Course Project Report

Product Reviews Summarizer

Parv Bhatt (pnb5078@psu.edu)
Namratha Sri Mateti (njm5914@psu.edu)
Dominic Thomas (dxt5349@psu.edu)

School of Graduate Professional Studies
Software Engineering Department
SWENG 545 – Data Mining

December, 2021

DOCUMENT CONTROL

Work carried out by:

Name	Email Address	Other
Parv Bhatt	pnb5078@psu.edu	
Namratha Sri Mateti	njm5914@psu.edu	
Dominic Thomas	dxt5349@psu.edu	

Project Timeline

Release No.	Date	Revision Description
1	11/07/2021	Built a scraper to collect data from Amazon.com website
2	11/14/2021	Performed pre-processing on the scraped data
3	11/21/2021	Clustered the reviews and extracted the top keywords for each cluster
4	12/04/2021	Summarized the cluster reviews
5	12/08/2021	Final Project Presentation

COURSE PROJECT REPORT

TABLE OF CONTENTS

1. INTRODUCTION

Companies like Amazon have millions of reviews on their website, from customers all over the world. Given the nature of the site and the fact that their users are looking for the best product to buy, having to sift through hundreds of reviews to find a product can be a real turn off. Text Analysis and Clustering can be used here to build tools that can summarize multiple properties in 1-2 sentences. Instead of scrolling through a list of reviews, you could simply say “Excellent product but has a history of bad packaging”.

2. DATA

2.1 Collection

Data was scraped from the Amazon.com website. Docker was used to create a Splash container, a light-weight browser with an HTTP API. The BeautifulSoup library and html.parser were used to collect the ‘product’, ‘title’, ‘rating’, and ‘body’ of the reviews for ten products. The products were selected in such a way that it covered a variety of departments and average star ratings. The scraped data was saved in .xlsx format using the Pandas library.

2.2 Pre-processing

To prevent the mismatch of datatype, all the saved data was explicitly converted into string datatype. The next step involved tokenizing each of the sentences from the pool of the reviews. Every non-alphabetical character in the pool was replaced by a blank character and the alphabetical characters were converted to lower-case. Using the stopwords module from the nltk library, about forty most common stopwords were removed from the data. The usage of the WordNetLemmatizer and PorterStemmer modules of the nltk.stem library reduced the efficiency of the model and thus were not used in this project.

	A	B	C	D
1	product	title	rating	body
11	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	Worth the money	4	Your browser does not support HTML5 video. *I was fully reimbursed for this iron* I will be honest with you guys, I thought I got a broken iron. I couldn't get it hot enough to iron a simple cotton dress or get the steam option to work. Took a good 20 minutes before I realized I was aligning the dial in the wrong spot (I noted where it should be aligned). Then things started working correctly. I could hear the steam getting ready and it was able to get wrinkles out almost immediately with no steam needed. This will most likely replace my 20$ Wal-Mart iron I bought 4 years ago. I used to covet Rowenta irons, still wouldn't mind one but this one will do fantastically until then. The auto shut off function does work in the horizontal potistion but I didn't test vertical becasue I have a curious toddler at home. It's got the option to vertically steam (which is so cool but it scared me the first time I tried), uses tap water only (woo no more jugs of distilled water!!) and the cord winds up in the base *this does make it a bit heavier than a standard iron. * It has an anti drip system but it will drip if you, like me, test out the steam button on the low setting. It needs to be on high. All said and done I would recommend buying for an "I just moved out and I need an iron" or "I need a replaced workhorse to complete my sewing projects". I already own a MUELLER slicer and this iron is just as good quality as that is.
12	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	Am I Just Stupid?	3	It seems like a good iron. It's heavy and much larger than I thought it would be (wish it were smaller). I don't do a lot of ironing even though I've been using one for 50 years. I dropped my old iron and it broke so I purchased this one. To be frank, this iron is a bit too complicated for my needs and since I don't use it that often, I have to pull out the manual each time I use it because I don't remember what the buttons and dials are for. There's no ON/OFF switch and the dial is behind the handle which makes it difficult to read and there's no easy to read arrow marking the dial setting. I had to take some nail polish and paint the raised arrow. The only easy to read label is the Mueller brand on the side of the iron!The cord doesn't retract like it should. I have to feed the cord into the hole for storage.I will keep this and use it until it dies but I would have preferred a much smaller, simpler iron with easy to read buttons and dials.
13	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	Big improvement over my last iron	5	I iron 5 days a week,and had a series of irons from a well know brand that starts with "sun". They were just OK and the last one quit on my with a puff of smoke after 2 years. Supposedly it was under warranty but the process was so cumbersome, I gave up and went in search of a new and different model. I like having a retractable cord and not winding it around a hot iron.So far, so great with this new Mueller! The water tank is much larger, the tip has a perfect shape to get in and around the buttons and tight spots on my shirts. My old iron was listed as "large" but this one has the same heated surface area and feels better in my hand.I would recommend to a friend
14	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	Mueller MU-77X Iron is JUNK 3 Months it Fails	1	Mueller MU-77X Iron. Well it looked awesome, felt awesome and performed awesome for 3 months and 2 days. Spouse went to iron and no green light and no heat. It decided to take a long nap.Perfectly handled and stored each night. We have 8 kids and professional jobs so we iron our business clothes almost daily. Sent us a replacement and 2 months later it started to smell like a burning mess so it was tossed. Updated review to reflect the soon to burn house down iron.Amazon is buying slots on google to push products that are junk.Do not buy the Mueller MU-77x if you need a long lasting iron.

Figure 1: Data Scraped from Amazon website

3. METHODOLOGY

3.1 Clustering

Kmeans clustering algorithm was used to group the similar reviews together and find the top keywords for each of the cluster. The cluster algorithm was biased towards the common words, used frequently for that product, and did not yield any satisfactory results. Grouping the reviews based on their star ratings proved to be a much more effective than using any of the clustering algorithms.

3.2 Cosine Similarity without tf-idf

The reviews classified based on the star ratings were then represented in the form of vectors. Using the pretrained GloVe Word Embeddings and the cosine similarity algorithm, a similarity matrix was prepared. The next step involved the application of a pagerank algorithm to extract the top 10 ranked sentences. The issue which was faced with this model was that it had an $O(n^2)$ time complexity. The model took around 120 seconds to rank about 1000 sentences. Increase this number to 10000 sentences (for some popular products), and the time increases to 12000 seconds or about $3^{1/4}$ hours!

3.3 Cosine Similarity with tf-idf

Before applying the cosine similarity algorithm, the tf-idf measure was used to calculate the frequency of each word in the pool of reviews and assign a score to the corresponding sentence. The nlargest module from the heapq library was used to capture the top 1000 scored sentences for the cosine similarity algorithm. Usage of the tf-idf measure, ensured that each product, irrespective of the number of sentences in the reviews, gave an output within 120 seconds. This method made sure no important feature is lost, giving similar results as the previous method but in considerately less time.

3.4 Hugging Face Transformers

The top ten ranked sentences after the cosine similarity algorithm were sent in the hugging face transformers pipeline to create a summary for that star rating. A common min and max length of the summary was specified. The output along with the product name, star rating, percentage/count of number of reviews for that star rating, the top 5 frequent words along with count, and the summary of the review were stored in .xlsx format.

4. CONCLUSIONS

Thousands of reviews were accurately summarized within 300 characters, retaining all the important features. Thus, helping the customers sift through the product reviews effectively and making a wise decision on buying a product or not.

	A	B	C	D	E	F
1	product	rating	summary	Counts	Percentage of Total	Most Common Words
2	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	1	the iron is not new it's used and it doesn't work . after two months the iron started smelling like the plastic was burning . the iron was dirty like a dusty hand grabbed it. this is the second time that I have bought an iron on amazon . it has come showing obvious where in tear from repeated use .	183	8.8	('Iron', 181) ('Water', 78) ('Cord', 58) ('Used', 56) ('Months', 39)
3	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	2	the iron is heavy, it never turns off and its awkward, and I don't care for the retraceable cord . it replaced a 1,500-watt iron on which the power cord had failed, but otherwise did a splendid job . i could deal with a $25 iron that worked ok but this cord messing up the fabric has taught me to look for one not in that position .	68	3.3	('Iron', 68) ('Cord', 45) ('Water', 31) ('Used', 20) ('Like', 20)
4	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	3	the iron works well but the retractable plug wouldn't go back in . the iron was crinkled and stiff like paper that got wet, and then dry . a water chamber was partially filled, and the source of the water damage . I don't do a lot of ironing even though I've been using one for 50 years .	80	3.9	('Iron', 86) ('Cord', 60) ('Water', 20) ('Heavy', 19) ('Steam', 19)
5	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	4	the iron is good the steam it does to remove wrinkles is good but the water comes out and wets the clothes more than normal . if you are using it for sewing I would use one that doesn't automatically turn off, but that's a nice thing about this iron if your forgetful.the iron does not fall and I was taking care of it so that it would not be damaged but it was.	207	10	('Iron', 249) ('Cord', 116) ('Steam', 74) ('Retractable', 64) ('Water', 61)
6	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	5	great iron,have it for a week and used every day, a very good product for its price,the water capacity is perfect and it makes good amount of steam .I have never had an iron with a retractable cord, it seems like a small detail but definitely helps for an easy clean up especially because I pull out the iron for last-minute situations . this iron is well made, it's light weight enough for easy use, but doesn't feel cheaply made .	1537	74.1	('Iron', 1418) ('Great', 611) ('Love', 520) ('Cord', 506) ('Steam', 442)

Figure 2: Summarized Output

5. IMPROVEMENTS AND FUTURE APPLICATIONS

The date and the location of the review can also be considered for effective summary building. A product may have a good review at a particular location/time but can have a totally different review at some other location or some other point of time. An end-to-end product like a mobile app can be developed which can enable the user to summarize any product just by inputting the product URL. This model can further be extended to movie reviews, restaurant reviews, tourist places reviews, etc.