2016/2017

Ingeniería basada en el conocimiento

Opiniones sobre hoteles

Héctor Rodríguez Salgado & Marta Loriente Nieves

Máster en Ingeniería Informática – ESEI – Universidad de Vigo

# Introducción

El objetivo principal del análisis de datos es la construcción y comparación de diferentes modelos de minería de datos, enfocando la discusión en sus resultados.

Se evaluarán un conjunto de datos siguiendo tres enfoques de análisis diferentes: modelos predictivos, modelos de agrupamiento y modelos de asociación.

Todos estos modelos se construirán utilizando la herramienta de minería de datos WEKA y sus algoritmos tal como se proporcionan, sin modificar ninguno de sus detalles de implementación.

# Exploración de datos

Para realizar los análisis se ha escogido un conjunto de datos de alrededor de 12.000 instancias, de un conjunto de datos que recogen las opiniones de los clientes que han visitado distintos hoteles. Dicha información se ha obtenido del buscador de hoteles *Tripadvisor*.

En un inicio, el *dataset* contenía la siguiente información:

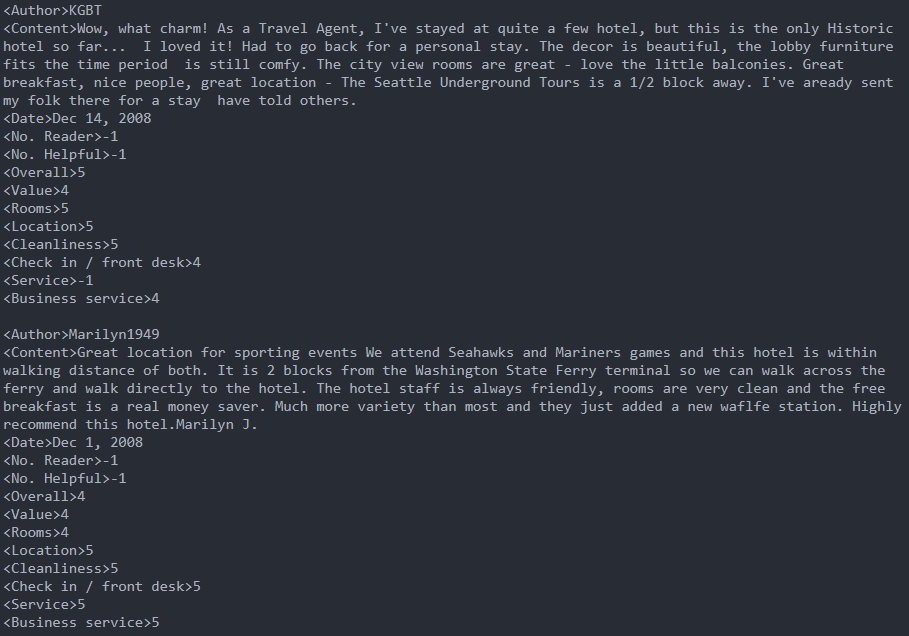


Ilustración 1 - Dataset inicial

Para el análisis, únicamente se requería la opinión del cliente. Por tanto, se ha realizado un preprocesado de los *datasets*, obteniendo como resultado:

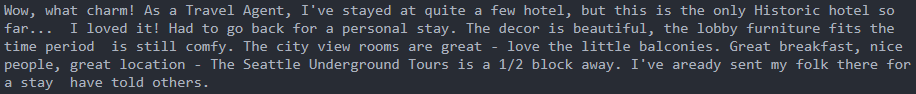


Ilustración 2 - Dataset tras el procesado

De esta forma, cada instancia contiene la opinión de un cliente en un hotel.

Una vez realizado dicho proceso, se han escogido las 50 palabras que más se repiten en todas las opiniones para determinar si la opinión del cliente ha sido buena o mala hacia el hotel en el que se ha hospedado. Dichas palabras pasarán a ser los atributos de interés para el análisis.

# Análisis de predicción de atributos

La primera tarea es determinar si es posible predecir si la opinión de un cliente es buena o mala en base a las palabras que contiene su opinión. Para ello se han ejecutado diferentes algoritmos predictivos y se ha observado el resultado.

## Lazy learning

El primer algoritmo que se ha decidido utilizar ha sido de tipo “lazy learning: (k-) nearest-neighbour”, conocido como IBK en WEKA. Se ha decidido realizar una predicción para cada atributo del dataset, variando el número de vecinos en cada análisis. La tendencia en cada uno de los análisis es que a más vecinos, mejor es la predicción realizada, ya que el porcentaje de acierto es mayor.