**INSTITUTO TECNOLÓGICO DE COSTA RICA II SEMESTRE 2021**

**INGENIERÍA EN COMPUTACIÓN**

**IC-8020 RECUPERACIÓN DE INFORMACIÓN TEXTUAL**

**Joseph Tenorio Pereira, 2019064588**

**Carlos Varela Ramírez, 2019077177**

**TAREA PROGRAMADA 3**

**DOCUMENTACIÓN**

1. Introducción

La tarea consiste en aplicar y evaluar dos clasificadores (Rocchio y bayesianos ingenuos) para una colección dada, la cuál se extrajo en por medio de dos archivos, los cuales contienen vectores de documentos extraídos de la colección de noticias Reuters21578. Dichos vectores se dan por medio de archivos tipo CSV con un formato especificado. El objetivo es leer de estos archivos los documentos correspondientes y aplicar los algoritmos de clasificación por medio del programa a crear. Sumado a esto, se realiza manualmente la evaluación de los resultados obtenidos por medio de las métricas de preicisión, recall, acierto y error.

1. Tabla de estado final del proyecto

|  |  |  |
| --- | --- | --- |
| **ETAPA** | **% DE COMPLET.** | **COMENTARIO O ACLARACIÓN** |
| **Lectura y conversion** |  |  |
| Puede leer y convertir datos de test\_set | 100 |  |
| Puede leer y convertir datos de train\_set | 100 |  |
| **CLASFICADOR ROCCHIO** |  |  |
| Calcula centroides | 100 |  |
| Calcula similitudes centroides-vectores para los tres casos | 100 |  |
| Produce un escalafón para cada documento del test\_set para cada uno de los tres casos | 100 |  |
| Asigna clases a los documentos del test\_set para cada uno de los tres casos | 100 |  |
| **CLASFICADOR BAYESIANO INGENUO** |  |  |
| Calcula los valores pip y qip para los términos | 100 |  |
| Calcula similitudes para los documentos del test\_set | 100 |  |
| Produce un escalafón para cada documento del test\_set | 100 |  |
| Asigna clases a los documentos del test\_set | 100 |  |
| Evaluación (puede ser manual) |  |  |
| Elabora la tabla de contingencia para cada clase | 100 |  |
| Calcula las métricas: acierto, error, precisión y exhaustividad | 100 |  |
| **EVALUACIÓN (puede ser manual)** |  |  |
| Elabora la tabla de contingencia para cada clase | 100 |  |
| Calcula las métricas: acierto, error, precisión y exhaustividad | 100 |  |

1. Comentarios finales

En general, se da por éxitosa la implementación de los dos clasificadores, así como la lectura de las colecciones de entrenamiento y prueba brindadas como material de pruebas para estos. No se encontraron mayores limitaciones respecto a la funcionalidad del programa, y se considera que este realiza lo esperado de manera razonablemente eficiente. Por otro lado, en cuanto a la evaluación manual realizada, se obvtuvo cabalmante los valores de presición, recall, acierto y error para las tres corrida de Rochhio y la corrida de bayesianos ingenuos efectuadas. Resulta de particular intéres el hecho de que las corridas de Rocchio tuvieran métricas mucho mejores si se le compara con el clasificador de bayesianos ingenuos.