

---

TEXT MINING & IMAGE RECOGNITION  
HOJA DE TRABAJO # 3

---

Instrucciones: A continuación verá una lista de ejercicios que debe completar para poder entregar el la hoja de trabajo #3. Al finalizar, todos sus archivos deben estar contenidos en un archivo ht3-sucarnet.zip. Este archivo lo debe entregar en el link del GES. Por favor cree una carpeta para cada ejercicio que usted realice.

**Problema 1:**

Realizar un breve resumen sobre las partes de una oración, utilizar el siguiente **link** como referencia.

**Problema 2:**

Realizar un resumen de 150 palabras sobre ¿Qué es NLP (Procesamiento de Lenguaje Natural)?

**Problema 3:**

Utilice expresiones regulares para validar las siguientes situaciones:

1. Implemente una regex para validar un correo electrónico en general, a continuación se muestran algunos ejemplos.
  - Guate.360-porelmundo@migate.com
  - Miercoles3@hotmail.com
  - Progra3.galileo@galileo.edu
2. implemente una regex para validar la dirección url de una página web con los tipos de domino (.com, .org, .edu). Note que la url incluye el protocolo (http o https) y los símbolos (//www.), a continuación se muestran algunos ejemplos:
  - https://www.guate360-porelmundo.com
  - http://www.a2.net
  - https://www.galileo.edu
  - http://www.8.org (No valida)
3. Implemente una regex para validar una MAC Address, notar que las mac addres están divididas en 6 bloques de caracteres hexadecimales, es decir que los símbolos solo pueden variar del 0 al 9 y las letras de la A a la F. a continuación se muestran algunos ejemplos:
  - 5A 6F AF 8C 9B 1D
  - 6D 6C 4D 3A EB 3F
  - 3A 7C FA C8 6D 4J (no valida por que el ultimo bloque contiene una J)

4. Implemente una regex para validar una dirección IPv4, notar que las direcciones IPv4 están divididas en 4 bloques de valores los cuales solo pueden ir desde 0 hasta 255, una ip donde algunos de sus bloques sea mayor a 255 no es valida, además tome en cuenta que cada bloque está separada por un punto. A continuación se muestran algunos ejemplos:
  - 192.16.8.1
  - 234.56.78.90
  - 1.2.3.4
  - 192.168.45.345 (no valida por que el ultimo bloque es mayor a 255)
5. Implemente una regex para validar una fecha con la secuencia día-mes-año donde el día, mes y año puedan estar separados ya sea por el caracter / o el caracter - o el caracter ., notar que las fechas son validas si los días están definidos desde el 1 al 31, el mes del 1 al 12 y el año de 2000 al 2019. También debe tomar en cuenta que los días y meses pueden estar escritos ya sea con uno o dos caracteres por ejemplo: Enero puede escribirse como 1 o como 01. Los años también pueden expresarse ya sea con dos o con cuatro caracteres por ejemplo: 19 o 2019 son validos. A continuación se muestran algunos ejemplos:
  - 20/1/2019
  - 12.03.2005
  - 31-11-08
  - 1-1-2012
  - 12-12-22 (no valida, por que el año supera al 2019).

Para su entrega solo es necesario que muestre los string que conforman la regex, para validar y armar sus regex puede utilizar el siguiente **link**.

#### **Problema 4:**

En la carpeta encontrará adjuntos 21 documentos que tiene 100 fechas en la secuencia días-mes-año pero con distinto separador y distinto formato de mes, en algunos casos aparece un numero y en otros el nombre del mes en ingles, por ejemplo: Enero puede aparecer como 1 o como Jan.

Utilice Python y expresiones regulares para encontrar el día, mes y año promedio total de los 21 archivos, los resultados deben ser un double.