PERSONAL REPORT

DATS 6312: Natural Language Processing

Liang Gao

2024-5-1

1 Introduction

Our project aimed to recommend books automatically with artificial intelligence. The dataset we used is Amazon book review [1]. Each of our three team members developed a model with natural language processing techniques for this project. Furthermore, all team members contributed significantly to the group report, presentation, and demo. Our collaboration was highly productive and immensely enjoyable, fostering a strong sense of camaraderie among us.

2 Individual work description

I mainly developed the KNN model.

Considering the limitations of our GPU capacity, the original size of our dataset was too large, so we only utilized a subset of it. We removed users with fewer than 10 reviews, dropped rows with null values, and dropped categories with a count of less than 5000 and greater than 20,000. The number of raw data observations is 3 million. After filtering, the dataset for later modeling has 58,199 observations.

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used for both classification and regression tasks. KNN works by finding the nearest neighbors to a query data point and then basing its prediction on the properties of these neighbors. KNN calculates the distance between points using metrics such as Euclidean, Manhattan, Minkowski, or Hamming distance to determine which known instances are closest to the new one. In classification, KNN assigns a class to the query point based on the majority class among its nearest neighbors. In this task, the KNN algorithm calculates the distance between a user's new review text and the existing reviews in our dataset. It then identifies and outputs the indices of the reviews closest to the input. We can effectively recommend books from our dataset that align with the user's preferences using these indices.

There are two steps in this part of the experiment. First, we used a pre-trained BertForSequence-Classification model to predict the category that the user potentially likes based on the input review. Second, we filtered the data frame to include only those rows where the 'categories' column matches the predicted category in the first step. Then, KNN will help to find the nearest books using the same input review. The details will be explained as follows.

One of the critical architectures in this part is BERT (Bidirectional Encoder Representations from Transformers), a groundbreaking model in the field of natural language processing (NLP) developed by researchers at Google AI. Introduced in their 2018 paper, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," BERT has significantly advanced the performance of various NLP tasks. BERT was designed to address the limitations of previous models that processed words in a sentence sequentially, either from left to right or right to left. BERT, by contrast, reads the entire sequence of words at once, which allows it to learn the context of a word based on all of its surroundings (left and right of the word).

For tokenization, a pre-trained BertTokenizer([2]) was used to tokenize the 'review_text', replacing classical tokenization methods such as NLTK and Spacy tokenization. The BertTokenizer is a crucial component of the BERT architecture, which has been instrumental in advancing stateoftheart NLP. The BertTokenizer is designed to effectively preprocess text data before it is fed into a BERT model. We can directly get the vector representation of each text review with BertTokenizer, which is content-based, instead of TFIDF, which is frequency-based.

The dataset was divided into three subsets: 80% for training, 10% for validation, and 10% for testing. We have six book categories: Religion(18,894), Business & Economics(10,813), Young Adult Fiction(10,582), Social Science(6,624), Philosophy(6,131), and Science(5,155).

A pretrained BertForSequenceClassification([2]) model was applied to solve this book category classification problem. We encoded the book categories with LabelEncoder and set the format of tokenized

reviews and labels for Pytorch. We only need to fine-tune (change hyperparameters) our model since it is pretrained. A validation set will test if the model is overfitted while training.

After getting the predicted category, we can filter the dataset and apply KNN to recommend books. There are many duplicate books in our dataset. Many books have more than one review. Thus, we added some post-processing to our predicted books to remove duplicates. We dropped rows with the same review summary, utilized the 're' package to delete brackets and inside content in the title, dropped duplicate books with the same title, and finally ranked the books based on review score and output the top 6 as our recommendation.

The assessment is similar to what we did in the SVD model. The training dataset only contains 6 categories. Thus, we used ChatGPT to generate reviews for books from the 6 categories (Table 1). Then, we manually assessed the similarities of recommended books to the review prompt. [3].

Category	Review
Religion	The Power of Myth" is a fascinating exploration of the universal themes woven into religious and cultural narratives. Joseph Campbell's insightful discussions with Bill Moyers offer a thought-provoking journey into the depths of human consciousness and the collective imagination. This book provides a profound understanding of the underlying structures of myths and their relevance to our modern lives, making it an enlightening read for anyone seeking wisdom and insight into the human experience.
Business & Economics	The author explores the two systems of thinking that govern our decision-making processes, shedding light on the biases and heuristics that often lead us astray in economic and business contexts. Readers praise the book for its engaging style and eye-opening insights, which challenge conventional wisdom and provide valuable lessons for navigating the complexities of the modern world.
Young Adult Fiction	This heart-wrenching tale follows the lives of two teenagers as they navigate love, loss, and the complexities of living with cancer. The book's authentic portrayal of the character's emotions, coupled with Green's poignant writing style, creates a deeply moving narrative that resonates with readers long after they've turned the final page.
Social Science	This insightful work delves into the complexities of human behavior, drawing upon psychology, sociology, and neuroscience to explore what drives our actions and relationships. Through engaging narratives and compelling research, he illuminates the subtle influences that shape our lives, offering a deeper understanding of the social forces at play in our everyday interactions.
Philosophy	The narrative seamlessly weaves together philosophical concepts with an engaging storyline, making complex ideas accessible and thought-provoking. It's a must-read for both beginners and seasoned philosophers alike, offering a delightful journey through the wonders of human thought.
Science	Sapiens is an eye-opening exploration of humanity's journey from ancient hunter-gatherer societies to the technologically advanced civilization we inhabit today. Harari's ability to distill complex ideas into accessible narratives makes this book informative and engaging. It's a must-read for anyone curious about the origins of our species and the forces that have shaped our societies

Table 1: Book reviews generated by ChatGPT for 6 categories.

3 Results

The results for the development and testing of each model are described below.

For the BertClassification model, the metric for the test dataset is F1-weighted at 0.846, F1-macro at 0.834, F1-mico at 0.843, and Cohen-kappa-score at 0.804.

Religion	3/6
Business & Economics	5/6
Young Adult Fiction	5/6
Social Science	4/6
Philosophy	6/6
Science	1/6

Table 2: Proportion of book recommendations similar to the review prompt regarding categories.

The artificial assessment results of the KNN model are in table 2. Results are subjective and represent personal views only.

4 Conclusions

Overall, my work is mainly about developing this BertClassification plus KNN model.

Instead of using KNN directly, identifying the book category first can narrow the book's range, which I personally think can increase the KNN performance. One critical limitation of the KNN model is it runs BertTokenizer every time, which requires high CPU computational power and is time-consuming.

References

[1] M. Bekheet, "Amazon books reviews. kaggle." 2022. [Online]. Available: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

[2] H. Face, "Transformers: State-of-the-art Natural Language Processing for PyTorch and TensorFlow," 2024. [Online]. Available: https://github.com/huggingface/transformers

[3] OpenAI, "Chatgpt, version 3.5," OpenAI, 2023, retrieved from. [Online]. Available: https://chat.openai.com