren un profundo conocimiento del procesamiento de textos y puede resultar difícil de aplicar para los no expertos.