**Chương 2: Tiền xử lý dữ liệu**

Dữ liệu chưa được xử lý: (gắn link google sheet)

**2.1. Phân tích dữ liệu**

Dữ liệu dataset chưa qua xử lý: có 10000 bản ghi

Các thuộc tính của bộ dữ liệu:

Book

Author

Description

Genres

Avg\_rating

Num\_rating:

Url:

\* Nguồn: (link nguồn dữ liệu)

- Phân tích dữ liệu

\* Dữ liệu thiếu (ở cột description)

\* Dữ liệu trùng lặp(ko có)

\* Dữ liệu nhiễu (cột Genres)

**2.2. Các bước xử lý dữ liệu(gg sheet)**

Tạo bộ lọc

\* B1: xử lý trùng lặp( ko có => ko cần xử lý)

\*B2: xử lý dữ liệu nhiễu(cột Genres, Book có một vài dữ liệu chứa dấu #, bỏ các cột dữ liệu ko sử dụng như: num\_rating, decripstion)

- ảnh và cách làm trên excel và vẽ biểu đồ

- ảnh code python và vẽ biểu đồ

**2.3. Chuẩn bị dữ liệu cho giai đoạn sau**

- gắn link google sheet kết quả sau khi xử lý

- xuất file csv

**Chương 3**

Thuật toán K-means

Thuật toán FP-Growth

- là gì

- tính chất

- các bước thực hiện ( bằng chữ )

- ưu điểm

- nhược điểm

- Áp dụng thuật toán và hệ thống bán sách

- tính confidence trung bình