Text Analytics Project

CIS 4680 - Advanced Data Analytics

by: Team Kee-Bored

Background

The keyboard is the most essential tool for any PC user, it is our main way of controlling technology. Other than the mouse, most users have no other way to interact with their computer. Since we spend a great amount of time using this tool, we have grown to prefer certain qualities of keyboard that fit our needs. These qualities can range from the size of the keyboard, the type of sound or feeling they produce, to whether or not it can produce light.

It's hard to know what kind of keyboard to buy as a consumer because there are so many different options. Most people have no idea how to describe their ideal keyboard. The consumer could always use the product description to determine the keyboard that they should buy. However, those descriptions are often marketing tactics and are usually formulated to appeal to as many people as possible. Some descriptions include words that may hold no meaning to the consumer at all. Words like "mushy", "tactile", "smooth" on keyboards would make no sense unless you have actually tried them.

The consumer can always rely on fellow consumer's reviews when they can't trust the seller. We can find reviewers on online retail sites attempting to describe why they like or dislike the keyboard they bought. Most consumers would use these reviews as reference when buying a new keyboard. However, it will be hard to find a keyboard that matches the user's preference without sacrificing a significant amount of time.

Proposed Solution

Using text mining and analytics, we can use the hundreds upon hundreds of reviews that exist on the internet to figure out what features do people care about in keyboards? The most appropriate text mining technique that we have decided to use is clustering and sentiment analysis. To cluster the text data, we will utilize topic modeling. By extracting the main themes in the text, we can figure out what the review is focusing on. Usually, the main focus of the review is often the reason the reviewer would either recommend or disapprove the product. These main themes could contain the features that people care about in the keyboard.

Similar research has been done using Amazon reviews for text mining. L. Jack and Y.D. Tsai utilizes various text mining techniques to try to understand these reviews (2015). Some in their paper, they explore associations of words used in the reviews to the overall rating of the product. Using these associations, they were able to highlight what differentiated products with higher review scores compared to products with lower review scores.

We will use the latent dirichlet allocation (LDA) model from SciKitLearn's decomposition library for topic modeling. The reason we choose LDA is because it can tell us the relationship between words in the documents. Using the weights of those relationships, we can see how closely related the document is to a certain topic. A tricky thing with user reviews is that they can either recommend the product or tell other consumers to stay away from the product. If we don't figure out the sentiment in the review, then we cannot tell whether the user is recommending or criticizing the feature in question. Knowing the sentiment value can help us identify if the feature is worth keeping since it's highly praised or it's a feature to avoid since it is often criticized.

For sentiment analysis, we will be using one of NLTK's sentiment models called Valence Aware Dictionary for Sentiment Reasoning (VADER). This model has a dictionary that allows it to determine the intensity of emotions with certain words. Based on the intensity and sentiment of the words, we can get an overall sentiment value of the document.

Data Collection

We have decided to use Amazon since they have a great source of user-generated reviews. We coded a web scraper that goes through a list of Amazon pages and grabs all the reviews in that page. With this web scraper we are able to get the product name, price, brand, and reviews. We gathered over 1700 unique reviews and now, we need to clean them.

0]:

	Name	Price	Brand	Category	username	Title	Time	Review
0	\n\n\n\n\n\n\n\nPICTEK RGB Gaming Keyboard USB...	$21.99	PICTEK	Gaming	KisnardOnline	Love the keypress sound	Reviewed in the United States on October 24, 2019	\n I love typing on this keyboard. It is so...
1	\n\n\n\n\n\n\n\nPICTEK RGB Gaming Keyboard USB...	$21.99	PICTEK	Gaming	leo	seems good so far.	Reviewed in the United States on October 21, 2019	\n I ordered this keyboard with low expectat...
2	\n\n\n\n\n\n\n\nPICTEK RGB Gaming Keyboard USB...	$21.99	PICTEK	Gaming	Abhishek Shah	If you are fond of backlit.....get this	Reviewed in the United States on October 24, 2019	This is one of the best keyboards I have had.....
3	\n\n\n\n\n\n\n\nPICTEK RGB Gaming Keyboard USB...	$21.99	PICTEK	Gaming	L.E.E. Family	update: manufacturer offered free replacement,...	Reviewed in the United States on April 23, 2020	\n Update:Pictek got in contact with me and o...
4	\n\n\n\n\n\n\n\nPICTEK RGB Gaming Keyboard USB...	$21.99	PICTEK	Gaming	Hugo	D- Best Keyboard for the Buck	Reviewed in the United States on June 10, 2020	\n Let me tell you as pro-gamer this keyboard...

Preprocessing and Cleaning

Before implementing any text analysis methods, it is crucial to clean your dataset to provide the most accurate results. We first imported the raw keyboard reviews from web scraping. Since the "title" and the "reviews" are in different columns, it is easier to clean the texts by combining them into one "comment" column. Once the "comment" column is added into the dataframe, we delete the "title" and "reviews" columns to eliminate redundancy.

	Time	wired	gaming	mechanical	ergonomic	membrane	backlit	Comment
0	2019-10-24	True	True	False	False	False	True	Love the keypress soundI love typing on this k...
1	2019-10-21	True	True	False	False	False	True	seems good so far.I ordered this keyboard with...
2	2019-10-24	True	True	False	False	False	True	If you are fond of backlit.....get thisThis is...
3	2020-04-23	True	True	False	False	False	True	update: manufacturer offered free replacement,...
4	2020-06-10	True	True	False	False	False	True	D- Best Keyboard for the BuckLet me tell you a...

While looking through our keyboard reviews, we noticed that some of the reviews are in different languages. This could cause an issue in the future because we are gathering the probability of important words based off of the english dictionary. So having non-english reviews is not necessary and should be dropped using langdetect. Afterwards, we must tokenize all of the reviews from sentences to words. Using NLTK, we implemented the stopwords and excluded the negative words such as "no", "nor", and "not". By doing this, we will then be able to apply stemming and lemmatization to uncover only the most important words in the review and put them all into a corpus.

Lastly, we had to normalize the corpus using regular expressions. For the corpus, all special characters, numbers, single characters, extra white spaces were removed from the corpus.

Having all of the words lowercase made it more efficient for the normalization process. We tokenized, filtered out all the stopwords, and lemmatized each remaining using a function called "normalize_docuement". After normalizing the corpus, we used a TF-IDF vectorizer to calculate the weighted score of each word and how frequently they show up in the reviews. This will then allow us to create a new dataframe containing only important words which will be used for sentiment analysis.

	also	feel	get	good	great	key	light	like	love	mouse	...	really	time	type	typing	use	used	using	well	work	would
0	0.17	0.15	0.00	0.00	0.00	0.33	0.17	0.40	0.51	0.00	...	0.00	0.00	0.36	0.18	0.14	0.18	0.00	0.00	0.13	0.16
1	0.25	0.45	0.26	0.45	0.00	0.17	0.26	0.40	0.00	0.00	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.20	0.00
2	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.00	0.00
3	0.27	0.25	0.00	0.25	0.22	0.37	0.00	0.00	0.00	0.00	...	0.28	0.00	0.00	0.00	0.46	0.00	0.00	0.28	0.22	0.00
4	0.22	0.20	0.23	0.20	0.00	0.00	0.00	0.70	0.22	0.00	...	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.22	0.00	0.00

Sentiment Analysis

Sentiment analysis is very important in our project to understand customers' feelings about the keyboards. There are two different ways to perform sentiment analysis. The first one is to use text classification to predict sentiment labels on pre-labeled dataset. This method requires us to have a predefined label in the dataset. Since our data set does not contain rating scores which is the label for supervised learning, and the size of our data is not big enough for using classification models or Word2Vec, we decided to use the second method -- a lexicon based model for our sentiment analysis. A lexicon based model is an unsupervised machine learning method which does not need a label. By matching each word to a predefined lexicon, each word in the comment will get a sentiment score.

After aggregating each word's sentiment score, we will be able to identify the sentiment level of each comment. NLTK provides a couple of built in models for sentiment analysis. We used VADER in our study. It can be directly applied to unlabeled text and return polarity and intensity of the emotion. VADER sentiment analysis used a predefined dictionary which maps words in the text to a sentiment score. A positive sentiment score means a positive emotion, while a negative sentiment score means negative emotions. The higher the absolute value means a stronger emotion.

Based on our study, we have 1663 positive reviews and 173 negative reviews. Negative reviews only count less than 10% of the total reviews. The unbalanced distribution of the negative and positive reviews increases the difficulty for us to build a recommendation system. We have 2 assumptions which explain the reason we get much fewer negative reviews than positive reviews:

1. By nature, positive reviews are more common than negative reviews. When we are collecting the data, we scrape the information of products in the first 20 pages. By default, the built in recommendation algorithms in Amazon already filtered our data. Therefore, our dataset may not contain products with lower rating scores.
2. It is hard for machines to detect negative sentiment. Even though we modified the list of stop words by removing 'no', 'not', 'nor' and n't', we are not sure the built in VADER module will reverse the sentiment score of the related text. For example, 'like' has a sentiment score of +2. If before the word 'like' has 'don't',

the sentiment score should be '-2'. Since we tokenize the sentence to each word and only use sum to calculate the sentiment score, We estimated that VADER is not able to classify negative comments accurately.

Therefore, It is convenient to have a built-in lexicon for performing sentiment analysis. However, this model is not very accurate and has some limitations.

	Comment	Topic	scores	compound	comp_category
0	Love the keypress soundI love typing on this k...	4	{'neg': 0.006, 'neu': 0.729, 'pos': 0.266, 'co...	0.9974	pos
1	seems good so far.I ordered this keyboard with...	0	{'neg': 0.043, 'neu': 0.803, 'pos': 0.154, 'co...	0.9749	pos
2	If you are fond of backlit.....get thisThis is...	4	{'neg': 0.0, 'neu': 0.705, 'pos': 0.295, 'comp...	0.9842	pos
3	update: manufacturer offered free replacement,...	2	{'neg': 0.079, 'neu': 0.771, 'pos': 0.15, 'com...	0.9469	pos
4	D- Best Keyboard for the BuckLet me tell you a...	2	{'neg': 0.014, 'neu': 0.794, 'pos': 0.192, 'co...	0.9828	pos
...	...	...	...	...	...
1854	just awesomenice and heavy, stays in place. RG...	0	{'neg': 0.0, 'neu': 0.718, 'pos': 0.282, 'comp...	0.8625	pos
1855	Worth the PriceWouldn't usually write a review...	1	{'neg': 0.0, 'neu': 0.474, 'pos': 0.526, 'comp...	0.9786	pos
1856	Sturdy, great action, thoughtful designThis Hy...	4	{'neg': 0.036, 'neu': 0.743, 'pos': 0.221, 'co...	0.9993	pos
1857	Love the lighting and keypress, but no wrist w...	3	{'neg': 0.056, 'neu': 0.805, 'pos': 0.139, 'co...	0.9973	pos
1858	The Most Disappointing Keyboard in 2020This is...	0	{'neg': 0.157, 'neu': 0.773, 'pos': 0.07, 'com...	-0.9771	neg

The table listed above is the result we got. The column 'scores' listed the details of sentiment analysis. It lists the score for negative, neutral, and positive sentiments. The column compound is the sum of negative, neutral and positive features. If the score is over 0, we classify it as a positive review. If the score is negative, we label it as a negative review. The 'comp_category' shows the category of reviews.

Clustering and Latent Dirichlet Allocation (LDA)

Cluster was the next step in our analysis that helps to answer our research question. Our question largely depends on the topics of the reviews, so this section was arguably the most important. After much experimentation with various clustering methods, we finally decided on the N Gram Bag of Words Model with 4 words. This model gave us the best idea of the topics that were found. In regard to the number of topics, we eventually settled on 5. We decided on this number as a group after experimenting with different numbers. We found 5 to be the "sweet spot" when creating topics that fit our subject.

As we used the Bag of Words technique, we needed to create a CountVectorizer for which we gave the properties of min_df=2 and max_df=0.95. We then fit the normalized corpus to the CountVectorizer which allowed us to finally pass the data onto the Latent Dirichlet Allocation (LDA) step.

Results

We defined our five topics based on the features extracted through clustering and Latent Dirichlet Allocation.

1. Features Extracted
    a. Wireless Mechanical Keyboard and Mouse Combo (Topic 0)
        i. A wireless mechanical keyboard and mouse combo is a great feature set for a product that is targeted mainly at casual users who need both items.

```
THE TOP 10 WORDS FOR TOPIC #0
['update review anything change', 'need two aaa battery', 'feel like laptop keyboard', 'would highly recommend anyone', 'wou
ld recommend keyboard anyone', 'left right mouse button', 'love keyboard mouse combo', 'really cant go wrong', 'cherry mx bl
ue switch', 'wireless keyboard mouse combo']
```

    b. Budget Wired Mechanical (Topic 1)
        i. Mechanical keyboards are preferred by typists, gamers, and enthusiasts. They are generally the most expensive keyboard type, so offering a budget option can attract consumers.

```
THE TOP 10 WORDS FOR TOPIC #1
['didnt want spend lot', 'key located top row', 'overall would highly recommend', 'dont even know key', 'key great feel lik
e', 'would highly recommend keyboard', 'bump cherry mx brown', 'tactile bump cherry mx', 'highly recommend keyboard anyone',
'small form factor keyboard']
```

    c. Basic Wireless Keyboard and Mouse Combo(Topic 2)
        i. A basic keyboard and mouse will appeal to the broadest audience searching for the most basic of computing accessories.

```
THE TOP 10 WORDS FOR TOPIC #2
['keyboard would recommend anyone', 'http smile amazon com', 'smile amazon com gp', 'keyboard mouse combo great', 'top row f
unction key', 'full size number pad', 'take bit get used', 'battery life pretty good', 'constantly go back correct', 'take t
ime get used']
```

    d. Basic Wireless(Topic 3)
        i. A basic wireless keyboard has mass appeal, especially as it is an item that is subject to a lot of wear and can be subject to frequent replacement.

```
THE TOP 10 WORDS FOR TOPIC #3
['http www amazon com', 'pbt double shot keycaps', 'best wireless keyboard ive', 'wireless keyboard ive ever', 'cap lock num
lock', 'keyboard ive ever owned', 'really like keyboard mouse', 'best keyboard ive ever', 'num lock cap lock', 'keyboard ive
ever used']
```

    e. Gaming Wired Keyboard and Mouse Combo (Topic 4)
        i. A wired keyboard and mouse are very popular with gamers as they do not require charging, and have low input latency.

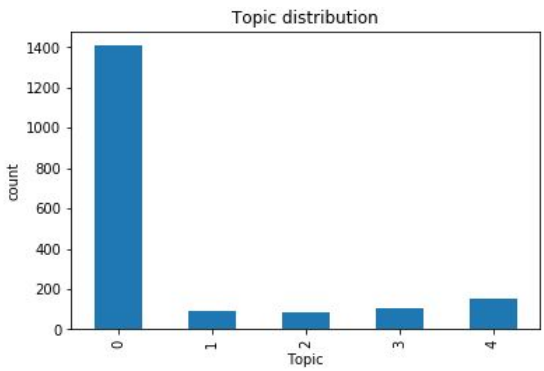
```
THE TOP 10 WORDS FOR TOPIC #4
['key take getting used', 'would definitely recommend keyboard', 'working great key easy', 'key nice tactile feel', 'led rgb
keyboard mouse', 'gaming led rgb keyboard', 'take little getting used', 'love keyboard mouse combo', 'keyboard mouse combo p
rice', 'nice keyboard mouse combo']
```

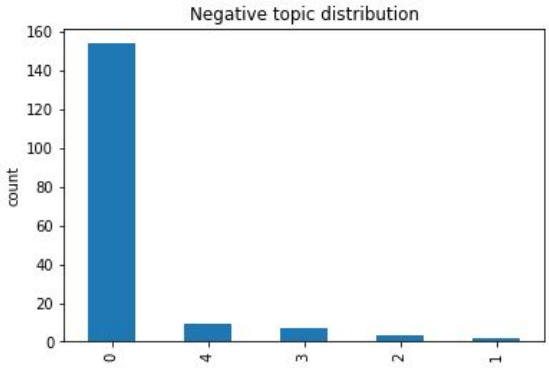
2. Features with Sentiment

After grouping topics and calculating the average sentiment scores in each keyboard category, we found that Topic 0 (Wireless Mechanical Keyboard and Mouse Combo) has the lowest sentiment score, while Topic 1 (Budget Mechanical) has the highest sentiment score. Our assumption is that people who bought budget keyboards care more about the price, not the quality. As long as the price meets their budget, they are happy with their shopping experience.

```
df.groupby('Topic').mean()['compound']
```

```
Topic
0    0.678401
1    0.869773
2    0.841532
3    0.822895
4    0.830539
Name: compound, dtype: float64
```





However, after looking into the distribution of the clusters, we found out the distribution is not balanced. Topic 0 accounts for the majority of the data, but at the same time has the lowest sentiment score. After we filtered out negative reviews, we found a similar distribution as topic distribution. Topic 0 has the majority of the negative reviews. This explained why Topic 0 has a lower sentiment score. However, the unbalanced distribution of different clusters is an issue we should improve in the future.

Discussion

In order to evaluate the results of our efforts, we first need to look at the topics that were modeled through clustering analysis. The most interesting topics are the ones that involve mouse combos. Since product bundling accounts for three of the two topics, we can consider it an important part of the consumer's priority when purchasing a keyboard. Another thing we noticed is that these topics are fairly general. It's separated by whether it is mechanical, wired, wireless, or a gaming keyboard. Although "gaming" is not really a "type" of keyboard, this could reveal how people are drawn to buzz words on the keyboard rather than the keyboard itself.

The sentiment around the keyboards shows how each topic is generally perceived as. Wireless Mechanical Keyboard and Mouse Combo has the lowest sentiment score which could be the result of trying to target a more general audience. Since there are more people that product is trying to target, the more people that could potentially dislike the product. We can see that the count of the reviews in that topic is exceptionally large compared to the other topics which could

also affect the sentiment score. In general however, the overall sentiment of all the topics are positive.

There are a lot of things that could be improved to the models and methodologies that we used. The main limitation that we had was our data. The large amount of positive reviews compared to negative reviews could be the result of only scrapping top result keyboards. Top result keyboards tend to have a high rating score which means that most of the reviews will be positive in favor of the product. To combat this limitation, we should actively try to introduce a diverse amount of product reviews, including reviews from products that have a really low score.

Another improvement we could try to make is the process of analyzing text reviews. One reason that the topics that we generated are so general is that general purpose products are more likely to sell than enthusiast products. Enthusiast reviews tend to be really comprehensive, but they are few in numbers. We would need to try to separate enthusiast products in order to try to extract some topics within that class of keyboards. However, it still stands that companies that make general purpose keyboards are reaching a wide market with generally positive sentiment.

Conclusion

From our analysis, the main feature that people tend to care about has nothing to do with specific aspects of the keyboard. Functionality such as being wired or wireless is more important than any style. There is also the distinction between normal and gaming keyboard that people will tend to emphasize. The most important insight that we are able to produce is that people are very vocal and receptive about product bundles with keyboards. In general, the sentiment around enthusiast and general use keyboards remain around the same level. While this can be due to the limitation of the data we collected, it does make sense that people would view any functional keyboard in a positive light.

Bibliography

Jack, L., & Tsai, Y. (2015, January). Using text mining of amazon reviews to explore user-defined product highlights and issues. In *Proceedings of the International Conference on Data Mining (DMIN)* (p. 92).

Portilla, Jose. NLP - Natural Language Processing with Python.