Universidad Nacional Autónoma de México

Facultad de Ciencias

Criptografía y Seguridad

Tarea Moral 2

García Ponce José Camilo - 319210536

1. Indice de Coincidencia encontrado
   IC = 0,07451973873818178 con 773688 letras (Los Pazos de Ulloa, sin limpiar)
   IC = 0,0743999101072844 con 694526 letras (El Código Da Vinci, sin limpiar)
   IC = 0,07535825484384993 con 796490 letras (Los Pazos de Ulloa, limpiado)
   IC = 0,0752270410742543 con 716071 letras (El Código Da Vinci, limpiado)
   Usamos dos textos diferentes y para cada texto una versión limpiada y la original.

2. Código usado
   El código usado para resolver este ejercicio se encuentra en el archivo intento1.py, los otros archivos importantes son pazos_ulloa.txt, codigo_da_vinci.txt (ahí están los textos usados, sin limpiar), pazos_ulloa_2.txt, codigo_da_vinci_2.txt (ahí están los textos usados, limpiados).
   En intento1.py esta todo el código usado. Los métodos usados fueron: método para limpiar un texto de caracteres espaciales, método para abrir un archivo de texto, un método para calcular el índice de coincidencia.

3. Anotaciones y proceso de resolver
   Para resolver este ejercicio primero buscamos diferentes textos en español con una larga cantidad de palabras, para esto usamos El Código Da Vinci y Los Pazos de Ulloa.
   Lo primero que hicimos fue limpiar a los textos de caracteres especiales (usando la biblioteca unidecode) para poder manejar mejor los textos.
   Después de obtener los textos hicimos un código para contar la frecuencia de cada letra (tomamos en cuenta solo estas letras abcdefghijklmnopqrstuvwxyz) usando un diccionario para esto (donde a cada letra le aumentamos un contador cuando la veamos en el texto) y luego contamos cuantas palabras leímos (sumando los contadores de la letras). Después calculamos el índice de coincidencia con la formula $IC = \frac{F_i(F_i-1)}{N(N-1)}$ donde F_i es la frecuencia del i-esimo carácter y N el total de caracteres (como lo vimos en la clase), para esto vamos recorriendo las letras del diccionario para calcular F_is y al final solo dividimos.
   De esta manera obtenemos el índice de coincidencia.
   Una observación interesante es que la cantidad de palabras es mayor en los texto limpiados, ya que así no se ignoran las vocales con acentos.