**Enunciado Trabajo Final**

**Optativa de Minería de Texto**

Una de las tareas más complejas en los métodos de clasificación supervisados es contar con un conjunto de datos etiquetado que permitan en una primera instancia entrenar un modelo y luego validar el mismo. Dentro del proceso de minería de texto existen muy pocos corpus que pueden ser usados para estos procesos. Por tanto el trabajo final consistirá en crear un conjunto de datos etiquetado que permita formar un modelo de clasificación para reconocer las reseñas positivas o negativas sobre una determinada temática

Se recomienda analizar las siguientes fuentes de datos como una opción para crear el corpus etiquetado:

<https://www.cinemablend.com/reviews/>

<https://timesofindia.indiatimes.com/entertainment/movie-reviews>

Para ejecutar esta actividad se recomienda implementar los siguientes procesos:

1. Recopilar los datos no estructurados. Para este trabajo se usará un conjunto de datos extraídos desde alguna página Web con las revisiones y la etiqueta sobre la valoración de la revisión ya sean positivas (pos) y negativas (neg). Los datos deberán ser grabados en algún repositorio (por ejemplo una hoja de cálculo), en el cual consten dos columnas. La primera columna es el contenido de la revisión y la segunda columna es la etiqueta de valoración. Para obtener la etiqueta de valoración se puede usar el ranking basado en estrellas el cual es un elemento de calificación muy común en distintas páginas web. Por ejemplo tres o más estrellas se pueden etiquetar como positivo y menos de tres estrellas se podría etiquetar como negativo.

En este proceso se recomienda las siguientes actividades obligatoriamente: **i) almacenar un conjunto de datos similar para las reseñas positivas o negativass, y ii) dividir el conjunto de datos de manera aleatoria en dos mitades: el primer 50% de los datos se usará como datos de entrenamiento y el 50% restante se reserva para verificar el rendimiento del algoritmo entrenado.**

Nota: RapidMiner ofrece algunos operadores para las actividades uno y dos respectivamente, por ejemplo el operador Filter Examples y el operador Split Data.

1. Preparación de los Datos. En este proceso la idea es convertir los datos no estructurados en un formato estructurado. Para ello ejecute todas las actividades de pre-procesamiento que considere conveniente. Recuerde que si esta tarea se ejecuta de forma equivocada puede afectar seriamente la precisión del modelo entrenado. El único punto a tener en cuenta aquí es que necesitará algún operador que permita eliminar cualquier "nbsp" (se usa "& nbsp" para representar un espacio de no separación en las páginas webs) y por supuesto los comandos propios de HTML. Por tanto una actividad importante a incluir durante este proceso es **i) eliminar estos caracteres del texto especificado dentro de cada fuente de datos.** La salida de este proceso es un vector de documentos que especifica las palabras que constan en el blog junto con sus pesos.
2. Identificar características clave. El resultado del proceso previo devuelve una tabla estructurada que constan de una fila por cada entrada extraida desde la página web y un número de columnas que indican los diferentes tokens que forman parte del documento y su peso de ocurrencia. No obstante intentar ejecutar un proceso de aprendizaje con tantas variables resulta una tarea muy costosa. Por tanto, en este proceso el objetivo es filtrar aquellas columnas (tokens) mediante algún método de selección de características.

La herramienta RapidMiner ofrece dos operadores para estos fines Weight of Information Gain y Weight by SVM. El primer operador intenta clasificar una característica o atributo por su relevancia con el atributo de etiqueta (en este caso, ranking del tema propuesto) basándose en la relación de ganancia de información y asignará ponderaciones según corresponda. El segundo operador establecerá los coeficientes del hiperplano como pesos de atributo. Identifique si alguno de los operadores puede ser usado en este contexto, entonces: **i) identifique el operador más adecuado para establecer un conjunto de atributos que tengan los pesos más altos y ii) seleccione los atributos con pesos más altos (por ejemplo los 30 mejores) para construir el modelo de entrenamiento. Hacerlo dará como resultado una reducción razonable en los costos de modelado.**

Nota: Rapidminer provee el operador Select by Weights el cual selecciona solo aquellos atributos de un conjunto de entrada cuyos pesos satisfacen el criterio especificado con respecto a los pesos de entrada. Este operador puede servir para la actividad dos de este proceso. No siempre este operador puede ser usado para los fines que persigue la representación del modelo, tomar en cuenta esta situación

1. Construir modelo de entrenamiento. Una vez que se tenga el vector del documento y los pesos de los atributos, se puede experimentar usando diferentes algoritmos de clasificación. Se pide aquí: **i) usar al menos dos algoritmos de entrenamiento (por ejemplo, Naive Bayes y SVM) para identificar cuál proporciona la mejor precisión, y ii) calcular el rendimiento de los algoritmos de entrenamiento utilizados para la clasificación de las valoraciones de los temas como positivos o negativos.**
2. Aplicar el modelo entrenado. Usando el 50% de los datos originales que fueron separados para tareas de validación es necesario validar el modelo entrenado. Para ello es necesario ejecutar las siguientes actividades: **i) convertir estos datos de prueba en un vector de documento. En otras palabras, tenemos que repetir el proceso del paso 2 (sin los operadores de filtrado y división de datos) en el 50% de los datos que se reservaron para las pruebas, y ii) aplicar el modelo entrenado sobre los datos de prueba y verificar su precisión.**

Nota: Recuerde establecer la variable de etiqueta en los datos a validar de forma que se a posible identificar las "Predicciones correctas" y "Predicciones incorrectas" del modelo. En RapidMiner mediante el operador SetRole se puede ejecutar esta tarea

Evidencias

Cada alumno deberá enviar una carpeta en formato empaquetada que contenga los siguientes directorios:

* Procesos: Donde se almacenará los procesos de manera secuencial que permitan ejecutar los numerales del 1 al 5. O los procesos adicionales creados de manera secuencial
* Resultados: En donde se almacenará los resultados intermedios de cada proceso en el formato que se considere conveniente.
* Preguntas: En un documento de texto se deberá incluir respuesta a las siguientes preguntas

1. Cuál es el tamaño del conjunto de datos luego de eliminar las filas que no contienen información pertinente como por ejemplo valores es castellano u otro información irrelevante.
2. Llenar el cuadro con la precisión del modelo entrenado con cada uno de los algoritmos utilizados:

|  |  |  |  |
| --- | --- | --- | --- |
| Algoritmo | Class Recall (M) | Class Recall (F) | Accuracy |
| Naive Bayes |  |  |  |
| SVM |  |  |  |

1. Identificar cuál es el número de predicciones correctas e incorrectas