File Preprocessing:

# Library:

# Preprocessing:

## Cột Author:

Vì 1 cuốn sách có thể lên đến 7 tác giả => cần processing

**Xóa những kí tự dấu linh tinh** => đồng nhất = dấu phẩy

**Thêm cột số lượng tác giả của từng cuốn sách (phục vụ model)**

## Cột Title và synopsis:

#### Dịch:

Ngôn ngữ của sypnosis chưa phải là tiếng anh hết (có tiếng Ấn)

* Gặp tiếng anh thì giữ nguyên, ngược lại thì dịch **(googletrans)**

#### Preprocessing ngôn ngữ tự nhiên:

**Strip:** xóa khoảng trắng dư khi crawl , gather data từ html

**Contraction:** xử lý viết tắt trong tiếng Anh (doesn’t, isn’t, i’ll,…)

**Punctuation:** loại bỏ kí tự dấu

**Stopword:** loại bỏ những từ cực phổ biến trong TA không mang lại ý nghĩa chính của câu (many, the, of, …)

**Stemming và Lemmatization:**  Là phương pháp đưa 1 chữ về dạng gốc nhằm mục đích rút trích được ý nghĩa chính của nó **(walking, walks, walked -> walk -> ta được ý nghĩa chính là đi bộ)**

* **Stemming**: Phương pháp này **nhanh, nhưng đơn giản**, xu hướng bỏ hoặc biến đổi các kí tự cuối. Không apply được các case đặc biệt như mouse -> mice hoặc foot -> feet….
* **Lemmatization:** Phức tạp hơn chạy lâu hơn, vì có database từ điển để tra -> Giải quyết được những case phức tạp

## Topics:

Dùng **LDA** để nhóm từng book vào các topic dựa trên các từ có trong **Sypnosis**

Có 25 topics

## Cột Edition:

Gộp nhiều thông tin như: **Quốc gia – Tòa soạn – ngày – tháng - năm =>** cần processing tạo cột mới

## Fill missing values:

#### Năm xuất bản:

Dùng trung vị để gán

#### Tháng xuất bản:

**Random có trọng số** để gán

**Trọng số dựa trên value\_counts** của từng tháng, tháng nào xuất hiện nhiều trọng số cao => Bằng cách này giữ nguyên được phân phối các tháng.

# Cyclical Encoding for Month:

Chưa rõ làm gì

Có liên quan đến model