# MSDS 453 Section 56

# Assignment 3: Clustering and Topic Modeling

# Benjamin Xiao

# Abstract

Adsf

# Introduction

The current pandemic has forced consumers to buy more products and services online. The increased activity has prompted users to leave more reviews and ask more questions (Niles, Barron, 2020). As e-commerce sales continue growing (Ali, 2020) and businesses continue developing e-commerce experiences, customer reviews become a valuable source of feedback and data. Reviews are typically coupled with a numerical rating from 1 to 5 with a lower number associating with a worse review. These reviews encapsulate the top down customer experience that include the product’s relative effectiveness to solve users’ problems, the physical delivery experience, relative expectations. Depending on website description and other reviews, potential customers have a more complete picture of a product’s effectiveness for their respective problems.

For business stakeholders, text reviews provide more granular information about what may or may not be working with their products. However, the gigantic volume of reviews and variation among review lengths make reading every review tedious and time-consuming. Natural language processing helps mitigate this problem by quickly going through each review. A sentiment analysis can flag particular reviews to see how many reviews seem problematic.

A subsequent step is to employ natural language understanding analysis. Among the flagged reviews, cluster analysis of individual reviews and associated words can provide us with meaningful insights. This can be abstracted into hierarchies such as among car brands, body types, and price points. Perhaps the good reviews highlight a feature of a product that performs better than its competition. The same could also highlight words or phrases associated with the opposite end of the spectrum, signaling areas for improvement. Deploying NLP processes can provide a more complete picture of review sentiment in a short amount of time.

# Literature Review

Amazon’s product reviews page is a common data source for textual reviews that also provides predetermined categorizations. It provides labels with every review in the form of 5-starred ratings which makes it ideal for sentiment analysis (Shafaee, Issa, Agne, Baumann, & Dengel, 2014). Vectorizing words can help determine how documents or reviews differ from one another through word frequency (Kim, Kim, & Cho, 2017).

Furthermore, research has been conducted to use entire sentences instead of processing words in order to maintain language structure and semantics (Büschken & Allenby, 2016). A mix of sentence and word clustering was tried to classify news topics (Yow & Tan, 2013).

Finally, the final topic that must be addressed is clustering algorithms. Ratings are rarely balanced and normally distributed. This may then require researchers vary their clustering algorithms to accommodate for unbalanced data (Qian & Saligrama, 2014).

# Methods

## Data Description and Processing

The dataset used is a set of scraped car reviews from Edmunds.com. It contains user information, the car manufacturer, car model, type, a text review, and a rating from 1-5 stars. To continue using text analysis, documents without review ratings and containing 50 total words in its text review are removed. Rows which were not formatted correctly into the comma separated formats were also removed.

# Results

asdf

# Conclusions

Asdf

# References

asdff