PRÁCTICA 2 (1/2)
Data Mining

Inteligencia Artificial en las Organizaciones

Grado en Ingeniería Informática

Curso 2023/2024

INTRODUCCIÓN

La minería de texto (text mining) es un área que en la actualidad está experimentando grandes avances. Una de las principales características de este campo es que los datos procesados son no estructurados. En esta práctica, nos centraremos en las técnicas clásicas para representar un texto como un vector de frecuencias de aparición de las palabras que lo componen. Con esta representación se pueden utilizar técnicas estándar de minería de datos para analizar los documentos, por ejemplo, clasificarlos o analizar la similitud entre dos textos.

Para alcanzar este objetivo, vamos a utilizar la herramienta RapidMiner con la que se pueden realizar procesos de minería de texto de forma rápida y eficiente.

OBJETIVO

El objetivo de la primera parte de la práctica es entrenar un clasificador que tome como entrada una reseña de un hotel en castellano y tenga como salida la puntuación del hotel otorgada por el cliente.

Para esta ello, utilizaremos una colección de reseñas sobre hoteles en Andalucía en español[1], compilada a partir de la web TripAdvisor y disponible en Kaggle.

Este corpus (en minería de texto, al conjunto de textos que se analizan se le denomina corpus) se llama AHR (andalusian hotel reviews) y ha sido publicado por Mariia Chizhikova. Incorpora otro anterior compilado por el grupo de investigación SINAI en 2014 (que está disponible en abierto en el sito web de SINAI[2])

[1] https://www.kaggle.com/code/chizhikchi/ahr-corpus-presentation/input

[2] Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A. (2014). Cross-domain sentiment analysis using spanish opinionated words. Natural Language Processing and Information Systems, Lecture Notes in Computer Science, vol. 8455, pp. 214-219. Springer International Publishing. DOI: 10.1007/978-3-319-07983-7_28

El corpus AHR contiene 18.172 reseñas. También hay disponible otro conjunto más pequeño (7,615) pero balanceado

El contenido de los ficheros, que están en formato csv, es el siguiente

- *title*– titulo de la reseña
- *rating*– calificación del establecimiento dada por el autor de la reseña, de una a cinco estrellas
- *review_text*- texto completo de la reseña
- *location*- ciudad y region donde se sitúa el otel
- *hotel*- nombre del hotel
- *label*- etiqueta generada por la autora del corpus, en el que se resume la clasificación en tres niveles en vez de los cinco originales

En la versión básica de la práctica debéis usar texto completo de la reseña (review_text), y entrenar un clasificador que dé como salida tres niveles (label). Podéis explorar también un clasificador que clasifique en los cinco niveles originales, y también ver si mejora el rendimiento usando también el título o sólo el título.

Se recomienda que todo el preprocesamiento de los datos se haga con RapidMiner para que los resultados sean más fácilmente reproducibles.

Sesión I: Clasificador de opiniones

1. Extensiones de RapidMiner

La funcionalidad básica de RapidMiner puede ampliarse fácilmente con el uso de extensiones, que son módulos externos que realizan tareas concretas. Estas extensiones se descargan a través de menú Extensions > Marketplace (updates and extensions). Las extension que necesitarás inicialmente para esta práctica es Text Processing (para el pre-procesado de texto), , que suele estar incluida en la instalación básica. Puedes comprobar si ya está instalada eligiendo Extensions > Manage Extensions.

2. Carga de texto en Rapid Miner

Lo primero que debemos hacer es descargar los datos desde Kaggle y cargarlos en RapidMIner para su análisis. El original está en formato csv. Podéis pasarlos a Excel, preprocesarlos y después cargarlos o cargarlos directamente desde csv para preprocesarlos.

RapidMiner puede manejar texto de diferentes formas. Las principales son colecciones de documentos, que son documentos como objetos independientes dentro de un objeto

(objeto “document collection”), y conjuntos de ejemplos en formato tabla, como los campos leídos de un fichero Excel (objeto “exampleSet”). Como veremos más adelante algunos operadores están disponibles en dos versiones para trabajar con cada tipo de datos. También existe un operador que convierte un “exampleSet”en “document collection”y viceversa.

En esta práctica vamos a trabajar con los datos leídos de una hoja Excel o csv, y cargados en un exampleSet.

Añade a tu proceso el operador Read Excel o Read csv (Figura 1). El camino más rápido para encontrar cualquier operador es teclear su nombre en la caja de texto search for operators que mostrará todos los operadores que coincidan con el texto buscado.

La **configuración del operador que lee los datos es muy importante** para no tener problemas más adelante. Esta configuración se puede hacer usando el wizard “import configuration wizard” o directamente editando los valeres en la ventana.

Es importante establecer lo siguiente

1. Qué campos se cargan de fichero y cuales no necesitamos
2. Qué papel juega cada campo y el tipo de datos. En nuestro caso el texto es un atributo (attribute), y la puntuación es una etiqueta (label).
3. El tipo de datos: es muy importante comprobar que el campo de texto tiene asignado el tipo de datos texto, pues en caso contrario los operadores de manejo de texto que vamos a usar no funcionarán.

Todos esos parámetros se definen en el botón **“dataset meta data information”** que lanza una ventana “pop up” para la configuración (un aviso, este botón queda oculto cuando se visualiza RapidMiner en una ventana que no ocupa la pantalla completa, conviene tener cuidado con esto pues a veces no se ven todos los campos u opciones).

Si preferís que esta configuración sea más visible, podéis usar el operador “Nominal to Text” para convertir los atributos a tipo texto, el operador “Set Role” para definir el rol de los campos y el operador “Filter attributes” para conservar sólo los campos que queremos analizar y eliminar el resto. También se puede usar “Filter examples” o “Sample” para quedarse con un subconjunto de los ejemplos para las pruebas iniciales o para descartar ejemplos no válidos (por ejemplo, que no tengan etiqueta)

Para comprobar si la carga de datos es correcta, conectamos la salida del operador al puerto res (resultados), y ejecutamos el proceso (icono Run). Si pasamos de la pestaña “Design” a la pestaña “Results”, veremos los datos leídos en forma de tabla.

3. Análisis descriptivo de los datos

En este punto es recomendable hacer un análisis descriptivo de los datos, ver el número de reseñas, numero de reseñas por clase, si el conjunto está o no balanceado, las palabras más frecuentes etc. Una forma es utilizar las herramientas (visualizations) disponible en la pestaña de resultados. Algunos ejemplos son un gráfico de barra en el que veamos cuantas entradas hay de cada tipo, o una nube de palabras (Word cloud) para los datos tipo texto.

4. Limpieza de los datos

Como estamos trabajando con datos reales, vamos a encontrar diferentes problemas e inconsistencia de los datos. En concreto en este corpus, veréis que las tildes en español han perdido el formato. Un posible enfoque es usar el operador “Replace - dictionary” para sustituirlos (en aula global tenéis un diccionario que podéis usar como base).

5. Generación de la matriz de términos por documento

El siguiente paso es procesar el texto, siguiendo los pasos típicos del proceso de minería de texto, y crear la matriz de términos por documento en el formato que elijamos. En nuestro problema cada línea del fichero Excel es un documento del corpus a efectos del procesamiento de la información.

Usaremos el operador “Process documents from Data” (Figura 1), que espera como entrada una tabla con los textos en formato “example Set” y da como salida, el texto procesado. Nótese hay otro operador equivalente, llamado “Process documents from file”, que es el que deberíamos usar si nuestros textos están en documentos en un directorio en vez de en un Excel o csv.

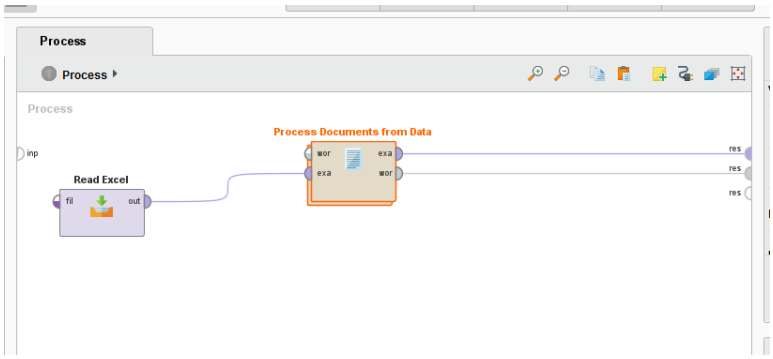


Figura 1. Procesar texto

Una vez añadido el operador "*Process documents from Data"*, lo desplegamos (doble click) para insertar los operadores correspondientes (es decir, se insertan subprocesos dentro del proceso). Estos son los pasos más comunes de pre-procesado del texto que puedes añadir (Figura 2)

1. Identificación de los términos individuales (tokens) en el texto. Para ello, se utiliza el operador *Tokenize*.
2. Filtro de palabras que no son de interés (Filter Stopwords). RapidMiner no tiene incorporado un filtro para palabras en español, aunque si para inglés. Puesto que nuestro texto está en español, podemos buscar una lista de stopwords, generar un fichero de texto con ella, y cargarla usando Filter Stopwords (Dictionary).
3. Poner todo en minúsculas: Operador *Transform cases*
4. Eliminar palabras de dos o menos letras: *Filter Tokens (By Length)* – el número de letras es configurable
5. Reducir las palabras a la raíz (stemming o lematización). El operador que realiza stemming con el algoritmo snowball da resultados razonables para español.

En función de la tarea a realizar, algunos de estos pasos son adecuados o no. Por ejemplo, stem puede no ser necesario y empeorar el resultado. Hay otras opciones de pre-procesado que pueden ser útiles, como eliminar palabras de un diccionario establecido por nosotros. Para ello habría que crear el diccionario

Cuando terminemos esta parte puedes probar a añadir o quitar algunos pasos del procesado de los textos y ver el efecto.

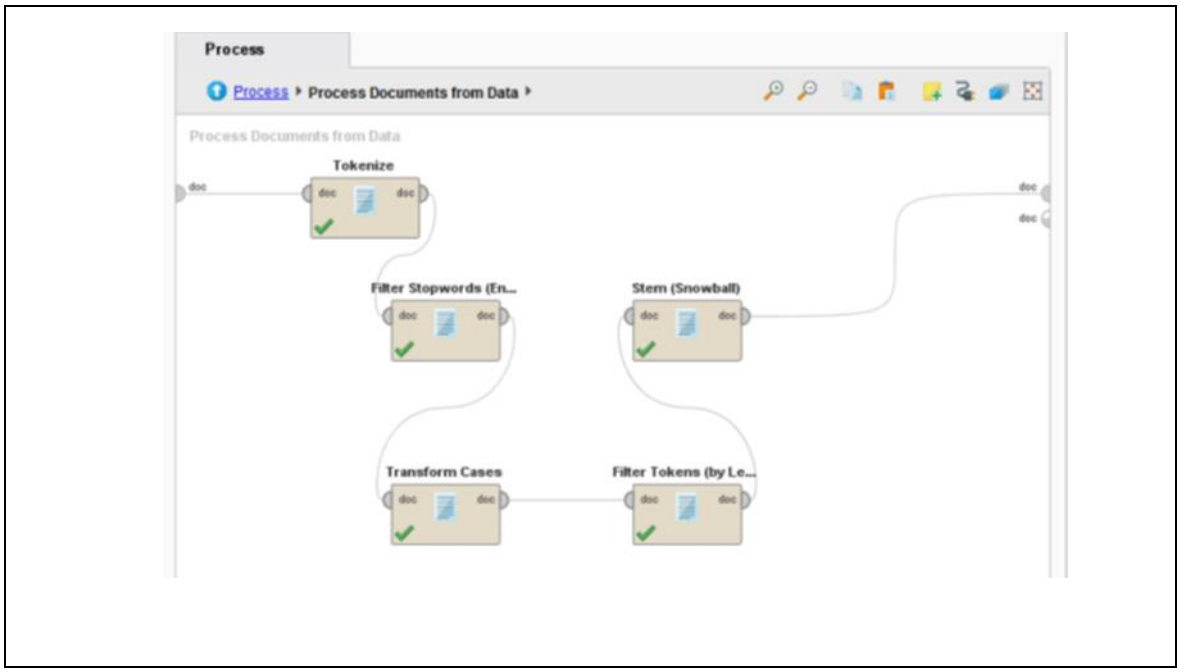


Figura 2. Ejemplo de sub-proceso englobado en el operador Process Documents from Data.

Volvamos a la pantalla principal de diseño (pulsando sobre el nombre del proceso o sobre la flecha que está junto al nombre) para configurar el operador “Process Documents from Data” y ver el resultado del proceso.

Como se puede ver en la Figura 1 el operador tiene dos puertos de salida exa y wor. En el primero se genera la tabla de términos por documento (tipo de datos Example Set) y en el segundo, una lista de las palabras del corpus con su frecuencia de aparición (tipo de datos Word List). Conectaremos ambos a dos puertos de resultados.

Antes de ejecutar el proceso, estableceremos la configuración del operador “Process Documents from Data”. El principal parámetro a configurar es la forma de crear la matriz de términos por documento, según como queramos que se represente la frecuencia de cada término: cuenta (term occurrence), frecuencia, frecuencia binaria (presencia vs ausencia) y tf-idf (frecuencia de término – frecuencia inversa del documento, term frecuency –inverse document frecuency). Estas opciones se establecen en el desplegable vector creation.

Comienza explorando la opción básica (cuenta, term occurrence), que es más facil de interpretar. Prueba después con otras representaciones.

Además, el operador “Process Documents from Data” permite establecer un método de poda (prune) que limite los términos que se utilizan para generar la matriz de términos por documento, eliminando los muy frecuentes y poco frecuentes. En función de la tarea a realizar, este paso será útil o no. Por ejemplo, se puede establecer una poda ligada al porcentaje (percentual), haciendo que la cota inferior (prune_below_percent) sea el 3%, dejando la cota superior (prune_above_percent) al 100%.

Ejecutemos el proceso para inspeccionar los dos resultados: matriz de términos por documento (de tipo example set) y lista de palabras (de tipo word list). Veamos primero la word list. Si ordenamos las palabras por el número de veces que aparecen, podemos ver los términos más comunes, y cómo de comunes son para cada una de las clases por separado. ¿Crees que bastaría con esta cuenta para clasificar las reseñas? Si inspeccionamos el example set, veremos la matriz en la que para cada entrada y término tenemos el número de veces que aparece. Podemos comprobar que es una matriz dispersa (la mayoría de los términos son 0). Puedes explorar diferentes visualizaciones de la matriz términos por documento en la ventana de respuestas, incluyendo la opción de visualizar nubes de palabras,

Ahora, repite la ejecución del proceso con otras representaciones de los documentos como term frequency y TF-IDF, y quizá con otros valores y métodos de pruning.

6. Construir y evaluar un clasificador

Una vez construida la matriz de términos por documento, los datos ya tienen estructura y por lo tanto se puede construir un modelo de clasificación con cualquiera de las técnicas que ya conocéis. Utiliza validación cruzada para los resultados.

7. Representación con n-gramas

Podéis comprobar si modificando el paso de procesamiento de texto para trabajar con 2-gramas (pares de dos palabras) mejoran vuestros resultados. Para ello usamos el operador generate n-grams dentro de Process Documents. Se puede eliminar el paso Stem y stopwords o colocarlo a continuación.

8. Representación de los textos con Word embeddings

Los word embeddings son una forma más avanzada de representar las palabras, que permiten que palabras similares tengan una representación vectorial similar. OPCIONALMENTE podéis explorar la generación de Word embeddings con este corpus y su uso para la clasificación.

9. Sentiment analytics

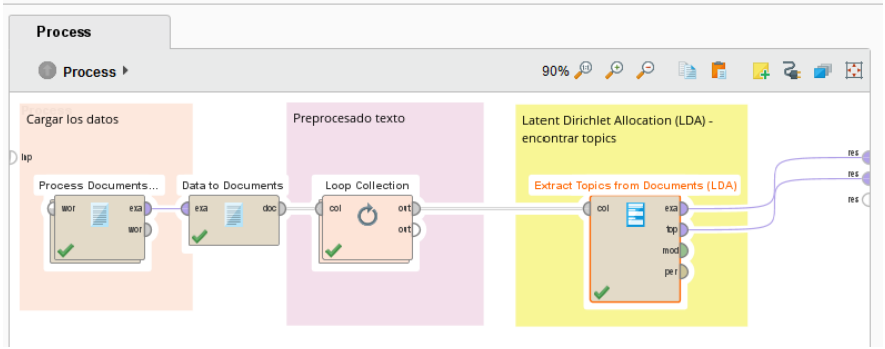
El análisis de sentimiento o sentiment analytics es un área de la minería de texto que tiene como objetivo asignar una valoración (sentimiento) a un texto, algo muy similar a lo que estamos haciendo en esta práctica. OPCIONALMENTE podéis explorar el marktet place de Rapid Miner para probar algún operador de sentiment analytics y ver qué tal se comporta con nuestros ejemplos

Sesión II: Clustering

El objetivo de esta parte de la práctica es analizar nuestros datos con técnicas de agrupamiento para identificar patrones característicos, como podría ser la identificación de los temas de los que se opina. En el contexto de minería de texto, a la identificación de los temas tratados en distintos documentos se le denomina “topic modelling”

Para ello, se puede representar los documentos como vectores como se hizo en la primera parte de la práctica y luego aplicar técnicas de clustering (como K medias).

Alternativamente, se puede recurrir a métodos de topic modeling como Latent Dirichlet Allocation, para lo que hay un operador específico en Rapid Miner. Está disponible en la extensión Operator Toolbox y se llama Extract Topics from Documents (LDA). Este operador tiene como entrada una serie de documentos, no puede trabajar con el texto en formato ejemplos (tabla). Una opción para poder usarlo es añadir un operador “Data to Documents” que convierte el texto disponible en el fichero csv a los documentos que necesitamos. Este operador puede mezclar diferentes campos de texto en un documento, así que hay que elegir el texto como atributo y asignarle un peso uno (si hubiera varios campos de texto, cada uno tendría un peso). A continuación, añade el un operador “Loop Collection”, que itera las acciones que lo componen a lo largo de todos los documentos. Dentro de este operador es donde colocamos ahora los operadores de pre-procesado de texto que sean necesarios (tokenize, filter stopwords). Ya tenemos los datos preparados y conectamos su salida con el operador. En la figura puedes ver un posible proceso.



Explora los resultados modificando diferentes parámetros, como el número de clústeres o la densidad de estos. Valora qué métrica de evaluación es más adecuada para elegir el modelo final. Entre los resultados a entregar deben estar la nube de palabras más representativas de cada uno de los temas.

Evaluación de la práctica

Aspectos a evaluar en la corrección de la práctica:

- Planteamiento y desarrollo del problema: 25%.
- Resultados del problema: 25%.
- Análisis de resultados y conclusiones: 25%.
- Presentación: 15%.
- Contexto de la práctica (Información complementaria sobre el desarrollo de la práctica): 10%.

Debe darse importancia a la presentación para mostrar los resultados, el análisis de los mismos, conclusiones, etc. Es importante tener en cuenta que el contexto de la práctica en la que se incluirá cualquier información relevante conocida por el estudiante, casos similares, noticias al respecto o cualquier otra información de interés y relacionada con la práctica.

Entrega de la práctica

- En esta parte de la práctica se requiere la realización de un documento explicando del conjunto de datos:
  - ✓ La descripción de los datos utilizados.
  - ✓ La descripción detallada de los diferentes filtros aplicados y su justificación.
  - ✓ El proceso de entrenamiento: Los diferentes clasificadores probados, los problemas encontrados, etc. Además, se debe incluir algún clasificador que pueda ser "interpretable".

- ✓ Descripción y análisis detallado de los resultados
- ✓ Conclusiones de la práctica.
- ✓ Contexto de la práctica.

- La práctica deberá realizarse en grupos de 3/4 personas (y entregarse únicamente por uno de los integrantes del grupo).
- La entrega de la práctica será un único documento en el que se detallen y analicen y relacionen las dos partes de la Práctica 2. Dicha entrega será un documento y todos los ficheros que se consideren relevantes para la evaluación de la práctica. Todos los ficheros que se entregan deberán describirse brevemente en un fichero de texto.
- La entrega se realizará por Aula Global con la fecha máxima que figure en el entregable.
- No hay un formato de entrega establecido. Sin embargo, es importante que toda la información se muestre de forma clara y su presentación también es considerada para la nota de la práctica.