Universidad Técnica Federico Santa María
Campus Santiago San Joaquín
Bases de Datos Avanzadas INF325
Primer Semestre 2019

Tarea I

Profesor: José Luis Martí Lara, jmarti@inf.usm.cl
Ayudante: Ignacio Tampe Palma, ignacio.tampe@sansano.usm.cl

1. Motivación

Se busca introducir al tema de las Bases de Datos Documentales a través del desarrollo de un caso de estudio utilizando el motor de datos **MongoDB** con el objetivo de conocer las características propias de este tipo de ambiente NoSQL. Adicionalmente, se deberá realizar un trabajo de minería de textos sobre parte del contenido de la base de datos construida.

2. Actividad: Base de datos Documental

Las Bases de Datos del tipo Documental permiten el almacenamiento y gestión de datos de un modo semi-estructurado que favorece por sobre todo el análisis de los meta-datos. Con el propósito de que el alumno adquiera una noción general de las bases de datos NoSQL del tipo Documental, los alumnos deberán construir una base de datos **MongoDB** y poblarla con uno de los datasets provistos:

2.1. Dataset A: Críticas de Apps del Google Play Store

Este Dataset contiene información de aplicaciones disponibles en el Google Play store, con su rating, categoría, número de instalaciones, etc. Junto a esto, se provee de un archivo de críticas de usuarios junto con su *sentiment analysis* por cada aplicación.

Se deberán asociar las críticas a cada objeto App provisto y poder hacer análisis tanto de la metadata como el corpus (Nombre de la app y críticas).
Disponible en: `https://www.kaggle.com/lava18/google-play-store-apps`

2.2. Dataset B: Críticas de empleados de empresas de tecnología

Este Dataset contiene críticas de empleados de compañías de tecnología como Amazon, Microsoft y Google. Cada crítica destaca pros y contras de trabajar ahí junto con distintos ratings e información adicional.

Se deberán asociar las críticas a cada empresa y poder hacer análisis de los comentarios positivos y negativos en base al resto de la información provista.
Disponible en: `https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews`

3. Actividad: Minería de texto

Una vez poblaba la base de datos, se debe realizar un trabajo analítico sobre el dataset. Para tal efecto, cada grupo de trabajo deberá desarrollar la experiencia vista en clases (minería de textos con R) sobre el dataset elegido que incluya resumen ejecutivo que presente el problema y el análisis hecho en un R Notebook. Su análisis deberá ser capaz de responder las siguientes preguntas:

- ¿Cuáles son las categorias (A)/ compañias (B) con mayor *ranking* promedio?
- ¿Cuáles son los términos descriptivos a los que más se hacen alusión?
- Posibles relaciones entre ellos.
- ¿Qué términos son los más comunes en comentarios de 3 categorias/compañias diferentes?
- ¿Qué términos son los más comunes en comentarios negativos y positivos?
- Generar una nube de palabras con los N términos más comunes.
- ¿Es posible identificar *clusters* de algún tipo entre los datos analizados?

Además, deberán **plantear una nueva interrogante** que haya surgido en el equipo y realizar el análisis/visualización necesaria para resolverla.

4. Evaluación y Entrega

Fecha de entrega: Domingo 14 de Abril 23:55.
Formato: Archivo comprimido en `tar.gz` con el R notebook y su versión en PDF.

4.1. Evaluación

Base de datos documental:

- Definir base de datos y colección de documentos: 10 ptos.
- Importar dataset con estructuras jerárquicas convenientes para el motor: 15 ptos.
- Búsqueda y filtrado de documentos: 10 ptos (se valorará la variedad y complejidad de las consultas).

Minería de texto:

- Análisis gráfico y cuantitativo de datos: 25 ptos.
- *Clustering* de textos: 15 ptos.

Resumen ejecutivo: 25 ptos. (A exponer ante el curso, 10 minutos por cada grupos - el profesor escogerá en el momento quien hará la presentación)

4.2. Consideraciones

- Equipos de 3 personas.
- Deberán elegir su dataset desde: `https://doodle.com/poll/v2y43tzgdy4nupk4` (cupos limitados)
- Puede que el dataset escogido necesite limpieza de datos.
- Importar el dataset sin usar una estructura jerárquica resultará en la **no revisión** de la tarea (se debe aprovechar el tipo de base de datos que se está usando).
- Copiar toda la base de datos a memoria en R no es válido, se deben hacer *queries* desde la base de datos para cada ítem del análisis.
- Consultas por moodle o correo.