Team NLPro Project Proposal

Members: Ethan Howes (edhowes2@illinois.edu), NetID: edhowes2
Project Leader/Captain: Ethan Howes

Summary

My task is to create a Chrome browser extension to augment product reviews on the Target website. This extension will provide an intelligent search function over reviews for a product using an inverted index and BM25 and a sentiment score for the reviews of a product using the Natural Language Toolkit (NLTK). Currently, the Target website does not allow users to search within the reviews of a product. A user may wonder if other customers had a particular problem with a product. Intelligent search using BM25 will allow users to search reviews for more nuanced questions about a product than simply using the find-on-page functionality within Chrome to search reviews. Additionally, the sentiment score computed using the NLTK will give users a second opinion on how users feel about the product instead of relying solely on the user's star ratings.

This project relates to the intelligent browsing theme as it provides a more intelligent way to query product reviews and provide sentiment scores for products via a browser extension. It relates to our class in that it incorporates fundamental text retrieval techniques covered in class such as BM25 and inverted index as well as a key text mining concept in sentiment analysis.

Datasets, Algorithms, and Techniques

The BM25 algorithm will be used to query product reviews. Meanwhile, an inverted index will be constructed to make the query process fast. BM25 and the inverted index will be implemented using the MeTA toolkit used in class. MeTA will also be used for the preprocessing of reviews including tokenization, removal of stop words, and stemming. Additionally, the Natural Language Toolkit will be used for sentiment analysis of each product review. To aggregate the sentiment scores an average of the scores will be calculated using the product review length to weight each score. Longer reviews are typically more detailed and may have a more telling sentiment score. This will be similar to document length normalization.

Demonstration of Approach

In the end, with this extension, a user should be able to navigate to a product on the Target website, open the extension, query reviews of the product within the extension, and view a sentiment score of the product based on the sentiment of reviews. A video demo will be recorded displaying these steps.

Frameworks and Languages

The primary language used for this project will be Python. However, some aspects of the Chrome extension may need to be constructed with JavaScript, CSS, and HTML. The MeTA and NLTK frameworks will be used for this project's NLP tasks. Finally, the RedCircle API will be used to retrieve product reviews on the Target website.

Tasks and Time Cost

- Learn Chrome extension basics: 3 hours
- Create my extension UI (as the rest of the features are developed): 4 hours
- Learn RedCircle API setup and basics: 2 hours
- Using RedCircle API to collect Target reviews for a given page: 2 hours
- Preprocess and parse reviews: 1 hour
- Implement BM25 and inverted index to search reviews: 3 hours
- Implement NLTK sentiment analysis on each review: 1 hour
- Intelligently aggregate the sentiment scores of each review: 3 hours
- Display comparison between star rating and sentiment score in extension UI: 1 hour
- General testing and tweaking BM25 parameters: 1+ hours

Total of at least 21 hours.