**INSTITUTO TECNOLÓGICO DE COSTA RICA I SEMESTRE 2020**

**INGENIERÍA EN COMPUTACIÓN**

**IC-8020 ELECTIVA TEXTO**

**TAREA PROGRAMADA 3**

**GUÍA DE DOCUMENTACIÓN**

1. Introducción

Presentar brevemente el problema. Puede “reciclar” partes del enunciado de la tarea programada.

En este proyecto se debe reducir el vocabulario de una colección de artículos de noticas, con el objetivo de hacer factible la aplicación de algoritmos de clasificación supervisada. Se debe escoger los términos que serán usados. Se debe genrar un txt con lista de las clases en que se clasifican los articulos. También uno para los documento y por úñtimo uno con el vocabulario de la colección.

1. Completar la siguiente tabla para describir el estado en que quedó el proyecto.

|  |  |  |
| --- | --- | --- |
| **Etapa** | **% de complet.** | **Comentario o aclaración** |
| **INDIZACIÓN** |  |  |
| Eliminación de stopwords | 100 |  |
| Extrae palabras **[a-z]+** | 100 |  |
| Extrae números **(\d){1,3}(,(\d){3})\*(\.\d+)?** | 100 |  |
| Descartar basura como **&#3** | 90 |  |
| **Archivos base** |  |  |
| Crea archivo ***clases.txt*** con datos <clase,nc> | 100 |  |
| Crea archivo ***docs.txt*** con datos <docid,clase> | 100 |  |
| Crea archivo ***dicc.txt*** con datos <termino,ni> | 100 |  |
| **Salidas calculadas** |  |  |
| Crea archivo ***gi.txt*** con datos  EntropíaTotal,  < ki, E(C,ki), GI(C,ki) > | 100 |  |
| Crea archivo ***mejores.txt*** con datos  < ki, E(C,ki), GI(C,ki) > en orden decreciente de GI | 100 |  |
| **Formato de las salidas** |  |  |
| Salidas son archivos de texto separado en líneas por cada ítem y con valores separadores por espacios. | 100 |  |
| Se usa un prefijo para todas las salidas. | 100 |  |
| **Pruebas** |  |  |
| Correr el programa para el archivo del enuncido  (***reut2-001.sgm***) usando los siguientes parámetros:  minNc=1 (aceptar todas las clases)  minNi=6  numMejores=200 | 100 |  |
| Correr el programa para el archivo de prueba que se les provee con esta guía (***TP3-Prueba.sgm***) usando los siguientes parámetros:  minNc=1 (aceptar todas las clases)  minNi=1 (aceptar todos los términos)  numMejores=8 |  |  |
| Comparar los resultados obtenidos en la prueba anterior con los valores esperados mostrados en el archivo ***TP3-Prueba-valores esperados.xlsx*** que se adjunta. |  |  |

1. Adjuntar los archivos producidos por las corridas con el archivo del enunciado y el archivo de prueba.
2. Describa brevemente cómo se hizo para eliminar los stopwords. Especialmente los stopwords tan particulares como los que tienen apóstrofe: "isn't", "aren't".

Se hizo un archivo txt que contiene todas las stopwords y se compara con cada palabra del texto original.

1. Comentarios finales (estado del programa)

Indicar el estado final en que quedó el programa, problemas encontrados y limitaciones adicionales.

1. Directorio con documentación

La tarea debe ser entregada presentando un archivo comprimido ZIP que incluya la siguiente estructura:  
 TP3-*ApellidosNombre1-ApellidosNombre2*  (directorio)  
 D*ocumentación*  
 *Programas*  
 *Archivos\_adicionales*

*Pruebas*

1. Entrega

**Enviar la tarea y su documentacion a la direccion** [**josee.arayamonge@gmail.com**](mailto:josee.arayamonge@gmail.com)**.  
Usar asunto: "** TP3-*ApellidosNombre1-ApellidosNombre2***"**