Naive Bayes

Se desea desarrollar un clasificador de género literario utilizando Naive Bayes Multinomial. Para ello se contará con un catálogo de libros de *epublibre.org*, y un archivo *tar* con libros con fecha de publicación anterior a 1964, en español. Dichos archivos pueden encontrarse en:

- https://web.csc.gob.ar/~jzuloaga/epub/catalog.csv
- https://web.csc.gob.ar/~jzuloaga/epub/compressed.tar

☞: Debido al tamaño del dataset, se recomienda que se verifique si los datos ya fueron descargados. De esta manera se evita descargar múltiples veces los mismos archivos.

(**a**) *Procesamiento del catálogo:*

1. Cargar el catálogo y explorar el contenido de sus columnas. ¿Qué representa cada una?
2. Filtrar las entradas del catálogo, de manera de quedarse solamente con los libros en idioma español.
3. Limitar las entradas del catálogo a las que tenga su correspondiente libro digital.

(**b**) *Definición de las clases:* El catálogo asigna múltiples géneros literarios a cada libro. Se desea elegir, para cada título, una sola clase representativa del género.

1. Analizar la distribución de libros por categoría.
2. Eliminar el género *Otros* por ser una categoría redundante.
3. Proponer y justificar un criterio para elegir un único género cuando un libro tenga varios.
4. Reportar la distribución final de libros por categoría.
5. Separar los libros para definir conjuntos de entrenamiento y testeo utilizando las proporciones 75/25. Fijar la semilla para reproducibilidad utilizando su número de padrón.

(**c**) *Preprocesamiento de texto:* ☞: Debido a que la base de datos es extensa, se recomienda reducir la base de datos mientras programa. Recordar, antes de entregar, volver a adaptar el código para trabajar con la base de datos completa (no se aceptarán entregas que utilicen el dataset reducido).

1. El formato de libros epub es un archivo comprimido *zip* que contiene la metada y estructura del libro, archivos multimedia y archivos *xhtml* con el texto del libro. Extraer el texto de esos archivos. Podrá realizarlo manualmente o valerse de bibliotecas.
2. Aplicar `CountVectorizer` (sklearn) al texto. Ajustar `max_df`, `min_df` y `stop_words` a criterio personal, justificando las decisiones. ☞: El corpus de texto completo es demasiado extenso para la memoria. Se sugiere el uso de *Generators* para procesar el texto plano on-demand.
3. Se desea comprobar el funcionamiento del vectorizador. Para ello, aplicar el vectorizador ya entrenado a dos obras clásicas del catálogo de diferentes géneros (por ejemplo, "Estudio en escarlata" y "Orgullo y prejuicio"). Descartar las palabras presentes en ambos libros. Luego, para cada libro, reportar las 40 palabras más frecuentes. Interpretar los resultados obtenidos.

(**d**) *Multinomial Naive Bayes:*

1. Utilizando solamente numpy y scipy, implementar el clasificador MNB. El mismo debe contener los métodos `fit`, `predict` y `predict_proba`.
2. Reportar el *Accuracy* y el *Macro F1*, tanto para entrenamiento como testeo. ¿Cuál sería la probabilidad de error asociada a un clasificador *dummy* en esta tarea?
3. Se desea efectuar un análisis cualitativo de los errores de clasificación. Para ello, seleccione las 10 obras más populares (de testeo) que hayan sido clasificadas incorrectamente. Interpretar los resultados.