ST0257 – Sistemas operativos
Proyecto 3: Paralelismo en un ambiente controlado

MBA, I.S. José Luis Montoya Pareja
Departamento de Informática y Sistemas
Universidad EAFIT
Medellín, Colombia, Suramérica

RESUMEN
El paralelismo permite que las tareas se puedan realizar de manera más eficiente, si se programa bien. Usando elementos de conceptos vistos en el curso, podemos encontrar patrones y paradigmas de programación que nos ayuden a resolver este reto de la manera más eficiente posible.

PALABRAS CLAVE
Formatos de audio, procesamiento en paralelo, paradigmas de programación en paralelismo, eficiencia, procesos, virtualización, memoria.

CONTEXTO

Trending YouTube Video Statistics (J, 2019)

Daily statistics for trending YouTube videos

Trending YouTube Video Statistics

About Dataset

UPDATE: Source code used for collecting this data released here

Context

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, "To determine the year's top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they're not the most-viewed videos overall for the calendar year". Top performers on the YouTube trending list are music videos (such as the famously virile "Gangam Style"), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for.

This dataset is a daily record of the top trending YouTube videos.

Note that this dataset is a structurally improved version of this dataset.

Content

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

EDIT: Now includes data from RU, MX, KR, JP and IN regions (Russia, Mexico, South Korea, Japan and India respectively) over the same time period.

Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

The data also includes a category_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

For more information on specific columns in the dataset refer to the column metadata.

Acknowledgements

This dataset was collected using the YouTube API.

Inspiration

Possible uses for this dataset could include:

- *Sentiment analysis in a variety of forms*
- *Categorising YouTube videos based on their comments and statistics.*
- *Training ML algorithms like RNNs to generate their own YouTube comments.*
- *Analysing what factors affect how popular a YouTube video will be.*
- *Statistical analysis over time.*

For further inspiration, see the kernels on this dataset!

Así como este, hay muchísimas más fuentes de información gratis que se pueden descargar para analizar información.

ACTIVIDADES

La práctica consiste en que se lean los 10 archivos .csv del dataset en memoria de tres maneras diferentes a través del mismo programa.

NOMBRE

dataload – Lector de datos.

SINOPSIS

dataload [OPCIONES] -f FOLDER

DESCRIPCION

Se verifica si FOLDER es una carpeta que está en la misma ruta donde está el ejecutable. Por cada archivo con extensión .csv que encuentre en la carpeta, debe cargarlo en memoria en un ArrayList o similar.

El objetivo de leer los archivos es responder las siguientes preguntas:

1. Cual es el video más popular en ese año
2. Cual es el video más impopular.
3. Cuales son los videos más populares e impopulares por cada una de las regiones de Youtube.

Cuando termine el proceso, se muestra un mensaje de resumen que debe indicar:

1. Hora de inicio del programa
2. Hora de inicio de la carga del primer archivo
3. Hora de finalización de la carga del último archivo
4. Tabla de resumen con la duración de la carga de todos los archivos según el orden que hayan sido procesados
5. Tiempo (en formato mm:ss) que tomó todo el proceso.

Si no tiene OPCIONES habilitadas, se lee cada archivo uno a la vez de manera secuencial hasta que termine.

OPCIONES

-s

Indica al programa que lea los n archivos con extensión .csv que encuentre en la carpeta al tiempo, donde cada archivo que deba leer debe asignarse a un hilo independiente asignado en el mismo proceso (y por ende, mismo core) donde corre el dataload inicial

-m

Al igual que el -s, cada proceso recibe un archivo para ser leído pero cada proceso puede asignarse a cualquiera de los cores que tenga el computador disponible. A su vez, cada proceso crea hilos para leer el archivo por pedazos (un hilo lee 100 registros por ejemplo, el otro hilo lee otros 10 y así sucesivamente).

Estado de salida del proceso:

0 Si el proceso termina OK

1 Si el proceso termina con errores

El dataset lo pueden descargar de Interactiva Virtual.

CONSIDERACIONES GENERALES

1. El desarrollo de la práctica puede ser individual o en equipos de máximo tres personas.

2. La entrega de la práctica se realizará entregando los fuentes en un archivo y el informe por el buzón recepción de trabajos de Eafit Interactiva (cualquier otro medio no será admitido).

3. Se debe informar al profesor a más tardar el 23 de Julio a las 6:00 p.m. los integrantes del equipo.

4. El informe final es una presentación que deberá contener una breve descripción de cómo funciona el programa, tablas o gráficos donde se muestre la ejecución de su programa en diferentes máquinas describiendo de cada una el tipo de procesador, cantidad de memoria RAM y sistema operativo que tienen instalado (puede ser en varios) y unas conclusiones que ustedes hagan sobre los datos obtenidos.

5. La práctica se puede realizar en cualquier lenguaje de programación. En el informe deben informar cual es la versión del compilador o del runtime que están usando.

6. Cada semana los jueves, se sacará un espacio de 10 a 15 minutos al inicio de la clase para hablar de la práctica y resolver dudas.

7. Criterios de evaluación (ver Anexo 1)

FECHA DE ENTREGA

Jueves 7 de noviembre en clase a través de Eafit Interactiva.

SUSTENTACIÓN

Jueves 7 y martes 12 de noviembre en clase. El mecanismo de sustentación es el siguiente:

1. Cada equipo muestra su desarrollo ejecutando en vivo, dejando ver que el programa se ejecuta con los tres modos solicitados.
2. Debe mostrar cómo solucionó cada uno de los retos y cómo lo relacionaron con conceptos vistos en clase.
3. Mostrar los resultados y explicar las conclusiones.
4. Se realizarán preguntas por parte del docente para validar el entendimiento individual de los conceptos aplicados. Esto significa que,

aunque el trabajo es en equipo, la nota del trabajo puede ser diferente para cada uno de acuerdo con la calidad de las respuestas.

Nombre de la asignatura: Sistemas Operativos

Competencia a la que aporta la asignatura: Conocer el sistema operativo del computador para un mejor desarrollo, diseño y ejecución de las aplicaciones y aplicar nuevas soluciones.

Resultado de asignatura evaluado: Lectura de archivos secuencial y en paralelo. Evento evaluativo: Proyecto 3 % del evento evaluativo: 30%

Criterios (que tributen al RA de asignatura)	**Cumple con altos estándares (4.5 -5)**	**Cumple a satisfacción (4 -4.4)**	**Cumple parcialmente (3.5-3.9)**	**Incumple parcialmente (2.5-3.4)**	**Incumple totalmente (0 -2.4)**	**Peso asignado al criterio sobre la calificación.**
Análisis de fundamentación para virtualización de procesos y memoria. Claridad en el concepto	Entiende completamente la necesidad y plantea varias opciones de solución. Utiliza conceptos adecuadamente para la solución.	Entiende completamente la necesidad y plantea una opción de solución.	Omitió un elemento clave para el entendimiento de la necesidad	Omitió varios elementos para el entendimiento de la necesidad.	Demuestra poco o nulo entendimiento del problema.	40%
Diseño de la solución Solución óptima	El diseño tiene en cuenta los conceptos vistos en clase y argumenta la elección de su solución. La solución elegida es la óptima para el problema.	El diseño tiene en cuenta los conceptos vistos en clase y argumenta la elección de su solución.	Aunque se tuvieron en cuenta conceptos vistos en clase, no hubo argumentación correcta en la elección de la solución.	No se tuvieron en cuenta los conceptos vistos en clase.	Demuestra poco o nulo entendimiento de patrones de paralelismo al momento de explicar la solución.	40%

Funcionalidad Calidad de la solución frente a necesidad planteada	El programa funciona correctamente y se entrega de la forma descrita en el documento.	La solución elegida no es la óptima pero el programa funciona correctamente y se entrega de la forma descrita en el documento.	El programa con los conceptos vistos en clase no funciona correctamente o no se entrega de la forma descrita en el documento.	El programa no tiene en cuenta ningún concepto visto en clase y la solución funciona correcta o parcialmente o no se entrega correctamente.	Se entrega la solución parcialmente o no se entrega ninguna solución o no se entrega correctamente.	20%