Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación

Entrega 1

Clusterización de Sesiones de Usuarios Web

Alumno: Daniel Soto
Profesor: Pablo Guerrero
Fecha de entrega: 15 de agosto de 2018
Santiago, Chile

1. Introducción

El objetivo de este proyecto es encontrar clusterizaciones para las sesiones de usuarios de una página web en específico. Estas sesiones fueron obtenidas desde una base de datos de un repositorio que también busca hacer esto, y están basadas en la página nosfuimos.cl. Cada dato contiene los índices de las páginas visitadas en cada sesión, en el orden en el que fueron visitadas, junto al usuario que realizó la sesión y tiempos de inicio y fin de la sesión. En la base de datos existen alrededor de 12000 sesiones totales, las cuales componen 2100 sesiones únicas (sin repetición de sesiones).

2. Metodología

El proyecto se está llevando a cabo en un computador personal, debido a que no es necesario demasiado poder computacional para calcular las distancias entre 2000 sesiones. El análisis se lleva a cabo en notebook de jupyter.

3. Repositorio

El repositorio de github donde se lleva a cabo el proyecto es el siguiente.

4. Avances

Hasta ahora se ha realizado una exploración de datos sobre el dataset, la cual reveló que las métricas más obvias del dataset, como tiempos de inicio, fin y longitud de sesiones, no revelan ningún patrón particularmente interesante sobre las sesiones. Se crearon por lo tanto medidas como la originalidad de una sesión, la cual logra segmentar los datos un poco mejor, por lo que se podría usar en alguna medida de distancia más adelante.

$$\theta = \Sigma c_{p_i}^{-1}$$

Figura 1: Originalidad: medida propuesta para describir una sesión compuesta por las páginas p_i, donde c_{p_i} es el número de veces que la página p_i fue visitada en una sesión.

Debido a que los datos son sesiones discretas de largo arbitrario, se tuvo que implementar una versión de K-Means que utiliza una función de distancia entregada por el usuario para el paso de asignación de clusters y cálculo de centroides.

Para definir un número de clusters apropiado para generar, se decidió hacer un gráfico de las distancias entre puntos de cada cluster, con la intención de encontrar el *codo* donde se encontraría un buen número de clusters. Los resultados fueron bastante inesperados, pues en las clusterizaciones generadas esta medida era estrictamente creciente.

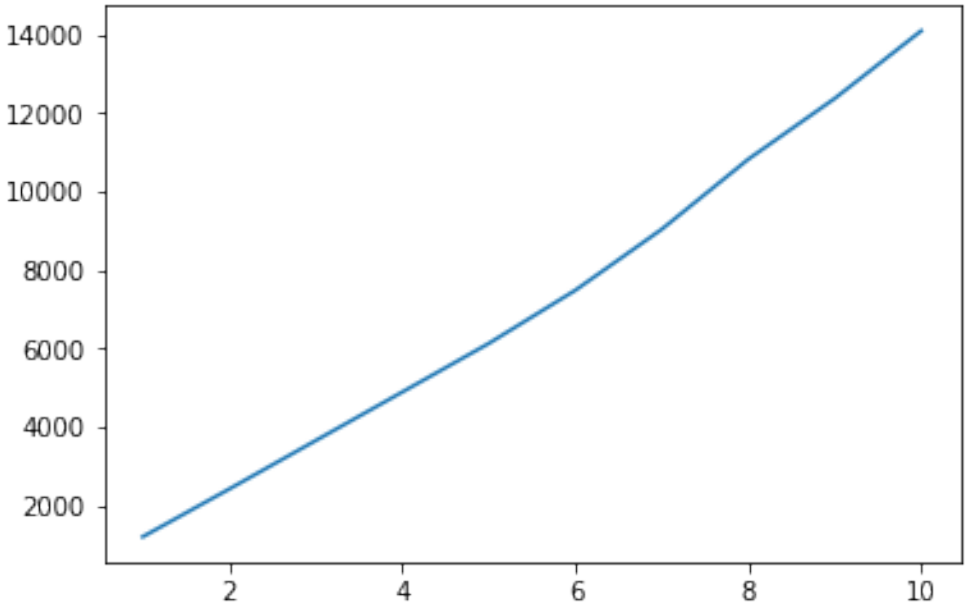

Figura 2: Gráfico de la WCSS generada, desde 1 hasta 10 clusters.

5. Tareas faltantes

Aún resta optimizar el algoritmo para realizar cálculos de centroides y generar el modelo y métricas sobre él más rápido. Luego de esto restaría evaluar la calidad de las clusterizaciones generadas y generar otras con distintas medidas de distancia entre las sesiones.

También podría ser un buen objetivo el analizar el porqué de los resultados inesperados obtenidos al aumentar el número de clusters a generar.

6. Instrucciones de uso

Para probar el avance actual, se debe generar una instancia del modelo DiscreteKMeans incluído en la entrega del notebook de jupyter, junto a una función de distancia, y luego se llama al método `fit` con un vector que contenga a los datos deseados. Este modelo luego tendrá un campo `clusters`, que contendrá un diccionario con los centroides como llaves, y una lista de sesiones asociadas a cada centroide como valor. El código del modelo se encuentra bien comentado para resolver dudas sobre su uso y/o funcionamiento.