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

**INGENIERÍA EN COMPUTACIÓN**

**IC-8060 RECUPERACIÓN DE INFORMACIÓN TEXTUAL**

**TAREA PROGRAMADA 3**

La tarea consiste en aplicar y evaluar el clasificador de vecinos próximos (K-NN) al caso que se describe a continuación.

Se les proveen tres archivos: ***training-set.csv***, ***test-set.csv*** y ***clases.csv***. Los dos primeros archivos contienen vectores de documentos extraídos de la colección ***Reuters21578*** la cual contiene artículos noticiosos. Los dos primeros archivos tienen una línea para cada documento. Cada línea contiene cuatro campos separados por tabulaciones:

* 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 elementos de esa lista están separados por un espacio en blanco. Cada elemento contiene el termino y su peso separado por el caracter '/'. 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 ...**

El archivo clases.csv contiene la cantidad de documentos para cada clase en los conjuntos de entrenamiento y de prueba:

**CLASS → N-TEST → N-TRAIN → TOTAL**

**coffee → 4 → 10 → 14**

**crude → 5 → 12 → 17**

**grain → 11 → 26 → 37**

**money-fx → 5 → 13 → 18**

**trade 6 → 15 → 21**

Se debe aplicar el clasificador K-NN con K=10 a los documentos de prueba y de entrenamiento. Luego evaluar los resultados para cada una de las 5 clases usando las métricas de ***precisión***, ***recall***, ***acierto*** y ***error***.

Para aplicar el clasificador se debe hacer lo siguiente:

* calcular la similitud entre cada documento del conjunto de prueba y cada documento del conjunto de entrenamiento; usar la similitud coseno (suma de productos de pesos); los vectores ya vienen con norma 1; de modo que solo hay que calcular la suma de productos, no hay que calcular normas ni dividir entre ellas
* para cada documento del conjunto de prueba tomar los mejores 10 documentos del escalafón y promediar por clase las similitudes obtenidas.
* asignar a cada documento del conjunto de prueba, la clase cuyo promedio de similitud sea mayor
* para cada clase evaluar los resultados usando la tabla de contingencia (la evaluación puede ser realizada manualmente, no requieren 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 en clase A** *(viene en clases.csv)* | **(# de docs en test-set) - (# docs en clase A)** | **(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)  
 recall = (a) / (c)  
 acierto = ( (a) + (e) ) / (g)  
 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 miércoles 16 de junio a las 11:59 pm.