Programa Ciencia de los Datos

Curso Minería de datos e Inteligencia de Negocios

Mayo 2020

Prof. Lorena Zúñiga

Tarea #4 Minería de datos - Web mining y clustering

Valor 13%

Objetivos

- Aplicar parcialmente la metodología CRISP-DM para resolver un caso básico de minería de datos
- Extraer datos de una página web para ser analizados posteriormente
- Aplicar expresiones regulares para procesar datos textuales
- Aplicar un algoritmo de clustering a los datos extraídos

Por hacer

- Según lo visto en clase y el material disponible realice lo siguiente:

Fases y actividades CRISP-DM

- **Entendimiento de los datos**
  - Exploración de los datos: busque en Internet una página web con datos que sean de su interés. Los datos deben ser parte de lo que la página presenta, es decir, deben ser tablas, párrafos, etc. visibles al cargar la página.

- **Preparación de los datos**
  - Seleccione los datos de su interés en la página previamente identificada
  - Limpieza de los datos: limpie lo que sea necesario aplicando expresiones regulares.
  - Construcción de nuevos datos (atributos). Si no aplica, indíquelo. Si construye nuevas columnas o atributos, explique.

- Transformaciones aplicadas a los datos. Describa las transformaciones realizadas.

Con los datos ya limpios y procesados forme un dataframe

- **Fase de modelado**

  Para esta fase se le solicita que seleccione dos de los algoritmos de clustering vistos en clase, o bien ejecutar dos veces sólo uno de los algoritmos estudiados pero usando diferentes parámetros cada vez.

  - Selección de técnicas
  - Construcción de cada modelo
    - Selección de los parámetros
    - Ejecución
    - Descripción del modelo obtenido (incluya al menos un gráfico por modelo)
  - Evaluación de los modelos
    - Compare los resultados obtenidos con cada modelo.

Exporte el documento RMarkdown a html

NOTA: tome en cuenta que no se tomará como válido utilizar datos provenientes de un archivo de texto o base de datos disponible en la página web seleccionada por el grupo de trabajo. Los datos completos a utilizar deben ser parte del texto que la página presenta.

Formato de entrega: archivo html únicamente

Forma de trabajo: grupos de 2 o 3 estudiantes únicamente

Forma de entrega: enviar documento html con la solución a través del TECDigital

Fecha de entrega: hasta el domingo 17 de mayo, a las 11:55 p.m.

Evaluación

Rubro	Valor
Entendimiento de los datos	**10**
Exploración de los datos	5
Verificación de la calidad de datos	5
Preparación de los datos	**45**
Selección de los datos	5
Limpieza de los datos	20
Construcción de nuevos datos (atributos)	5
Transformaciones aplicadas a los datos	15
Fase de modelado	**45**
Selección de técnicas	2
Construcción del modelo #1	4
Selección de los parámetros modelo #1	4
Ejecución modelo#1	3
Descripción de l modelo obtenido (además de la descripción, incluya al menos un gráfico)	9.5
Selección de técnicas	2
Construcción del modelo #2	4
Selección de los parámetros modelo #2	4
Ejecución modelo#2	3
Descripción de l modelo obtenido (además de la descripción, incluya al menos un gráfico)	9.5
Total	**100**