

RETO 3: Crímenes en Los Ángeles

Objetivo

En este reto se pretende poner en práctica los conceptos aprendidos en clase acerca de las estructuras de datos que implementan los árboles y colas de prioridad para realizar búsquedas eficientes de información. Específicamente se espera:

1. **Practicar** los conceptos sobre estructuras de árboles binario de búsqueda, árboles balanceados y colas de prioridad.
2. **Integrar** el uso de las estructuras lineales (tablas de hash, Lista, pilas y colas) a los árboles binarios.
3. **Practicar** los algoritmos de ordenamientos y búsqueda binaria.
4. **Practicar** la carga y procesamiento de datos en formato CSV en memoria.
5. **Utilizar** adecuadamente el administrador de versiones GIT y GitHub.
6. **Trabajar en equipo.**

Fecha Límite de Entrega

El reto podrá entregarse hasta el martes **29 de abril**, 11:59 p.m. Ver las condiciones de entrega en la sección Entrega.

Contexto

Crime in LA es un proyecto que recopila los reportes de crímenes en la ciudad de los Ángeles desde el 2010.

Crime in LA los ha seleccionado para llevar a cabo un análisis del histórico de los crímenes almacenados en la plataforma. El conjunto de datos contiene un total de 315.000 reportes de crimen con un total de 20 columnas de información en cada reporte (como la fecha y hora, el área, ubicación (latitud y longitud) o el código del crimen).

Carga de Datos

Los datos del reto fueron tomados del proyecto **Kaggle** publicado por *la ciudad de Los Ángeles* titulado “Crime in Los Angeles.”¹ Los datos útiles para el proyecto están preparados y disponibles para los estudiantes en el aula **unificada** en Bloque Neón (BrightSpace).

¹ Crime in Los Angeles, Kaggle, URL: <https://www.kaggle.com/datasets/cityofLA/crime-in-los-angeles>

En particular, la información está conformada por cuatro archivos CSV en formato UTF-8 que se describen a continuación:

- a) El archivo **Crime_in_LA_*.csv**, contiene la información relevante a los crímenes reportados. La **Tabla 1** muestra las columnas incluidas dentro del reporte, con la explicación de cada una de ellas. El * en el nombre anteriormente mencionado hace referencia a el porcentaje de datos incluidos en el archivo.

NOMBRE DE LA COLUMNA	DESCRIPCIÓN
DR_NO	ID unico (número de expediente de registros) es un número de archivo oficial compuesto por un número de 2 dígitos del año, un ID de área y 5 dígitos adicionales.
Date Rptd	Fecha en que se reportó el incidente (formato MM/DD/AAAA).
DATE OCC	Fecha en que ocurrió el incidente (formato MM/DD/AAAA).
TIME OCC	Hora en la que ocurrió el incidente en formato de 24 horas.
AREA	Área geográfica asignada por el Departamento de Policía de Los Ángeles (LAPD). Existen 21 áreas numeradas del 1 al 21.
AREA NAME	Nombre del área geográfica o división policial dentro del LAPD. Estas divisiones pueden hacer referencia a un punto de referencia o a la comunidad circundante.
Rpt Dist No	Código de cuatro dígitos que representa un subárea dentro de un área geográfica. Este código se usa para estadísticas de delitos.
Part 1-2	Clasificación del delito según el sistema del FBI: Parte 1 (delitos graves) y Parte 2 (otros delitos menores). Corresponde a la gravedad de los delitos.
Crm Cd	Código del crimen cometido.
Crm Cd Desc	Descripción del código del crimen cometido.
Vict Age	Edad de la víctima
Vict Sex	Sexo de la víctima: F (Femenino), M (Masculino)
Vict Descent	Origen étnico de la víctima.
Premis Cd	Código del tipo de estructura, vehículo o ubicación donde ocurrió el crimen.
Premis Desc	Descripción del lugar donde ocurrió el crimen.
Status	Código del estado del caso.
Status Desc	Descripción del código estado del caso.
LOCATION	Dirección del incidente del crimen.
LAT	Latitud de la ubicación del incidente.
LON	Longitud de la ubicación del incidente.

Tabla 1. Nombres y descripción de las propiedades de los datos.

Para evitar problemas de buffer en la lectura de los archivos se recomienda aumentar el tamaño de los campos de lectura de la librería **Python CSV** al máximo posible para el sistema con el siguiente comando en la librería CSV en el **controller.py** del Reto.

```
import csv
...
csv.field_size_limit(2147483647)
```

En algunos casos experimentales puede que Python y el IDE declaren que se alcanzó el límite de recursión con un mensaje “**RecursionError: maximum recursion depth exceeded in comparison**”, en este caso se recomienda actualizar en el **view.py** este límite con las siguientes líneas de código:

```
import sys
...
default_limit = 1000
sys.setrecursionlimit(default_limit*10)
```

Trabajo Propuesto

Parte 1: Configuración Repositorio

Complete los siguientes pasos para configurar su repositorio de trabajo:

- p.1.** Cree en GitHub un repositorio basado en la plantilla propuesta para el reto, el cual se encuentra en el URL: <https://github.com/ISIS1225DEVS/ISIS1225-Reto-Template>
- p.2.** Renombre el repositorio de su reto con el formato **Reto3-G<<Número del grupo>>** ej.: **Reto3-G01** para el grupo 1 de la sección.
- p.3.** Edite el **README** del repositorio e incluya los nombres completos, correo Uniandes y códigos de los miembros del equipo de trabajo.
- p.4.** Descargue los datos desde la sección unificada del curso y cópielos en la carpeta **data** del repositorio local.

Parte 2: Carga de Datos

En la sección de **Reto 3**, de la sección **unificada** en Bloque Neón, encontrarán los datos oficiales del reto. El ZIP contiene varios archivos con los sufijos **_10**, **_20**, **_40**, **_60**, **_80** y **_100**. Estos son archivos con diferente cantidad de registros. Esto facilita la implementación y pruebas en computadores con memoria RAM y procesadores reducidos. En particular el archivo **_10** contiene únicamente 10 datos para poder realizar pruebas y debuggear su trabajo.

Para responder los requerimientos deberán cargar la información de los archivos entregados; recuerde que solo se permite leer una vez la información de cada archivo y que las pruebas finales sobre sus algoritmos serán sobre el archivo ***_100.csv**.

Al completar la carga de datos se debe reportar la siguiente información:

- El total de reportes cargados.
- La información de los cinco primeros y cinco últimos reportes cargados. Cada elemento debe contener:
 - DR_NO
 - Date Rptd
 - DATE OCC
 - AREA NAME
 - Crm Cd

Recomendaciones:

- Se recomienda utilizar la librería de Python `datetime`² para manejar el procesamiento y comparaciones de los datos temporales dentro de los archivos.
- Se recomienda utilizar librerías como `tabulate`³ para imprimir adecuadamente los resultados como se muestran en los ejemplos.
- En caso de que no exista algún valor dentro de los campos solicitados, para cualquiera de los requerimientos, completelos con el valor **“Unknown”**.

Parte 3: Desarrollo de los Requerimientos

Para este reto se ha identificado **ocho requerimientos: siete obligatorios y un bono**. Divididos según dificultad en básico, intermedio y avanzado. Adicionalmente, es importante resaltar que **tres** de estos requerimientos se deben desarrollar de forma **individual (I)** y los **cuatro** restantes son **grupales (G)**. El bono es grupal.

El resumen de los requerimientos se muestra en la siguiente tabla y se explican detalladamente en la siguiente sección.

Básico	Intermedio	Avanzado
REQ 1. Listar los crímenes ocurridos entre dos fechas (G)	REQ 3. Consultar los N crímenes reportados más recientemente para un área, dado su nombre (I)	REQ 6. Consultar las N áreas más seguras para un sexo, en un mes específico del año (G)
REQ 2. Listar los crímenes graves resueltos, que fueron reportados en un rango de fechas (G)	REQ 4. Consultar los N crímenes con víctimas en un rango de edad dado y clasificarlos por parte (I)	REQ 7. Determinar los crímenes más comunes para víctimas de un sexo en un rango de edad dado (G)
	REQ 5. Consultar las N áreas con mayor cantidad de crímenes no resueltos ocurridos en un rango de fechas (I)	REQ 8. (BONO) Determinar los N crímenes más cercanos y más lejanos del mismo tipo que ocurren en áreas distintas a partir de un área de interés (G)

NOTA: Si el equipo este conformado solamente por **dos** integrantes deberán resolver solamente **cinco** **Requerimientos obligatorios: uno** de los requerimientos **básicos**, **dos** de los requerimientos **intermedios**, y los **dos** primeros requerimientos **avanzados**.

² Librería Python `datetime`, Python URL: <https://docs.python.org/3/library/datetime.html>

³ librería Python `tabulate`, PyPi URL: <https://pypi.org/project/tabulate/>

Requerimiento 1 (Grupal): Listar los crímenes ocurridos entre dos fechas

Los **parámetros de entrada** de este requerimiento son:

- La fecha inicial del periodo a consultar (con formato "%Y-%m-%d").
- La fecha final del periodo a consultar (con formato "%Y-%m-%d").

La **respuesta esperada** debe contener:

- Identificador del reporte
- Fecha en que ocurrió el crimen
- Hora en que ocurrió el crimen
- Nombre del área en que ocurrió el crimen
- Código del crimen
- Dirección del crimen

Recomendaciones:

- Despliegue los reportes en orden cronológico del más reciente al más antiguo teniendo en cuenta el año, mes, día, hora y minuto en los datos.
- En caso de que existan varios reportes con la misma fecha y hora, ordénelos descendientemente por área en la que sucedieron.

Requerimiento 2 (Grupal): Listar los crímenes graves resueltos, que fueron reportados en un rango de fechas.

Los **parámetros de entrada** de este requerimiento son:

- La fecha inicial del periodo a consultar (con formato "%Y-%m-%d").
- La fecha final del periodo a consultar (con formato "%Y-%m-%d").

La **respuesta esperada** debe contener:

- Identificador del reporte
- Fecha en que ocurrió el crimen
- Hora en que ocurrió el crimen
- Área en que ocurrió el crimen
- Subárea en que ocurrió el crimen
- Gravedad del crimen ("Part 1-2")
- Código del crimen
- Estado del caso

Recomendaciones:

- Despliegue los reportes en orden cronológico del más reciente al más antiguo teniendo en cuenta el año, mes, día, hora y minuto en los datos.
- En caso de que existan varios reportes con la misma fecha y hora, ordénelos descendientemente por área en la que sucedieron.
- En caso de obtener una respuesta demasiado extensa (más de 10 elementos) se deben desplegar solo los cinco primeros y cinco últimos elementos de la respuesta con los detalles especificados.

Requerimiento 3 (Individual): Consultar los N crímenes reportados más recientemente para un área, dado su nombre

Los **parámetros de entrada** de este requerimiento son:

- El número (N) de crímenes reportados a consultar.
- Nombre del área de la ciudad a analizar

La **respuesta esperada** debe contener:

- Número total de crímenes que cumplen el criterio.
- Para cada crimen se debe presentar la siguiente información en orden cronológico de más reciente a más antiguo:
 - Identificador del reporte
 - Fecha en que ocurrió el crimen
 - Hora en que ocurrió el crimen
 - Área en que ocurrió el crimen
 - Subárea en que ocurrió el crimen
 - Parte del crimen
 - Código del crimen
 - Estado del caso
 - Dirección del crimen

Recomendaciones:

- En caso de que existan varios reportes con la misma fecha y hora, ordénelos descendientemente por área en la que sucedieron.

Requerimiento 4 (Individual): Consultar los N crímenes con víctimas en un rango de edad dado y clasificarlos por gravedad ("Part 1-2")

Los **parámetros de entrada** de este requerimiento son:

- El número (N) de crímenes a consultar
- Edad inicial del rango a consultar (con formato entero en años)
- Edad final del rango a consultar (con formato entero en años)

La **respuesta esperada** debe contener:

- Número total de crímenes que cumplen el criterio.
- Para cada crimen se debe presentar la siguiente información ordenados descendentemente (mayor a menor) por edad de la víctima:
 - Identificador del reporte
 - Fecha en que ocurrió el crimen
 - Hora en que ocurrió el crimen
 - Área en que ocurrió el crimen
 - Subárea en que ocurrió el crimen
 - Gravedad del crimen ("Part 1-2")
 - Código del crimen
 - Edad de la víctima
 - Estado del caso
 - Dirección del crimen

Recomendaciones:

- Primero deben aparecer todas las víctimas de crímenes graves, ordenadas descendentemente (mayor a menor) por edad, y luego todas las víctimas de crímenes menores, ordenados descendentemente (mayor a menor) por edad.
- En caso de tener víctimas con la misma edad, estas se deben ordenar cronológicamente con la fecha en que ocurrió el crimen (del más antiguo al más reciente).

Requerimiento 5 (Individual): Consultar las N áreas con mayor cantidad de crímenes no resueltos ocurridos en un rango de fechas

Los **parámetros de entrada** de este requerimiento son:

- La cantidad N de áreas a consultar
- La fecha inicial del periodo a consultar (con formato "%Y-%m-%d").
- La fecha final del periodo a consultar (con formato "%Y-%m-%d").

La **respuesta esperada** debe contener:

- Área en que ocurrió el crimen
- Nombre del área en que ocurrió el crimen
- Cantidad de crímenes no resueltos reportados en el rango de fechas
- Fecha del primer crimen
- Fecha del último crimen

Recomendaciones:

- Las áreas deben estar ordenadas por cantidad de mayor a menor crímenes. Si hay múltiples áreas con la misma cantidad de crímenes, entonces se debe ordenar por nombre del área.

Requerimiento 6 (Grupal): Consultar las N áreas más seguras para un sexo, en un mes específico del año

Los **parámetros de entrada** de este requerimiento son:

- El número N de áreas a consultar.
- El sexo por consultar
- El mes del año para realizar la consulta (como un entero)

La **respuesta esperada** debe contener:

- Área en que ocurrió el crimen
- Nombre del área en que ocurrió el crimen
- Cantidad de crímenes ocurridos en el mes dado
- Cantidad de crímenes ocurridos para cada año, en formato (crímenes, año)

Recomendaciones:

- Las áreas se deben ordenar de menor a mayor por cantidad de crímenes, si hay múltiples áreas con la misma cantidad de crímenes reportados, estas se deben ordenar por cantidad de años en lo que ocurrieron crímenes, y por último se deben ordenar lexicográficamente por el nombre del área

Requerimiento 7 (Grupal): Determinar los crímenes más comunes para las víctimas de un sexo en un rango de edad dado

Los **parámetros de entrada** de este requerimiento son:

- El número N de los crímenes más comunes a calcular
- El sexo de la víctima
- Edad inicial del rango a consultar (con formato entero en años)
- Edad final del rango a consultar (con formato entero en años)

La **respuesta esperada** debe contener:

- Por cada crimen listado entre los N más comunes presente la siguiente información:
 - Código el crimen
 - Cantidad de crímenes cometidos
 - Cantidad de crímenes ocurridos por edad de la víctima, en formato (crímenes, edad)
 - Cantidad de crímenes ocurridos para cada año, en formato (crímenes, año)

Requerimiento 8 (Grupal) - BONO: Determinar los N crímenes más cercanos y lejanos del mismo tipo que ocurren en otras áreas a partir de un área de interés.

Los **parámetros de entrada** de este requerimiento son:

- El número (N) de crímenes a consultar.
- Nombre del área de interés inicial.
- Tipo de crimen a consultar

La **respuesta esperada** debe contener:

- Tipo de crimen
- El nombre de la otra área
- Fecha del crimen en el área de interés
- Fecha del crimen en la otra área
- La distancia entre los crímenes

Recomendaciones:

- La lista de crímenes debe estar ordenada por la distancia entre las ubicaciones (latitud, longitud) de los crímenes.
- Para el cálculo de la distancia use la distancia de Harvesine
- El primer crimen entre los dos crímenes seleccionados debe ser el crimen más antiguo

Parte 4: Análisis de resultados

Dentro del proyecto debe incluir, en la carpeta **Docs**, un documento en formato **PDF** donde se evidencie el análisis de complejidad y las pruebas de tiempos de ejecución para cada requerimiento. Se sugiere que el documento tenga la siguiente distribución del contenido:

- Nombres, código y correo Uniandes de los integrantes del grupo.
- Para cada requerimiento individual se debe indicar quien lo realizó.
- Para los requerimientos grupales se debe indicar el porcentaje de contribución de cada estudiante

- Análisis de complejidad de los requerimientos en Notación O.
- Pruebas de tiempos de ejecución, con el conjunto completo de datos, para cada uno de los requerimientos. En estas pruebas se deben incluir:
 - Las tablas de tiempos de ejecución registrados.
 - Las gráficas comparativas de los experimentos.
 - Un análisis de resultados comparándolo los resultados obtenidos con el análisis de complejidad realizado.

Entrega

Para realizar la entrega del reto deben:

- 1) Agregar los usuarios de los monitores del curso a su organización de GitHub para hacer la entrega adecuada de la actividad
- 2) Dar permisos adecuados al repositorio a los monitores y al profesor, de lo contrario el reto **NO** podrá ser calificado
- 3) Asegurarse que la visibilidad del repositorio entregado sea privada y que solo pueda accederse con los permisos configurados para los integrantes del grupo
- 4) Enviar el enlace de GitHub en la actividad correspondiente dentro de Bloque Neón Uniandes (BrightSpace) de su sección
- 5) Incluir en el repositorio GIT todo el material, código y documentos solicitados durante la actividad.

IMPORTANTE: Recuerde que solo se calificará el material hasta el último **COMMIT** realizado previo a la Fecha/Hora Límite de Entrega indicada al inicio de este enunciado.