

Objective

Build a data visualization tool for the consumer product review dataset in order to explore various aspects related to the dataset.

Introduction

For this tool, I have used customer product review data from CDs_and_Vinyl_5.csv.gz. I assume, product managers and/or marketing leaders are my target audience. Through this visualization tool I am trying to answer the following questions:

1. What is the average rating of a product (main analysed product) ?
2. What are the top products reviewed immediately before and after this product reviewed, and how they are rated overall.
3. How do the same set customers rate those products (those are reviewed immediately before and after) ?
4. Do people rate a product much higher or lower than their average ratings.

Tool Overview

Input : Product ID and number of connected products wanted to see.

Output : Tree graph with the intended product at the root of the tree.

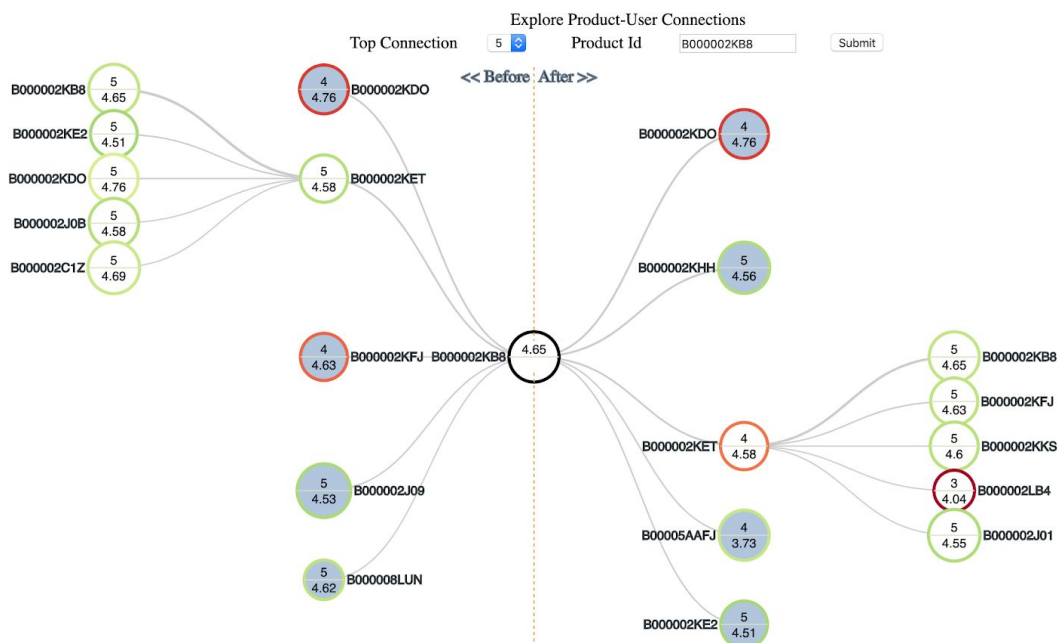
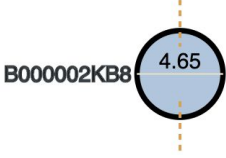
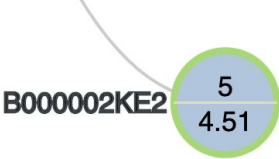





Figure 1

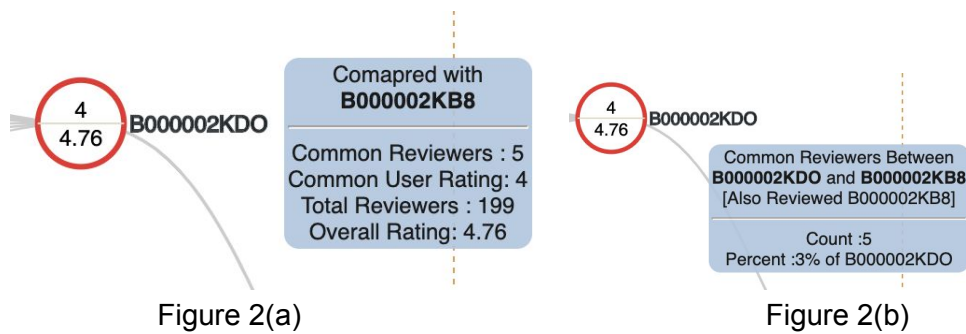
Web Access : <https://cs765-dc2.herokuapp.com/>

Each node represents a product and the product name is displayed beside it. The intended product plays a central role here and it's placed at the center of the visualization. The products got reviewed immediately before this product (focused product) got reviewed are placed on the left side of it and similarly, the products got reviewed immediately after this

product got reviewed are placed on the right side of it. The thickness of the edges shows number of the customers on that path. In figure 1, five nodes (products : B000002KDO, B000002KET, B000002KFJ, B000002J09, and B000008LUN) left of the central line shows; they are reviewed immediately before the product B000002KB8 got reviewed. Similarly there are five nodes (products : B000002KDO, B000002KHH, B000002KET, B00005AAFJ, and B000002KE2) right of the central lines shows, they got reviewed immediately after the product B000002KB8 got reviewed.

	<ul style="list-style-type: none"> • The central node is the product that queried. The edge color of this node is black. • The number inside this node shows the average rating received for this product.
	<p>Other product nodes (right or left of the central line):</p> <ul style="list-style-type: none"> • The number at the lower half of node shows the average rating for the product received from all customers. • The number at the upper half of the node shows the average rating for the product received from these set of users (those reviewed products in that chronological order)
	<p>Filled node means there are children nodes exist and not displayed. On mouse click of this node, children nodes are displayed. Empty node suggests either there are no child node or they are explored (displayed).</p>
	<p>Colors on node boundary shows relative difference* between average ratings by all reviewers and selected set of users (those reviewed products in that chronological order). The colors are from continuous scale. The more negative relative difference it appears more red, similarly more positive relative difference it appears more green.</p>
	<p>Size of the node shows, number of reviewers reviewed the product.</p>

*relative difference:
$$\frac{(Average\ Rating)_{selected\ Users} - (Average\ Rating)_{Overall}}{(Average\ Rating)_{selected\ Users}}$$



More attributes of a product are displayed with mouseover on nodes and edges. Figure 2(a) shows mouseover features on a node. The popup shows the number of reviewers associated for the ratings in addition to the average ratings (also shown inside the nodes). Figure 2(b) shows the number of users reviewed in that chronological order and what percentage of the total reviewers of that product.



To further analyse any product on the visualization, there are features provided on right click. Figure 3(a) and 3(b). With 'Explore product' option, the product is queried and the visualization updated for that product. Using 'Explore Rating' option, we can see the rating given by it's reviewers and compared with their average ratings in an additional chart (Figure 4).

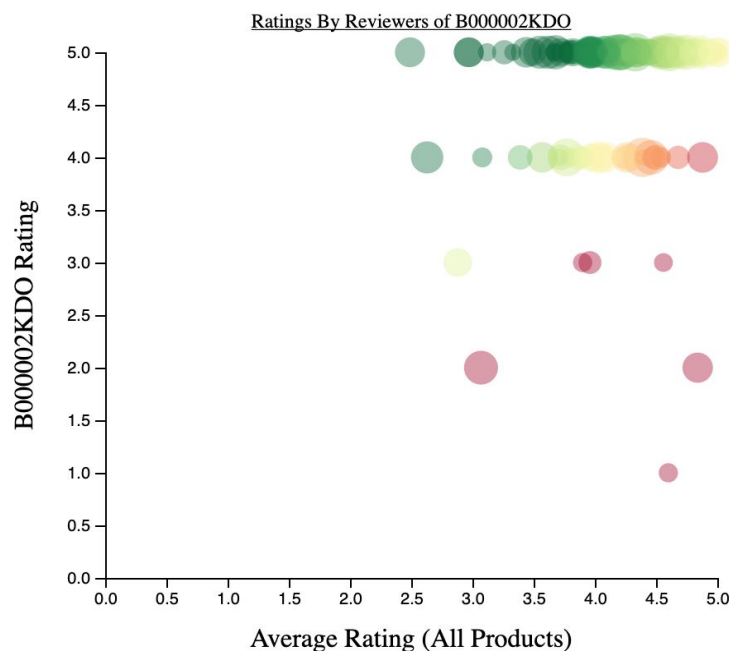


Figure 4

In figure 4, every dot represents a reviewer. X-axis shows average rating i.e. averaged over all the products that a user has reviewed. Y-axis shows rating given for the interested product. The size represents the number of products the reviewer has reviewed. The color shows relative difference** of the ratings. Details are provided on mouse over as well.

$$\text{**relative difference} = \frac{\text{Rating given for the product} - \text{Average of ratings given in all reviews}}{\text{Rating given for the product}}$$

Descriptions of the Designs and Their Intents

The overall intention of this tool is to explore a product and to understand its reviewers' choices before or after reviewing the current product. This tool serves the purpose well.

A products may be connected to other products if they are reviewed by the same customer. So, just to understand these connections, graph representation would be a nice fit. However, with graph representation we would lose the sequential effect if there is any. To answer the questions that we are looking for, we need to preserve the sequence of reviews. This makes tree design a natural choice than a graph design.

To quickly and correctly see the average ratings received for a product, they are written on the nodes. More often popular products get more number of reviews. Hence, size encoding is used for number of reviews. Moreover, our focus is to see to what extent the selected set of customers like/dislike other products. For this reason, divergent color encoding is used to encode the relative difference of ratings. Additional features of the tool such as a) on click explores the downstream/upstream products, b) mouse over provides more information, and c) explore details with right-click satisfies 'details on demand' design principle.

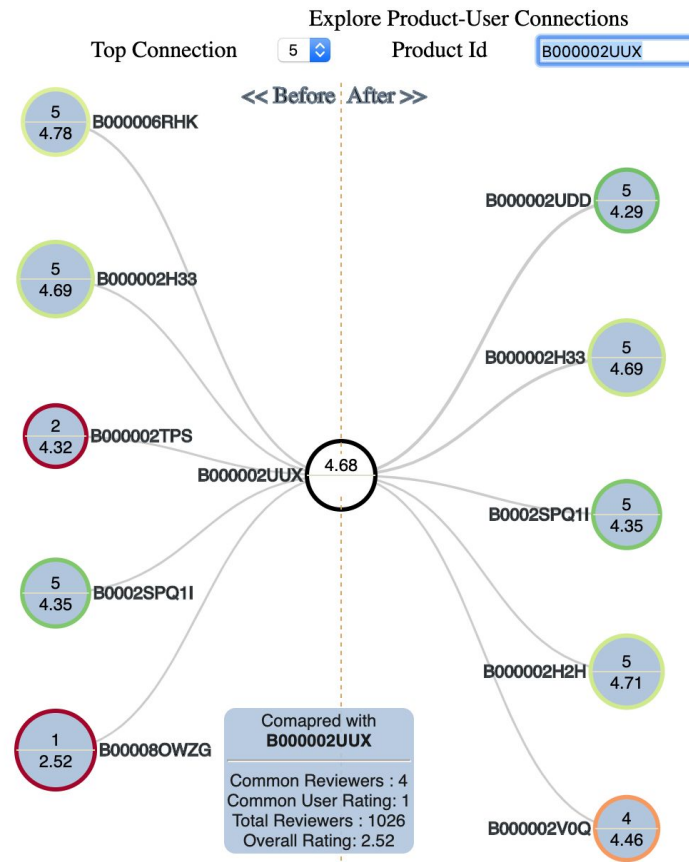


Figure 5

Use Case Evaluation

In this example, I am explaining how we can answer the questions using this tool.

- We are interested to explore product B000002UUX and wanted to see 5 most reviewed products before and after B000002UUX got reviewed.
- 'Top connections' 5 was selected and product Id B000002UUX was searched. As a result we saw Figure 5.
- From figure 5, we can see products reviewed before B000002UUX got reviewed are placed left of it and products reviewed after B000002UUX got reviewed are placed right of it.
- The average rating of B000002UUX is 4.68 (shown on the product node). Similarly average ratings for other products are mentioned on the corresponding nodes.

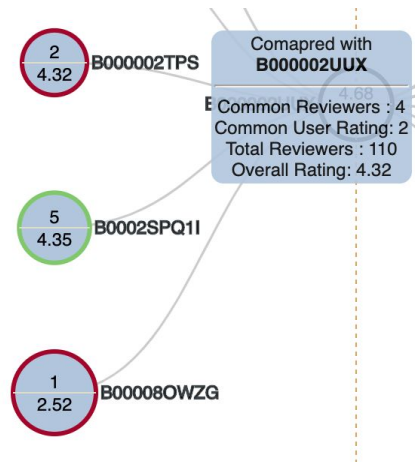


Figure 6

- Product B000002TPS has received much lower average rating by 4 customers who reviewed B000002UUX. B000002TPS received average rating 2, as compared to its overall average 4.32
- To further understand reviews for B000002TPS, we can see the details by using 'Explore Ratings' option by right clicking on the product.

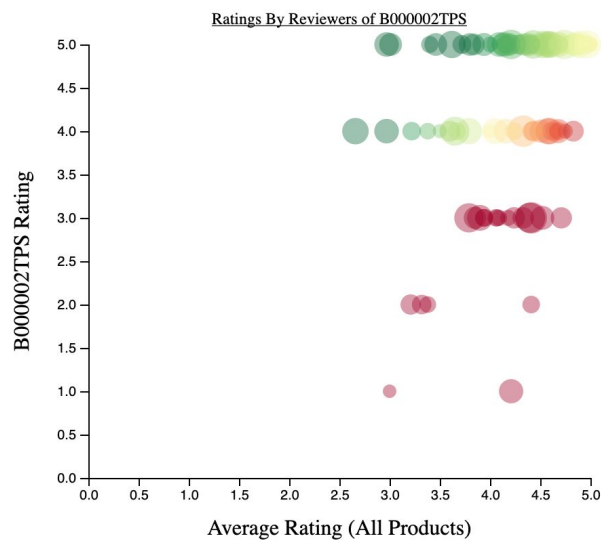
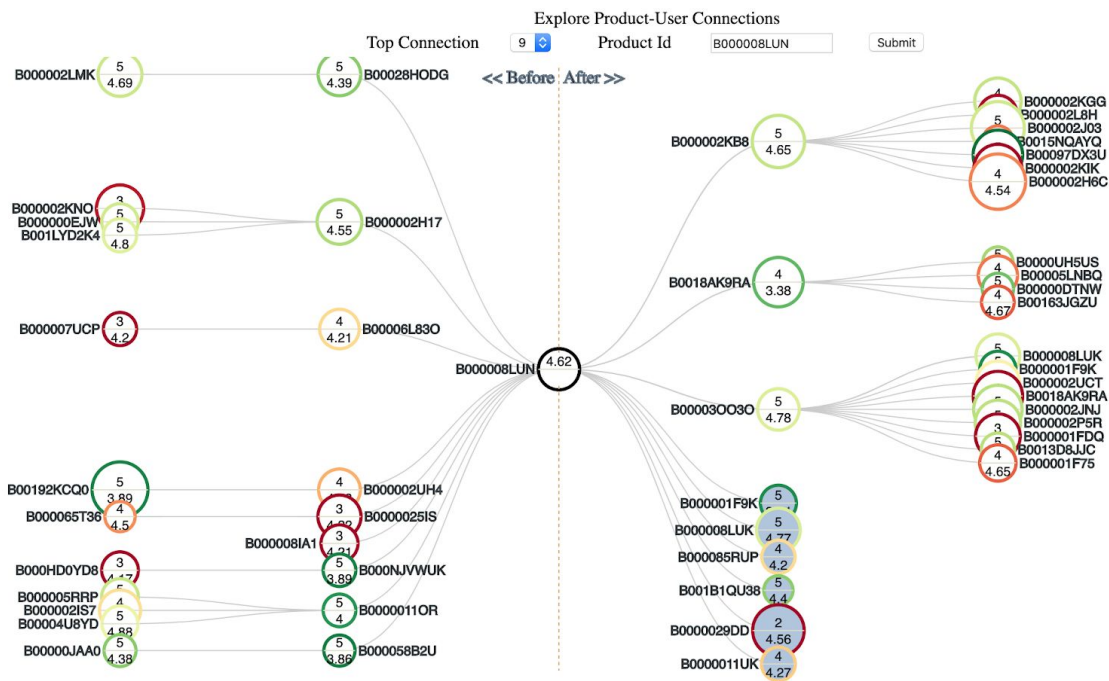


Figure 7

A Scalability Discussion

The design can scale to any data size. However tool has limited scalability in order to explore number of top products those are associated to our interest. With top 9 connections, it's becoming difficult to see the details on the nodes.



Information about the programs

Languages and libraries used :

Python : numpy, pandas, Flask

Javascript L D3, d3-context-menu

HTML

CSS

1. To run locally Python and mentioned python libraries need to be installed. A web browser is required to access it (Tool is tested only on Google Chrome browser). No additional software is required to install. Once the required python packages are installed, the following command has to be used from the root folder of the source code to run the tool.

> *python viz_serve.py*

web browser address : <http://127.0.0.1:5000/>

2. Additionally, the tool can also be accessed from the internet <https://cs765-dc2.herokuapp.com/>

Interaction

The tool is interactive. The details are covered in 'Tool Overview' section.

A short screencast recording is also uploaded to Canvas.

Data Sets

Customer product review data from `CDs_and_Vinyl_5.csv.gz` used. The application reads the compressed data file. Application read other datafiles as long as the column header and data types are consistent with current dataset format.

Self-Assessment

I am an experienced Python programmer and also used D3 in past (4 years back). However, implementing D3 was not an easy task for me. I spent considerable time in experimenting other libraries and searching solutions for D3 so that I would get desired results. I also spent a good amount of time in python coding.