Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Tarea 1: Minería de Datos IIC2433

Profesor : Karim Pichara Baksai

Fecha de entrega : 24 de Septiembre 2018, 23:59

1 Descripción de la actividad

Esta tarea consiste en estudiar en profundidad e implementar el algoritmo *Apriori*, el cual tiene por objetivo encontrar itemsets frecuentes dentro de una base de datos y generar las reglas de asociación que superan umbrales de soporte y confianza. Posterior a la implementación, dicho algoritmo se someterá a prueba en la extracción de información en una base de datos real.

En *data mining*, las reglas de asociación son ampliamente utilizadas para descubrir relaciones entre variables en bases de datos de gran tamaño [Agrawal et al., 1993]. Aplicaciones clásicas de este tipo de estrategias pueden ser encontradas en análisis de compras y de características socio-demográficas desde bases de datos censales.

Con el fin de hacer un recorrido desde los detalles algorítmicos hasta la aplicación de las reglas obtenidas, usted deberá:

- **Implementar** el algoritmo *Apriori*. En este punto es fundamental que el código cuente con las funciones *fit* y *generate*. Donde *fit* debe aplicar el algoritmo a la base de datos y *generate* debe entregar las reglas de asociación. Para la implementación solo podrá utilizar las librerías *numpy* y *pandas*.
- **Aplicar** el algoritmo a la base de datos entregada y filtrar las mejores 10 reglas de acuerdo a dos criterios de calidad definidos por usted. El alumno también deberá presentar en los resultados dos filtros que involucren más de un criterio, por ejemplo $support \geq 0.1$ & $confidence \geq 0.5$.
- **Explicar** las reglas obtenidas. Seleccionar 4 reglas y comentar su calidad de acuerdo a los diferentes indicadores disponibles (*support*, *confidence* y *lift*).
- **Visualizar** las reglas, es decir, dado un conjunto de reglas proponer una gráfica que permita entenderlas y discriminarlas de manera directa. En este punto usted podrá hacer uso de todas las librerías de visualización disponibles.

2 Base de datos

La base de datos a utilizar corresponde a una parte de la información liberada por *Spotify* para el *RecSys Challenge 2018*[1] y contiene información de listas de reproducción creadas por usuarios de *Spotify*.

La base de datos que ustedes usarán tiene un total de 10.000 listas de reproducción con un número variable de canciones por lista y estará disponible en la página del curso junto al enunciado en el archivo *spotify.npy*.

3 Entrega

- La entrega debe ser realizada en un .zip con todos los archivos necesarios.
- El archivo de entrega debe ser subido al cuestionario abierto en el sistema SIDING específicamente para esta tarea hasta el día y hora señalada con el nombre $[numero\_alumno]\_T1$.
- En caso de atraso, se aplicará un descuento lineal de nota 7 a 1 en 24 horas.
- La tarea es estrictamente individual y el algoritmo debe ser implementado 100% (no usar funciones previamente implementadas o re-utilizar código).
- El documento principal debe ser un jupyter notebook con el código.
- Cualquier instrucción adicional y necesaria para la revisión debe ser escrita en un archivo README.txt contenido en el .zip

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[1] Consiste en implementar un recomendador de canciones a usuarios usando parte de su historial de reproducción.