IC-3002 Análisis de Algoritmos
Ingeniería en Computación
Instituto Tecnológico de Costa Rica
Prof. Diego Munguía

Tercer Examen Parcial

Instrucciones

Debe realizar los ejercicios de programación que se especifican a continuación, utilice todos los conocimientos que ha adquirido en el curso hasta el momento.

Para todos estos ejercicios debe trabajar con un archivo de datos del proyecto GDELT. Este proyecto recolecta diariamente reportes de noticias sobre eventos socio-políticos que ocurren alrededor del mundo, los sistematiza y los publica en una plataforma abierta.

El archivo de datos sobre el que se trabajará para esta prueba contiene más de 230 mil eventos registrados el día 11 de mayo del 2017. Cada evento se registra como una diada, donde un actor 1 realiza una acción sobre un actor 2. Por ejemplo: "Rusia despliega tropas en Crimea, Ucrania" y "Estados Unidos critica a Rusia por sus acciones en Ucrania".

El archivo contiene datos tabulados en texto plano, cada fila de la tabla está separada por un cambio de línea (`\n`) y cada columna está separada por un caracter de tabulación (`\t`).

La especificación de cada una de las columnas de la tabla se puede encontrar en http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf .

Ejercicio #1: Ordenamiento en sitio

Implemente un algoritmo para ordenar la tabla utilizando la columna `SOURCEURL`. El algoritmo recibe como entrada el archivo que contiene la tabla de eventos GDELT y produce un nuevo archivo que contiene las mismas filas que el archivo original pero ordenadas ascendentemente por `SOURCEURL` como salida.

Restricciones

- Utilice python 2.7 para su implementación
- Su implementación no debe sobrepasar 64M de memoria utilizada
- Su implementación no debe tardar más de 3s en ejecutar
- No puede utilizar ninguna biblioteca. Debe implementar sus propios algoritmos auxiliares.

(65 pts)

Ejercicio #2: Agrupación

Implemente un algoritmo para encontrar la noticia, representadas por la columna `SOURCEURL`, que involucren la mayor cantidad de países distintos ya sea como actor 1 o actor 2. Sólo debe tomar en cuenta los valores con formato de URL para la columna `SOURCEURL`. El algoritmo recibe como entrada el nombre del archivo que contiene la tabla de eventos GDELT y como salida produce una tupla `(url, número de países)`.

Restricciones

- Utilice python 2.7 para su implementación
- Su implementación no debe sobrepasar 64M de memoria utilizada
- Su implementación no debe tardar más de 2s en ejecutar
- No puede utilizar ninguna biblioteca. Debe implementar sus propios algoritmos auxiliares.

(35 pts)

Puntos extra: Compresión

Implemente un algoritmo que comprima un archivo utilizando códigos de Huffman. El algoritmo recibe como entrada el nombre del archivo a comprimir y el nombre del archivo comprimido resultante; produce como salida un archivo comprimido almacenado bajo el nombre indicado en la entrada.

Implemente además un algoritmo que descomprima un archivo comprimido con el algoritmo anterior. Recibe como entrada el nombre del archivo comprimido y el nombre del archivo descomprimido resultante; produce como salida un archivo descomprimido almacenado bajo el nombre indicado en la entrada.

Restricciones

- Utilice python 2.7 para su implementación
- Su implementación no debe sobrepasar 64M de memoria utilizada
- Su implementación no debe tardar más de 2s en ejecutar
- No puede utilizar ninguna biblioteca. Debe implementar sus propios algoritmos auxiliares.

(35 pts extra)

Logística

El examen debe ser trabajado individualmente en un repositorio privado de GitLab. Debe proveer permiso de acceso al profesor. Fecha de entrega: miércoles 24 de mayo de 2017, a más tardar a las 11:59pm; cualquier commit posterior a esta fecha no será tomado en cuenta en la revisión.

Debe basar su solución en la plantilla de código adjunta `examen03.py`.