

Práctica 6.

Natural Language Processing

*Máster en Big Data*

*Curso 2019-2020*

***Data Science & Advanced Analytics***

*Montse Llos i Bombardó*

*Práctica 6.3*

La cadena hotelera VamosAúnMejor quiere conocer fácilmente las experiencias de sus clientes en los diversos hoteles de su cadena.

Para ello, nos pide que les desarrollemos un sistema automatizado donde se puedan clasificar los comentarios en redes sociales sobre sus hoteles. Les interesa obtener una clasificación de los mensajes en 5 grupos, donde 5 suponga mejor experiencia y 1 peor.

Para el proceso de aprendizaje partiremos de unos textos ya clasificados del 1 al 5. Nosotros tendremos que crear el modelo que aprenda de las características que determinan la nota de experiencia para después poderlo aplicar sobre los nuevos mensajes que se publiquen en las redes sociales.

De esta forma, podrán evaluar más fácilmente qué opinan los clientes de sus hoteles y qué deben mejorar o enfatizar.

Los datos ya los tenemos cargados dentro del repositorio de RapidMiner.

Deberemos crear:

* Un proceso que genere el modelo
* Un proceso que use el modelo para clasificar los nuevos mensajes

*Práctica 6.4*

Vamos a realizar un web scrapping para identificar automáticamente páginas web que puedan tener un contenido similar

*Ejercicio 6.5*

En la cadena hotelera VamosAúnMejor quieren hacer un estudio sobre su personal a partir de los datos que disponen de sus trabajadores más las estadísticas que anualmente les hacen para conocer qué mejorarían en su puesto de trabajo.

El objetivo es ayudarles a predecir cuándo un empleado tiene una alta probabilidad de marchar de la empresa.

Los datos a usar los encontrareis en el campo virtual en la carpeta de Prácticas 6.

Los datos que disponemos son los siguientes:

* CanDoBetter: campo de comentario libre donde cada trabajador opina de aspectos a mejorar en la empresa
* ID: Identificador del trabajador
* Training: indica si tiene alguna formación
* Est\_income: salario bruto anual
* Benefit: indica si ha recibido algún beneficio por parte de la empresa
* Lenght\_of\_hire: número de días que lleva trabajando en la empresa
* Facility: tipo de facilidades que dispone
* DriveTime: Tiempo de desplazamiento entre domicilio y área de trabajo
* Marital: Estado civil
* SpousalBenefit: Beneficios conyugales
* StockPlan: Indica si tiene acciones de la empresa
* TrainingReceived: indica si ha realizado alguna formación dentro de la empresa
* PositionType: posición en la empresa
* Education: educación
* Status: Current indica que continúa trabajando en la empresa. Past indica que ha marchado de la empresa

La práctica se debe realizar en **Python** o **RapidMiner**. Se deben comentar todas y cada una de las líneas.

Se deberán probar un mínimo de 3 algoritmos distintos.

Número límite de personas: 3

Entregables:

* Un fichero con el código Python o RapidMiner
  + Nombre del fichero: <Apellidos+InicalNombres>\_Ejercicio\_6.5. .ipynb
  + Ejemplo si sois dos: *LLosM\_RooneyM\_Ejercicio\_6.5.ipynb*
* Un documento Pdf con nombre y foto de la persona que ha realizado el ejercicio y conclusiones del ejercicio con los resultados con cada uno de los 3 algoritmos probados
  + Nombre del fichero: <Apellidos+Nombres>\_Ejercicio\_6.5.pdf
  + Ejemplo si sois dos: *LLosM\_RooneyM\_Ejercicio\_6.5.pdf*

Fecha límite entrega: 6 abril 2020 a las 24h

Subir al campus