LABORATORIO #2

DATOS ESTRUCTURADOS Y PROCESAMIENTO DE TEXTO

Este enunciado consta de TRES problemas independientes entre sí. Entregue la solución a cada problema en archivos independiente, de nombres `P1.py`, `P2.py` y `P3.py`.

1. La empresa RawInput S.A. desea hacer una segmentación de sus clientes según su ubicación geográfica. Para esto, analizará su base de datos de correos electrónicos con el fin obtener información sobre el lugar de procedencia de cada cliente.

   En una dirección de correo electrónico, el dominio es la parte que va después de la arroba, y el TLD (*top-level domain*) es lo que va después del último punto. Por ejemplo, en `usuario@dcc.uchile.cl`, el dominio es `dcc.uchile.cl` y el TLD es `cl`.

   - Siguiendo la receta de diseño, escriba la función `obtenerDominios(correos)`, tal que dada una lista de correos electrónicos, retorne la lista de todos los dominios, sin repetir y en orden alfabético.
   - Siguiendo la receta de diseño, escriba la función `contarTLD(correos)`, tal que dada una lista de correos electrónicos, retorne un diccionario que asocie a cada TLD la cantidad de veces que aparece.

2. En este problema buscamos analizar algunas propiedades léxicas en el idioma español. Para ello, considere el archivo `palabras.txt` adjunto a este enunciado, que contiene todas las palabras actualmente aceptadas por la RAE (una palabra por línea).

   - Dos palabras son anagramas si tienen las mismas letras, pero en otro orden. Por ejemplo, “torpes” y “postre” son anagramas, mientras que “aparta” y “raptar” no lo son, ya que “raptar” tiene una “r” de más y una “a” de menos. Siguiendo la receta de diseño, escriba la función `sonAnagramas(p1,p2)`, tal que indique `True` si `p1` y `p2` son anagramas, y `False` en caso contrario.
   - Las palabras panvocálicas son aquellas que tienen las cinco vocales. Por ejemplo, “centrifugado”, “bisabuelo” y “hipotenusa”. Siguiendo la receta de diseño, escriba la función `esPanvocalica(palabra)`, que indique si una palabra dada es o no panvocálica.
   - Siguiendo la receta de diseño, escriba la función `tieneLetrasEnOrden(palabra)` que indique si las letras de la palabra dada están o no en orden alfabético. Por ejemplo, “himnos” y “abenuz” están en orden, pero “zapato” no.
   - Use las funciones anteriores en un programa que:
     - Pida al usuario una palabra e imprima en pantalla todas aquellas en el diccionario español que son sus anagramas. Si no tiene anagramas, indíquelo apropiadamente con un mensaje ad hoc.
     - Escriba en el archivo “panvocalicas.txt” todas las palabras panvocálicas en el diccionario.
     - Escriba en el archivo “letrasEnOrden.txt” todas las palabras cuyas letras están en orden.

3. En el archivo “donQuijote.txt” se encuentra una transcripción de la obra “Don Quijote de la Mancha” en texto plano. Cargue este archivo en Python y procese el texto para imprimir en pantalla:

   - Las 10 palabras más frecuentes, junto a su frecuencia, ordenadas alfabéticamente
   - Las 10 palabras más frecuentes, junto a su frecuencia, ordenadas por frecuencia
   - Las 10 palabras menos frecuentes, junto a su frecuencia, ordenadas alfabéticamente
   - Las 10 palabras menos frecuentes, junto a su frecuencia, ordenadas por frecuencia
   - La cantidad total de palabras distintas
   - La cantidad total de palabras distintas, con al menos 4 caracteres