Prueba técnica Talento B - Ciencia de datos
Bancolombia
Mayo 2024
Emmanuel Botero Osorio

Bitácora *BookMatcher*

- **Comprensión del problema**

  Inicialmente se lee detenidamente los detalles de lo que debe cumplir el proyecto, las condiciones y los resultados que esperan por parte del equipo evaluador. Posteriormente se realiza un primer acercamiento a los datos, la estructura, entre otros. Este paso inicialmente se hace con la descripción que ofrece Kaggle propiamente del Dataset, donde podemos acerca de que tipo de dato se tiene en cada campo, el formato en que se encuentra almacenado, además de cómo está cada campo respecto a datos nulos o inválidos. Esta parte es muy importante ya que cuando se esté realizando el EDA puede recortar un poco el camino a seguir con los datos. En esta etapa también se pueden identificar algunos campos a los que se les debe realizar alguna transformación o 'limpieza', como es el caso particular del campo de la tabla booksdata.csv 'categories', esta a pesar de ser un string tiene una estructura '[categoria]', la cual más adelante se deberá organizar.

- **Definir alcance del proyecto** Luego de tener este primer acercamiento a los datos y conocer el problema que se quiere resolver, es fundamental cuál será el principal objetivo que se quiere lograr, la manera en particular como se va a abordar el problema. Para esta fase inicial decidí solo tener en cuenta las variables que nos dan una descripción dentro de una categorización que se tiene de los libros, como por ejemplo: autor, categoría. Pues dada la limitante del tiempo no decidí abordar los campos a los cuáles se les debe hace un procesamiento de lenguaje, igualmente se especifica en futuros pasos que se desea implementar al sistema de recomendación estos análisis para ver como se comporta el modelo, mirar que pasa en términos de eficiencia, recursos computacionales y mejora en la recomendación.

  Teniendo esto en cuenta, se establece como meta un sistema de recomendación que se base en la categorización de los libros a través de sus diferentes características.

- **Configurar el entorno de trabajo (entorno virtual y repositorio)** Como en todo proyecto, ya sea colaborativo o en grupo, es fundamental llevar un manejo de versiones además de establecer entornos de ejecución, donde se evite tener problemas por inconsistencias en las versiones de las diferentes librerías y dependencias que necesita el proyecto para poder funcionar correctamente. En esta etapa se establece la estructura del repositorio y se crea el entorno virtual. La estructura del proyecto sería la siguiente:

```
Data/
Docs/
 Bitacora.pdf
src/
  Notebooks/
       EDA.ipynb
       Pruebas_modelos.ipynb
  model.py
  requirements.txt
README.md
```

- **EDA**

  Todo este proceso se desarrolla dentro del archivo 'EDA.ipynb' donde inicialmente se cargan las librerias necesarias para posteriormente cargar los dos archivos csv que tiene el dataset. Luego de cargar estos dos archivos se empieza a mirar su estructura, realizar un análisis de cómo están los datos, esto principalmente con algunas funciones de la librería pandas de Python. Con esto podemos ver la relación de valores nulos por cada uno de los campos que se está estudiando. Inicialmente se descartan algunos campos como *'image , previewLink , infoLink* pues de entrada no voy a considerar la información que estos almacenan, no son relevantes para el análisis, simplemente son vínculos a páginas de internet.

  Se destaca que dentro de la tabla *bookdata* la variable con mayor número de valores nulos es la de *ratingsCount* al rededor de un 80% no tiene este valor asignado, para estas situaciones se pueden abordar de varias maneras, descartar los nulos, mirar si se pueden rellenar (dependiento del tipo de dato), crear una nueva columna que nos hable sobre la disponibilidad o no de este dato.

  En esta etapa se realiza además una limpieza a las columnas *authors, categories* pues tienen una estructura inadecuada, esta limpieza se hace a través de un función que reemplaza los caracteres que no se desean, esto dado que es constante en toda la columna, tratamiento similar se da a la columna *año* pues al tener una variedad de formatos y algunos con información faltante se decide dejar solamente el año, que está presente en casi todas las entradas, esto además de que en general el día y el mes pueden ser un poco irrelevantes para nuestro fin.

  Proceso similar es llevado a cabo con la variable *review/helpfulness* pues a pesar de que se interpreta como un valor numérico esta está almacenada como dato tipo string, donde está la particularidad que se quiere expresar una relación tipo 3/4 , 5/5 , 7/7 pero sin ninguna consistencia o escala establecida, la decisión que se toma entonces es realizar un split por el divisor '/' y luego dividir los dos valores resultantes, en caso tal de que no se pueda realizar la operación, se asigna 0.

  Finalmente se realiza un merge de los dos dataframe sobre el campo 'Titulo' y se guarda en un nuevo .csv Esto dado el peso de los archivos es bastante grande.

- **Modelos a probar**

  En la parte del modelo se empezó a implementar un modelo basado en el contenido, donde se trata de en base a las características de los libros poder hacer la sugerencia.

- **Para mejorar**

  Para mejorar la implementación hay varios aspectos que por temas de tiempo no pude alcanzar a cubrir,

  En general, podría implementarse diferentes modelos, ver como son los resultados, como mencioné anteriormente se puede implemente NLP para ver como puede mejorar el modelo de recomendación, llevar seguimiento a los resultados además de eficiencia del código, me parece fundamental implementar el tema del "logging" y/o manejo de errores, a la hora de entornos de producción es muy importante ver que parte del proceso está fallando, que funciones o métodos están fallando, esto de la mano teniendo una programación modular (clases) puede mantenerse en buenos términos. Respecto al volumen de los datos, se podría implementar Spark, que da un mejor manejo de grandes volúmenes de datos. En general variar un poco las variables que se tienen en cuenta para el modelo, hacer un seguimiento a las diferentes métricas para buscar escoger el mejor resultado.

  De ante mano pido disculpas, ocasionalmente suelo trabajar con eventos y este fin de semana no tuve mucho tiempo libre, por eso la hora de mandar todo y las partes faltantes en el poryecto, sobre todo en la parte de modelos, de igual forma muchas gracias.