Ruslan Rozenfeld

Brandon Nolan

5/9/2016

Text Mining Online Shopping Communities

1. **Introduction**

The increasing presence of online shopping has encouraged many different websites to allow their users to not only browse their catalog but to purchase it online as well, this has led these websites to see increased sales because of the convenience for users to purchase their products without traveling to the store, waiting in lines, and locating the items they like. Furthermore, this concept has been taken one step further with shopping websites that are purely online, giving their users a way to shop for items wherever they may be, such convenience and simplicity has been very well met by the user base that has been purchasing more and more.

In such, these purely online shopping websites have seen a dedicated user base community form that solely shop at their website. The online catalog for the two websites examined closely contain well over a million different items ranging from many different categories, this enormous selection is able to give most users the ability to purchase any item wanted for often a very low price compared to actual stores. Because of this relatively recent concept, not much has been studied about how each website's community compares to the other. Some of the questions that are more closely examined are which community rates similar items more positively or negatively, or which community tends to leave more reviews for similar items. Such analysis can give insight as to why the discrepancies happen in each community.

This project explores the trends of two online shopping communities, specifically NewEgg and Amazon electronics section. Thousands of items from the electronics catalog were mined from both sites with the item’s various attributes, such as rating, amount of reviews, and price, attached to each record delimited by a tilde. With this structured dataset mined, it is then used for visualization purposes to discern differences in how the two different communities act on similar items between the websites. To mine the catalog data, this project used a web scraping java library called jsoup [3] . Jsoup is an open source HTML5 parser that is capable of extracting web data and manipulating it [4]. In analyzing the visualization output, many discrepancies were found among the two datasets among similar items, mostly in number of reviews and ratings.

1. **Related Work**

Related studies of analyzing the trends of two different online communities using text mining have been done [1] that concluded positive opinions on topics that the website focuses on, another study concluded a similar thing [2]. These studies looked at message boards, so examining online shopping communities should yield interesting conclusions if the same trends follow for multiple categories of online communities, if so, future studies would have to examine if the same holds for websites that try to take no focus on a specific topic.

1. **Data**

This project text mined thousands of electronics items and their attributes between the two websites, NewEgg and Amazon, further, more specialized, datasets were created on narrowing the mining parameters to be able to compare similar items such as computer processors. In such multiple datasets were used in creating visualizations out of the two catalogs and hence drawing conclusions about trends in the online communities.

1. **Methods**

To mine the data, jsoup, a Java library for web scraping, located in the Maven repository, was used to mine the catalog data and their attributes [3]. The program takes a couple of parameters before it begins scraping the catalog data. Parameters such as the URL, output file location, and the CSS selectors that point to the data to be mined are a few of the tweakable parameters that make the program modular for both communities. Additionally, the ratings distribution per each catalog item, a useful piece of data that is located within the page of the item self, can also be mined with the program that can be tweaked in the config script. The output is a tilde delimited, clean formatted file that illustrates the catalog item name and/or description, the price of the item, number of reviews the item has, number of stars it has been given by the reviewers, and the category the item was mined in.

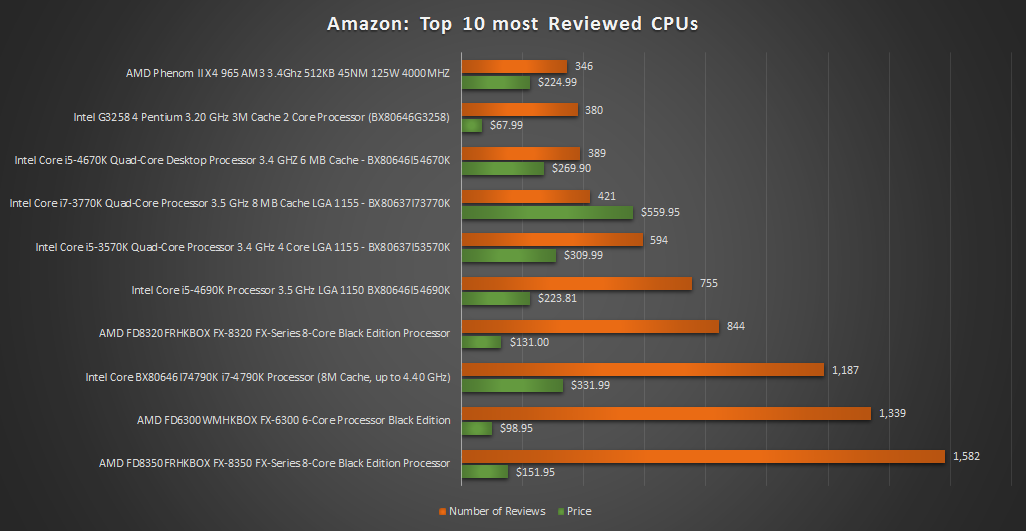
The delimited datasets were used in a visualization program to graph various attributes spanning across each community and look for trends in the data and compare this trend with the other community and see if it shares that same trend, if not, then this finding can be used as a premise to future studies. The visualizations focused on graphing all of the mined attributes in ways to illustrate any discrepancies across similar items. Number of reviews and overall item ratings were examined closely for both similar items and overall community activity throughout the electronics section.

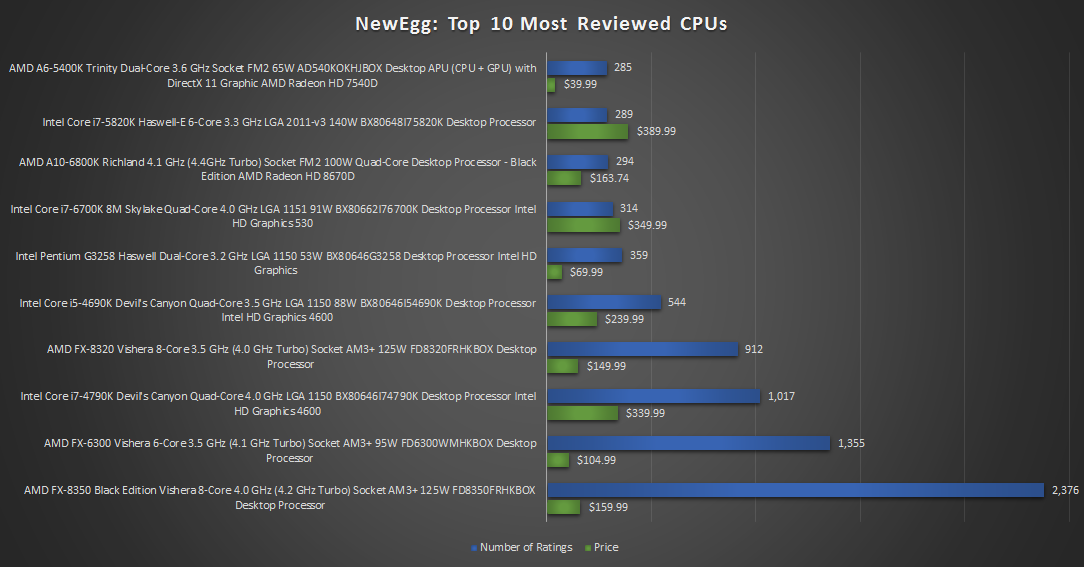
1. **Results**

In analyzing the graph data, some of the results that were encountered illustrated that the user base of Amazon generally leaves more reviews per item than on NewEgg while the user base of NewEgg tends to leave more reviews for a specific brand of computer processor, specifically AMD products, in contrast to Amazon, where Intel products are more reviewed and purchased. In addition, the correlation of NewEgg’s user base to AMD products is supported by the trend of AMD products leaning on the cheaper side than on Amazon, whereas Intel products are more expensive than on Amazon.

Furthermore, the star distribution, meaning how many people gave a star one through five on an item, is fairly consistent among the most relevant electronics catalog items. Additionally, the most relevant electronics results on both sites have fairly consistent number of reviews and ratings as well. Amazon also had a lower average price per item on the most relevant electronics items than NewEgg and also a higher average number of reviews for the same items.

Visualization of dataset:





NewEgg had slight tendency toward AMD products from the datasets than Amazon

1. **Conclusion**

With the graphs analyzed, some conclusions that can be made about both communities is that for the more focused computer processor category, users rated AMD products more than Intel ones with those products being slightly cheaper than on Amazon, on the contrary, Amazon had lower reviews for AMD products with those items being priced slightly higher than on NewEgg and Intel products were more sought after and reviewed higher than on NewEgg. Furthermore, Amazon had, in general, lower average price per item in the electronics section than NewEgg and generally more reviews showing that the Amazon community is more active and is purchasing different types of electronics items more than on NewEgg, while the user base of NewEgg mostly purchases only a few of the most relevant items with many of the other items in its electronics section do not have any reviews and are generally priced higher than similar items on Amazon. Overall, the price of a product played a large part in the reviews and ratings related products got in both communities.

1. **Future Work**

Related future studies can focus on finding correlations in other online communities instead of online shopping and discussion sites and seeing if the same hold for the conclusion of this and other studies on the topic. Additionally, tracking a sample of users across sites can lead to insight as to if their trends remain consistent among each community.

1. **References**

[1] Zhou, Lina, Liwei Dai, and Dongsong Zhang. "Online shopping acceptance model-A critical survey of consumer factors in online shopping." *Journal of Electronic Commerce Research* 8.1 (2007): 41.

[2] Dringus, Laurie P., and Timothy Ellis. "Using data mining as a strategy for assessing asynchronous discussion forums." *Computers & Education* 45.1 (2005): 141-160.

[3] "Jsoup: Java HTML Parser." Jsoup Java HTML Parser, with Best of DOM, CSS, and Jquery. Web. 08 May 2016.

[4] "3 Examples of Parsing HTML File in Java Using Jsoup." 3 Examples of Parsing HTML File in Java Using Jsoup. Web. 08 May 2016.