Ejercicio 3: Procesamiento de texto

Dra. Sara Elena Garza Villarreal

Agosto 2019

Con este ejercicio conocerá comandos útiles para dar formato rápidamente a grandes cantidades de texto, así como obtener información valiosa del mismo. Para ello, utilizará comandos como head, tail, cut, sort, uniq y wc, así como el redireccionamiento de la salida y pipes.

Instrucciones.-

- 1. Documente su proceso utilizando el comando script. Nombre el archivo script-ej3.txt.
- 2. Descargue el archivo metadata.txt a home/Documents. El archivo se encuentra disponible en Recursos. Este archivo representa información sobre una corrida de una técnica de agrupamiento en un conjunto de datos extraído de Wikipedia. Los registros tienen los siguientes campos (separados por comas): clusterName, inGroupSize, outGroupSize, internalWeight, externalWeight, cohesion, rd, totalTimeMinutes.
- 3. Cuente la cantidad de registros del archivo.
- 4. Obtenga los siguientes datos:
 - a) Cluster (clusterName) con mayor inGroupSize.
 - b) Cluster con menor inGroupSize.
 - c) Tamaño del cluster (inGroupSize) con mayor rd.
 - d) Tamaño del cluster (inGroupSize) con menor rd.
 - e) Top 10 de los clusters con mayor rd (clusterName, inGroupSize, rd) y ordenar por tamaño.
 - f) Top 10 en orden descendente también por tamaño.

- g) Guarde estos dos rankings en un solo archivo llamado resultados-ej3.txt.
- 5. Recuerde que el script-ej3.txt debe ir en la carpeta scripts y resultado-ej3.txt en la carpeta resultados de su home/Documents.