

Laboratorio 9

Programación 1

Primer semestre 2017

1. Escribir una función que sanitiza una cadena. El método de sanitización es el siguiente:
 - a. Todas las letras con tilde se cambian por la misma letra sin tilde.
 - b. Símbolos de pregunta y de exclamación se reemplazan por guiones (" ")
 - c. Los espacios se reemplazan por guiones bajos (" ")
 - d. La "ñ" se reemplaza por la "n"
 - e. Cualquier otro símbolo (que no sean letras) se remueve.

2. Escribir una función que quita los elementos duplicados de una lista utilizando un diccionario. Debe retornar una nueva lista sin elementos repetidos.
3. Escribir una función que recibe el nombre de una canción y retorna el/los artista que la interpreta(n) (utilizando diccionarios). Por ejemplo:

`get_artistas('All my loving') -> 'The Beatles'`

`get_artistas('All along the watchtower') -> ['Bob Dylan', 'Jimi Hendrix']`

4. Crear una función que permita determinar probabilísticamente resultados que se pueden obtener en una tirada de un dado. Para eso se deben generar 10.000 tiradas aleatorias y obtener el porcentaje que resulta para cada número posible en el dado
5. Definir una función que abra un archivo (cuyo nombre es dado como parámetro) y cuente las palabras en el archivo que tienen "la misma letra inicial". La función debe imprimir el resultado en forma de una tabla ordenada por letra y luego ordenada por frecuencia en la siguiente manera:

Letra inicial	Frecuencia
a	78346
b	1289
c	821
.	
.	
.	

Desafío

- 1) Escriba un programa que toma un archivo como parámetro y cuenta las ocurrencias de todas las letras ('a' hasta 'z') sin distinguir entre mayúsculas y minúsculas y sin tener en cuenta acentos y otros símbolos. Calcular el porcentaje correspondiente a cada letra.

En el idioma español, la letra con mayor frecuencia es la letra "e" seguida por las "a", "o", "n", "i", "d", "l", "t" (*)

En el idioma inglés, la letra con mayor frecuencia es la letra "e" seguida por la "t", "a", "o", "i", "n", "s", "h", "r", "d" (*)

(*) Fuente: https://en.wikipedia.org/wiki/Letter_frequency

Diseñar un algoritmo que permita detectar el idioma de un texto arbitrariamente largo.

Es posible descargar libros en distintos idiomas completos en forma gratuita desde el siguiente sitio:

<http://www.gutenberg.org/>

- 2) 2) Un log del servidor de web Apache típicamente contiene entradas (líneas) similares a la siguiente línea, que contiene una dirección IP:

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

El formato exacto depende de las posiciones del archivo de configuración del servidor.

Es posible que una entrada contenga un mensaje de error, en cuyo caso el primer elemento no es una dirección IP.

Escribir una aplicación Python que procesa un log (dado por su nombre, ej. "C:\Apache\bin\logs") y genera un archivo con las frecuencias de las diversas direcciones IP en el log.

El formato del archivo de salida es el siguiente: en cada línea da la dirección IP, seguido de ":" y la frecuencia de las líneas del log que comienzan con dicha IP.

Por ejemplo: 127.0.0.1: 6791

Es necesario verificar que al comienzo de cada línea del log hay una dirección IP. Para eso, se puede utilizar `split()`, para obtener las cuatro cadenas que están separadas por puntos. Si estas cadenas son números entre 0 y 255, es una dirección IP válida.