Tópicos Avanzados en Analítica

PROCESAMIENTO DE LENGUAJE NATURAL

Origen de datos: booksummaries.zip

Se recomienda el uso de un notebook para el proyecto.

Responda a las siguientes preguntas haciendo uso de tablas, gráficos y/o texto. Según como considere necesario.

1. Descripción inicial de la córpora (35%):
    a. ¿Qué campos son los más interesantes desde una perspectiva de procesamiento de lenguaje natural? ¿Por qué?
    b. ¿Qué autores son los más comunes en la córpora analizada? ¿Qué porcentaje de los libros tienen un autor asociado?
    c. ¿Qué periodo de tiempo cubre los libros analizados? ¿Cuál es el libro más reciente? ¿más antiguo?
    d. ¿Cuál es el título más largo? ¿más corto?
    e. ¿Qué géneros son los más comunes?
    f. ¿Hay problemas de calidad en la córpora? ¿Qué tipo de problemas?
    g. ¿Algún otro análisis qué nos permita entender mejor la córpora? (Bono por propuestas interesantes)

2. Preprocesamiento de Texto:
    a. ¿Cuántas palabras tiene en total la sección de 'Plot Summary'? ¿Cuál es el promedio por libro? Realice el mismo análisis con palabras únicas. Tip: Recuerde trabajar todo en minúsculas o mayúsculas y eliminar signos de puntuación.
    b. Use la lista de stopwords de NLTK. De las palabras que no fueron filtradas por esta lista, ¿qué palabras se deberían eliminar? Justifique su decisión.
    c. Realice stemming y lematización. ¿En cuánto se reduce el número de palabras únicas en cada caso? ¿Por qué la diferencia?
    d. Con base en el tipo de análisis a realizar en la próxima sección y otras restricciones, ¿preferiría usar stemming o lematización para trabajar con la columna de 'Plot Summary'? Justifique su respuesta.

3. Text Classification:

Ustedes son un equipo que pertenece a ANA-litika Colombia, una empresa de consultoría enfocada en analítica avanzada, y uno de sus clientes, una librería virtual llamada libritospocopiratas.co desea una prueba de concepto para clasificar libros según su género a partir de sus resúmenes.

Para esta prueba de concepto, el cliente desea conocer si es posible clasificar un libro como "Science Fiction". **Desarrolle un algoritmo que a partir de los resúmenes clasifique los**

libros de la córpora según este criterio, adicionalmente determine si un proyecto con un alcance más amplio sería factible.

Este punto se evaluará según los criterios de Entendimiento del objetivo de negocio, preparación de datos, modelamiento analítico y apropiación de conceptos de NLP.

Explique uno a uno los pasos seguidos para este proceso, y muestre los resultados de su algoritmo. Incluya el código en el archivo a entregar.

Tips:

- Haga uso de normalización de texto y de tf-idf para generar los features necesarios para realizar la clasificación.
- Genere una columna dónde cada libro tendrá un valor de 1 o 0 dependiendo de si alguno de sus géneros es "Science Fiction".
- Esta sección será evaluada según el entendimiento que tenga del proceso utilizado para generar la clasificación y no por las métricas de su modelo. Por favor siga el proceso necesario para realizar un buen ejercicio de clasificación. **Recuerde que, a pesar de estar trabajando con texto, finalmente este es un problema de analítica como a los que se ha enfrentado previamente.**