Escuela Politécnica Nacional
Computación Distribuida
Proyecto 01

Coeficiente Jaccard/Tanimoto:

Se desea calcular la similitud química entre compuestos de acuerdo a la métrica denominada "Coeficiente Jaccard/Tanimoto".

Se tienen dos archivos de texto: `chemicals.tsv`, y `solution.tsv`

El archivo `chemicals.tsv` contiene un listado de compuestos, expresados como secuencias de caracteres. Cada línea del archivo corresponde a la información de un compuesto; dentro de cada línea, la primera columna asigna un identificador al compuesto, y la segunda columna indica la fórmula química del compuesto.

El archivo `solution.tsv` es un ejemplo de cómo se deben expresar los valores de similitud química entre compuestos. Cada línea expresa la similitud entre un par de compuestos, así, en la primera columna se escribe el primer identificador de compuesto (compuesto **a**), la segunda columna expresa el segundo identificador de compuesto (compuesto **b**), y la tercera columna expresa el valor del coeficiente Jaccard/Tanimoto entre los dos compuestos (valor **T(a,b)**).

El coeficiente Jaccard/Tanimoto **T(a,b)** se calcula mediante la siguiente fórmula:

$$T(a,b) = \frac{N_c}{N_a + N_b - N_c}$$

Donde **T(a,b)** es coeficiente de similitud entre el compuesto **a** y **b**, **Na** es el número de elementos en el compuesto **a**, **Nb** es el número de elementos en el compuesto **b**, y **Nc** es el número de elementos comunes entre los compuestos **a** y **b**.

Ejemplo:
Si el compuesto **a** tiene la fórmula "Cccc@@" y el compuesto **b** tiene la fórmula "CCcc@":

Na = 5 (1C, 3c, 1@)
Nb = 5 (2C, 2c, 1@)
Nc = 4 (1C, 2c, 1@)
T(a,b) = 0,67

Los símbolos '`@`' en las fórmulas son comodines, y se cuenta uno solo, aunque existan repeticiones. Es decir: "`@@`", y "`@@@`" se consideran equivalentes a '`@`'

La salida del programa se muestra en la consola.

Rúbrica de evaluación:

Los estudiantes deberán implementar el algoritmo que calcule la similitud química entre cada par de compuestos expresados en el archivo de entrada. Se debe calcular **T(a,b)** mediante una función que reciba como parámetros las fórmulas químicas de los compuestos **a** y **b**.

El problema tiene una valoración de 10 puntos.

- *3 puntos* por implementar un algoritmo paralelo en lenguaje Python

- *3 puntos* por implementar un algoritmo paralelo en lenguaje C, usando la librería OpenMP
- *1 punto* por aproximar el valor de **T(a,b)** para expresarlo únicamente con dos decimales.
- *1 punto* por implementar la condición de no considerar las repeticiones del símbolo '@'.
- *1 punto* por imprimir la salida del algoritmo en orden alfabético según el identificador del compuesto **a**.
- *1 punto* al código más elegante.