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

**INGENIERÍA EN COMPUTACIÓN**

**IC-8060 RECUPERACIÓN DE INFORMACIÓN TEXTUAL**

**TAREA PROGRAMADA 3**

La tarea consiste en aplicar y evaluar dos clasificadores (Rocchio y bayesianos ingenuos) al caso que se describe a continuación.

Se les proveen dos archivos: ***training-set.csv*** y ***test-set.csv***. Esos archivos contienen vectores de documentos extraídos de la colección de noticias ***Reuters21578***. Cada línea de estos archivos contiene el vector de un artículo noticioso. Las líneas están separadas en cuatro campos. Se usan tabulaciones para separar los campos. Además, la primera línea de ambos archivos contiene los nombres de los campos:

* DOCID
* CLASE
* NUM-TERMINOS
* TERMINOS[termino/peso]

DOCID es el identificador del documento. CLASE es la clase a la que pertenece el documento. NUM-TERMINOS es el número de términos del documento. El último campo contiene una lista con los términos y sus correspondientes pesos. Los pares (término,peso) de esa lista están separados por un espacio en blanco. El caracter '/' separa cada término de su peso. A continuación se muestra un ejemplo de dichas líneas:

**3 → money-fx → 49 → addition/0.102383 afternoon/0.158492 ...**

**13 → trade → 110 → added/0.042619 admit/0.102966 also/0.028315 ...**

**16 → coffee → 143 → added/0.10897 already/0.059756 arap/0.101845 ...**

Clasificador Rocchio

Para cada clase Cp, se debe calcular su centroide usando la fórmula  
usando los documentos de la colección de entrenamiento. Luego aplicar la similitud entre cada uno de los documentos de prueba y los centroides. Para cada documento de prueba, escoger la clase con la similitud más alta. Ejecutar el clasificador Rocchio para los siguientes tres pares de parámetros:  
(β=0,75, γ=0,25), (β=0,85, γ=0,15), (β=0,95, γ=0,05).

Bayesianos ingenuos

Para cada clase p y cada término ki, calcular los valores pip y qip.  
usando los documentos de la colección de entrenamiento. Luego para cada clase Cp y cada documento dj del conjunto de prueba, calcule su similitud:  
Para cada documento de prueba, escoger la clase con la similitud más alta.

Evaluación

Cada corrida de un clasificador debe producir una salida con los siguientes datos:  
 docid **→** clase\_real **→** clase\_asignada\_clasificador

A partir de esas salidas se deben evaluar los resultados usando las métricas de ***precisión***, ***recall***, ***acierto*** y ***error***. Usar la tabla de contingencia. La evaluación puede ser realizada manualmente, no se requiere programarla.

|  |  |  |  |
| --- | --- | --- | --- |
| **Evaluación clase A** | **documentos que realmente pertenecen a la clase A** | **documentos que realmente no pertenecen a la clase A** |  |
| **documentos que el clasif. dice pertenecen a la clase A** | **(a) # docs de clase A que el clasif. asignó a clase A** | (d) #docs que el clasif. asignó a clase A y NO son de clase A | **(f) contar los casos en que el clasif. asigna clase A** |
| **documentos que el clasif. dice no pertenecen a la clase A** | (b) # docs de clase A que el clasif. asignó a otra clase | (e) #docs que el clasif. NO asignó a clase A y NO son de clase A *(no importa si las clases no coinciden, lo importante es que ambas sean ¬A)* | **contar los casos en que el clasif. asigna otra clase** |
|  | **(c) # docs de clase A en el test-set** | **(# de docs en test-set) -  (# docs en clase A en test-set)** | **(g) # de docs en test-set** |

Para calcular toda la tabla basta conocer los valores señalados en rojo. Los otros se calculan como diferencias por filas y por columnas.

Las métricas son:

|  |  |
| --- | --- |
| precisión = (a) / (f) | acierto = ( (a) + (e) ) / (g) |
| recall = (a) / (c) | error = ( (b) + (d) ) / (g) |

**Consideraciones finales**

Pueden usar cualquier lenguaje de programación. Se recomienda Python.

La tarea puede ser realizada en grupos de dos personas.

La fecha límite de entrega es el martes 16 de noviembre a las 11:55 pm.