Proyecto TID

Alejandro Campoy Nieves
Gema Correa Fernández
Luis Gallego Quero
Jonathan Martín Valera
Andrea Morales Garzón

14 de noviembre de 2018

1. Comprender el problema a resolver

Para la realización y aplicación de las técnicas explicadas a lo largo del curso, hemos seleccionado un *dataset* proporcionado por *UCI Machine Learning Repository*. En concreto, hemos escogido **Drug Review Dataset**, una exhaustiva base de datos de medicamentos organizada por relevancia para medicamentos específicos. El conjunto de datos proporciona revisiones de pacientes sobre medicamentos específicos junto con las condiciones relacionadas. Además, las revisiones se agrupan en informes sobre tres aspectos: beneficios, efectos secundarios y comentarios generales. De igual modo, las calificaciones están disponibles con respecto a la satisfacción general, así como una calificación de efectos secundarios de 5 pasos y una calificación de eficacia de 5 pasos. Los datos se obtuvieron rastreando los sitios de revisión farmacéutica en línea.

DataSet Characteristics:	Multivariate, Text	Number of Instances:	4143	Area:	N/A
Attribute Characteristics:	Integer	Number of Attributes:	8	Date Donated	2018-10-02
Associated Tasks:	Classification, Regression, Clustering	Missing Values?	N/A	Number of Web Hits:	7001

Los datos se dividen en un conjunto train (75%) y otro conjunto test (25%) y se almacenan en dos archivos.tsv (tab-separated-values), respectivamente. Los atributos que tenemos en este dataset son:

1. **urlDrugName** (categorical): nombre de la droga
2. **condition** (categorical): nombre de la condición
3. **benefitsReview** (text): paciente sobre beneficios
4. **sideEffectsReview** (text): paciente sobre los efectos secundarios
5. **commentsReview** (text): comentario general del paciente
6. **rating** (numerical): clasificación de paciente de 10 estrellas
7. **sideEffects** (categorical): clasificación de 5 pasos de efectos secundarios
8. **effectiveness** (categorical): clasificación de efectividad de 5 pasos

2. Prepocesamiento de datos

Para poder analizar el dataset y realizar el prepocesamiento al mismo, lo primero que se ve hacer es explicado previamente, debemos hacer uso

```
# Para leer datos train
datos_train <- read.table("./datos/drugLibTrain_raw.tsv", sep="\t", comment.char="",
                          quote = "\"", header=TRUE)
head(datos_train, 5) # visualizar solo las 5 primeras filas
#summary(datos_train) # información sobre los datos
#View(datos_train) # ver la tabla

# Para leer datos test
datos_test <- read.table("./datos/drugLibTest_raw.tsv", sep="\t", comment.char="",
                         quote = "\"", header=TRUE)
#head(datos_test, 5) # visualizar solo las 5 primeras filas
#summary(datos_test) # información sobre los datos
#View(datos_test) # ver la tabla
```

Prepocesamiento de datos

Ver si algún atributo (columna) no nos hace falta

- Indicar que es relevante para nosotros, ver como están agrupados.
- Comprobar la columna "sideEffectsReview" ya que contiene NONE
- Comprobar la columna de ID si hace falta, ya que tenemos 8 atributos, y aparecen que tenemos 9 atributos.

Correlación

- Crear patrones en los textos, para saber que documentos hacen referencia a un tipo de fármaco, y podíamos sustituir los fármacos por un número.
- Leyendo el comentario que nos diga pertenece a este fármaco con una x probabilidad.
- Técnica de agrupamiento: coger los textos y agrupar (sin hacer clasificación), haciendo un "unit" a la columna de fármacos.

Para el analisis de sentimientos solo hace falta texto, tener en cuenta "benefitsReview" y "commentsReview". Los efectos secundarios y la efectividad que tienen las drogas, hacer un análisis de sentimientos.

El analisis textual hacer para: - los comentarios en funcion de la droga - en función del rating

Comprobación de celdas NA

```
#View(is.na(datos_test))
```