Práctica 2. Procesamiento de Lenguaje Natural

Fecha de entrega: 3 de mayo de 2020

Esta práctica vamos a realizar procesamiento de lenguaje natural con la aproximación de bolsa de palabras.

Se elaborará una memoria mediante notebooks de jupyter que incluirán los apartados debidamente señalizados, con su código, la salida del mismo cuando corresponda, y las respuestas a las preguntas planteadas.

IMPORTANTE: Usa la opción **random_state** en las funciones que generan las particiones de datos y que implementan los algoritmos de aprendizaje automático para que los resultados del notebook sean reproducibles en todo momento.

Parte 1. Análisis de sentimiento

En esta práctica vamos a usar el fichero "yelp_labelled.txt" que contiene un conjunto de datos de opiniones de restaurantes del recomendador Yelp. Las opiniones está categorizadas como positivas o negativas y hay 500 de cada tipo.

El objetivo es comparar el funcionamiento de distintos clasificadores y entender cómo trabajan estos con texto.

Para ello, vamos a usar distintas opciones a la hora de transformar el texto en bolsa de palabras y observar su efecto en este caso concreto en el que los textos son cortos y escritos con lenguaje informal.

Apartado a)

Configura una partición train-test usando el 75% de los datos para entrenamiento y el 25% restante para test.

Vamos a estudiar varias representaciones de bolsa de palabras, pero todas ellas utilizarán **`countVectorizer`** con el diccionario que se crea a partir de los términos del propio corpus y la lista de palabras vacías (**`stop_words`**) que proporciona sklearn para el inglés. Las 4 posibilidades que estudiaremos surgen de combinar los siguientes 2 parámetros:

- Bolsa de palabras binaria y bolsa de palabras con TF/IDF (parámetro **`binary`**).
- Usando un rango de n-gramas de (1,1) y de (1,2) (parámetro **`ngram_range).`**

Para cada una de esas 4 combinaciones entrenaremos dos clasificadores:

1. Un clasificador naive bayes, eligiendo el más adecuado para cada caso.
2. Un árbol de decisión buscando un valor óptimo para uno de los siguientes parámetros para que se maximice la tasa de aciertos en el conjunto de test: **`max_depth`**, **`min_samples_leaf`** o **`max_leaf_nodes`** (siempre el mismo).

Analiza la tasa de aciertos de entrenamiento y test de los 2 clasificadores en las 4 representaciones de bolsa de palabras (8 configuraciones en total) y contesta a las siguientes preguntas:

- ¿Hay un clasificador que sea superior al otro? ¿por qué crees que sucede?
- Para cada clasificador, ¿tiene un efecto positivo el añadir "complejidad" a la vectorización? Es decir, añadir bigramas y añadir tf-idf. ¿Por qué crees que sucede este efecto positivo o la falta del mismo?

Selecciona el mejor árbol de decisión y obtén las 25 variables con más poder discriminante:

- ¿Predominan más las palabras de uno u otro sentimiento? ¿por qué? ¿hay ruido?

Selecciona el mejor clasificador naive bayes y obtén las 25 variables con más presencia en cada clase:

- ¿Tienen sentido las palabras seleccionadas? ¿hay ruido (palabras sin sentimiento o de sentimiento opuesto al esperado)? ¿por qué crees que suceden estos fenómenos?

Finalmente, explica de manera razonada las conclusiones que has extraído de todo el estudio realizado en este apartado.

Apartado b)

Toma el mejor clasificador Naive Bayes y el mejor árbol de decisión y analiza a fondo sus resultados en el conjunto de test.

1. Analiza la precisión y la exhaustividad de cada clasificador en cada una de las clases (opiniones positivas y negativas).
    - Para cada clasificador, ¿tiene un comportamiento homogéneo a la hora de clasificar ambas clases?
    - ¿Cuáles son las fortalezas y debilidades de cada uno de los clasificadores?
    - ¿Hay algún clasificador que sea mejor que el otro en todo?
    - ¿Coinciden ambos clasificadores a la hora de clasificar mejor una clase que la otra?
2. Pinta los 8 primeros niveles del árbol de decisión y comenta lo que ves.
    - ¿Qué estructura tiene el árbol?
    - ¿Cómo interpretas los niveles que has pintado? ¿tienen algún sentido con respecto a la tasa de aciertos, o la precisión y exhaustividad del clasificador?
    - ¿Hay nodos impuros?
3. Por cada clasificador identifica 2 críticas que hayan sido falsas positivas (malas críticas calificadas como buenas) y 2 críticas que han sido falsas negativas (buenas críticas clasificadas como malas). Analiza tanto su texto original, como el vector de palabras resultante (solamente los términos activos).
    - ¿Por qué crees que ha fallado el clasificador en cada uno de los casos?
    - ¿Se te ocurre alguna idea sobre cómo mejorar el clasificador de sentimiento?

Entrega

La entrega se realizará a través del campus virtual subiendo el notebook de jupyter. En la primera celda del notebook debe aparecer el número de grupo y los nombres completos de sus integrantes. Además, el nombre del archivo será P2GXX, siendo XX el número de grupo.