## INSTITUTO TECNOLÓGICO DE COSTA RICA II SEMESTRE 2017

**INGENIERÍA EN COMPUTACIÓN**

## IC-8020 RECUPERACIÓN DE INFORMACIÓN TEXTUAL TAREAS CORTAS 2-3

Esta tarea consiste en procesar la colección “a billion spanish words” y obtener su diccionario. Esto es, se debe obtener una lista de todos los términos de la colección. **Solo se aceptan como términos las raíces derivadas de palabras que son secuencias de letras consecutivas del alfabeto español: [a-zA-ZáéíóúüÁÉÍÓÚÜñÑ]**.

Se deben realizar las siguientes transformaciones a las palabras para transformarlas en términos:

* Convertir a minúsculas preservando las tildes y la eñe.
* Someterlas a un proceso de extracción de raíces usando la implementación **snowball** del algoritmo de Porter para español para el lenguaje de programación escogido. Esta herramienta está disponible para muchos lenguajes de programación. En la documentación deben indicar cuál stemmer usaron.

El resultado final debe ser un diccionario ordenado alfabéticamente por término que contenga la siguiente información por cada término distinto encontrado:

|  |  |
| --- | --- |
| **término** | forma final de las palabras luego de extraer su raíz |
| **número de variantes** | cantidad de palabras distintas encontradas que tienen a término como su raíz |
| **lista de variantes** | lista de palabras originales que tienen a término como su raíz |
| **frecuencia total** | conteo de la cantidad de veces que aparece un término en toda la colección |

El diccionario debe ser un archivo de texto con una línea por término en la que se incluye la información anterior.

La colección spanish\_billion\_words 100 directorios, 9436 archivos y 8.810.101.297 bytes.

Correr el programa para la siguiente cantidad de directorios y archivos (tomarlos en el orden en que vienen). Completar la información de la tabla. Llenar la tabla hasta donde el tiempo sea razonable (1 hora). El programa se considerará ineficiente si no puede procesar 16 directorios en el lapso de una hora.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **# directorios procesados (total o parcialmente)** | **#archivos tomados de esos directorios** | **# bytes procesados** | **# términos encontrados** | **# palabras distintas encontradas** | **tiempo requerido para procesar** |
| 1 | 1 | 734.346 |  |  |  |
| 1 | 3 | 1.610.733 |  |  |  |
| 1 | 6 | 4.516.877 |  |  |  |
| 1 | 16 | 11.305.210 |  |  |  |
| 1 | 40 | 27.786.527 |  |  |  |
| 1 | 101 | 70.269.147 |  |  |  |
| 3 | 303 | 216.713.045 |  |  |  |
| 6 | 588 | 414.603.183 |  |  |  |
| 16 | 1534 | 1.212.641.742 |  |  |  |
| 40 | 4002 | 3.632.227.195 |  |  |  |
| 100 | 9436 | 8.810.101.297 |  |  |  |

Con este enunciado se adjuntan los primeros 16 archivos del primer subdirectorio. El resto de la colección lo pueden pedir en clase o en horas de oficina.

**Esta tarea es individual.   
Esta tarea vale por dos tareas cortas / pruebas cortas.  
La fecha de entrega de esta tarea es el sábado 28 de octubre a medianoche.**

Deben entregar un directorio comprimido llamado ***T2-ApellidoNombre*** que contenga los siguientes archivos:

* documento que describe cómo correr el programa y que incluye la tabla descrita anteriormente; además debe indicar cuál stemmer usaron; finalmente incluir algún comentario o aclaración sobre el estado del programa
* código fuente del programa; puede ser un subdirectorio que contenga varios archivos
* diccionario con el resultado del caso más grande que lograron correr.