NlP Project Summery Of Books

Saba Razi

June 2023

Git Repository

Git repository address: `https://github.com/sabarz/NLP_PROJECT_summary`

Project Overview

In this project, we used the website `www.goodreads.com` for crawling summaries of three genres: crime, romance, and psychology. The best parts of the website for crawling were identified for each genre. Here are the lists of books for each genre:

- Crime books: `https://www.goodreads.com/list/show/11.Best_Crime_Mystery_Books?page=1`
- Romance books: `https://www.goodreads.com/list/show/10762.Best_Book_Boyfriends`
- Psychology books:
  - `https://www.goodreads.com/list/show/41846.Inspiring_Books`
  - `https://www.goodreads.com/list/show/691.Best_Self_Help_Books`
  - `https://www.goodreads.com/list/show/86863.Life_Transformation_Books`

We used the scrapy library for crawling. In the "spiders" folder, the "books.py" file was used for crawling the raw data. You can run it with the command `chmod u+x raw.sh`. The raw data is a JSON file, where each genre has a list of summaries.

The "cleaning.py" file was used to clean the data. You can run it with the command `chmod u+x cleaning.sh`. The cleaned data is a JSON file, where each genre has a list of cleaned summaries. The cleaning process involved removing numbers, unicode, undefined words with numbers, and most punctuations (except the period). Additionally, any summaries that were labeled as "None" were deleted.

The "sentenceBroken.py" file was used for breaking the data into sentences. You can run it with the command `chmod u+x sentenceBroken.sh`. The "sentenceBroken" data is a JSON file that includes a key, genre, summary, and a list of sentences. Sentences were created using the `split()` function by splitting at periods.

The "wordBroken.py" file was used for breaking the data into words. You can run it with the command `chmod u+x wordBroken.sh`. The "wordBroken" data is a JSON file that includes a key, genre, summary, and a list of words. Words were tokenized using the `split()` function by splitting at spaces.

Huggingface Repository

Link to the Huggingface repository dataset: `https://huggingface.co/datasets/sabarzii/NLP_summery_Books`

stats:

	crime	romance	psychology
number of summeries raw data:	6900	6747	4528
number of summeries clean data	2679	2679	2679
number of sentences	777488	1131130	686036
number of words	13680017	18325126	12040883
number of unique words	702759	811452	624185

	number
number of unique common words of romance and crime	63
number of unique common words of romance and psychology	63
number of unique common words of sychology and crime	63
number of unique uncommon words of romance and crime	1514085
number of unique uncommon words of romance and psychology	1514085
number of unique uncommon words of sychology and crime	1326818