CS410 Text Information Systems - Fall 2022

Project proposal

Team SALE

- Eunbi Go - eunbigo2 (Team captain)
- Akarsh Bhagavath - akarshb2
- Lingfei Yang - lyang26
- Sruthi Jayanti - sjayan3

Motivation

We have chosen the topic of intelligent browsing for this project. We plan to build a browser extension that takes in book title keywords as user input and outputs a sentiment score for reviews about that book after performing sentiment analysis on a book reviews dataset. The extension will also serve as a recommendation system by recommending books similar to the book title entered based on sentiment scores.

This project solves the problem of having to read through multiple book reviews to gauge how good a book is. This can be a tedious experience, and our proposed project aims to solve this problem intelligently. Similarly, a recommendation system for books pushes useful information to the user, which minimizes the need for users to go find similar books on their own.

This topic and project relate to the theme of the text information systems because the Chrome extension we propose involves sentiment analysis on a textual dataset, as well as pushes information to the user based on a recommendation system model. These are topics directly relevant to the course.

Datasets, algorithms or techniques

Dataset

We plan to use content-based filtering techniques on the books Data set. We will begin with this dataset as it has a descriptive review of the books on amazon.

- https://jmcauley.ucsd.edu/data/amazon/ (Under books 5-core)

Algorithms/techniques

For the recommender system, we will look into using the Lenskit framework for python. If lenskit does not work for our use case, we will look into other approaches in python. We will evaluate our model using standard evaluation techniques (precision, recall, etc). For the frontend, we will need to develop a chrome extension using javascript to create. We will need to develop an approach to wire the extension with the backend model.

How will you demonstrate that your approach will work as expected?

Our application will create a Chrome extension (limited to the English language) that can have the following main features and approaches.

First, users can type in a book title/keywords. Also, the alternative is that sentences or terms can be copied from the website where the extension plug-in is. We will use text retrieval techniques (e.g. BM25) to find the top 5 books that match the book title/keywords provided by the user based on the book reviews dataset and/or book descriptions dataset.

Meanwhile, for each book in the list, the extension will search the database to obtain corresponding book reviews. It will then determine the sentiment score for the book based on a 1-5 scale using sentiment analysis, where 1 represents negative sentiment and 5 represents positive. The main approaches can be Naive Bayes or VADER Models, etc.

Finally, we will provide suggested book names to the user based on the content of the book descriptions in order of relevance. The methods can be content-based filtering, collaborative filtering or hybrid filtering.

Programming language(s)

- Backend: Python
- Chrome extension: Javascript

Estimated workload

We are a team of 4 and a total of 88 hours of work is expected for this project. The expected required time is as follows.

Task	**Time required**
Pick a topic & Write a proposal	5 hours
Work on the project	-
Build and clean the dataset related to the book review	10 hours
Build sentiment analysis algorithm	15 hours
Build recommendation system	15 hours
Build Chrome extension - Backend	15 hours

Build Chrome extension - Frontend	10 hours
Progress report	3 hours
Software code submission with documentation	5 hours
Software usage tutorial presentation	10 hours
Total	**88 hours**