

# Práctica obligatoria y evaluable

Experto en Big Data: Módulo Big Data y Analítica de Datos con R

1. En la detección de Spam se utilizan con frecuencia técnicas de machine learning para mejorar los índices de detección de correos no deseados. En el dataset adjunto, se han seleccionado para cada mensaje una serie de términos clave que suelen aparecer con frecuencia en los mensajes spam. Posteriormente, se ha realizado una codificación vectorial de los correos electrónicos considerando esos términos clave. Para cada correo disponemos de la clasificación por parte de los expertos humanos. Se pide realizar las siguientes tareas:

- Sustituir un 2 % de valores de la matriz de datos por NAs, de manera aleatoria. Imputar dichos valores faltantes y justificar la elección del método utilizado.

1 punto

- Realizar un análisis exploratorio de la matriz de datos. Comentar los resultados y utilizar visualizaciones cuando sea necesario.

2 puntos

- Eliminar aquellas palabras que tengan una correlación elevada con otras. Calcular el número de documentos en que aparece cada palabra y eliminar aquellas de menor frecuencia. Dibujar un histograma de dichas frecuencias calculadas.

Nota: Con la codificación “bag of words” utilizada, el número de documentos en que aparece una palabra se obtiene sumando para cada variable todas las filas.

2 puntos

- Proyectar los datos sobre un subespacio de dimensión menor utilizando PCA. ¿Cuántas componentes principales se deben utilizar para poder visualizar la estructura semántica de los documentos ? Obtener un plot utilizando dichas componentes principales y comentar si hay estructura de grupo.

2 puntos

- Realizar un clustering de los mensajes de spam atendiendo a su contenido semántico. Discutir y comparar el resultado para dos algoritmos diferentes.

2 puntos

- Aplicar los mapas autoorganizativos para visualizar la estructura semántica de la colección de documentos utilizada. ¿Qué ventajas tienen los mapas autoorganizativos con respecto a los algoritmos de clustering utilizados anteriormente ?

1 punto

Nota: Se debe entregar un comprimido con el pdf resultado de compilar el fichero rmarkdown y el .Rmd. El documento pdf debe incluir comentarios y discusión de los resultados obtenidos. Se recomienda utilizar tablas para comparar los diferentes modelos atendiendo a los índices de error pedidos y visualizaciones cuando se consideren necesarias.

**Criterios de evaluación:** La resolución de las tareas pedidas se evaluará considerando los siguientes criterios:

1. Planteamiento y justificación del método elegido para resolver el problema
2. Implementación en R
3. Discusión de los resultados
4. Organización y presentación del informe técnico