Text Analysis on Amazon Product Reviews of Office Product

1. Introduction to Data and Research Question

1.1 Data Introduction

This research uses the dataset of Amazon Reviews from the following link: https://nijianmo.github.io/amazon/index.html#sample-review. Because this Analysis focus on category of Office Product, only the Office Product subset provided is downloaded for analyzing. This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features). The review data has been reduced to extract the 5-core, such that each of the remaining users and items have k reviews each.

Table 1: Column Explaination of the Metadata Dataset

Column	Explaination
reviewerID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin	ID of the product, e.g. 0000013714
reviewerName	name of the reviewer
vote	helpful votes of the review
style	a disctionary of the product metadata, e.g., "Format" is "Hardcover"
reviewText	text of the review
overall	rating of the product
summary	summary of the review
unixReviewTime	time of the review (unix time)
reviewTime	time of the review (raw)
image	images that users post after they have received the product

Table 2: Column Explaination of the Review Data Dataset

Column	Explaination
reviewerID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin	ID of the product, e.g. 0000013714
reviewerName	name of the reviewer
vote	helpful votes of the review
style	a disctionary of the product metadata, e.g., "Format" is "Hardcover"
reviewText	text of the review
overall	rating of the product
summary	summary of the review
unixReviewTime	time of the review (unix time)
reviewTime	time of the review (raw)
image	images that users post after they have received the product

1.2 Research Question

The research question is **what rating performance, text topics are of products**. As for the application, first, **rating performance and text topics show marketing trend**, which can help amazon to develop related sellers and make better brand advertisements. Second, the research provides **concise improvement advice**, since it extracts low-rating products and some text topics reveal accurate optimization methods. Thus, it can be applied to the real business environment effectively.

The research conducts Natural Language Processing (NLP) to determine the review text topics of each product. For low rating products, in some degree, product optimization direction can be extracted from text topics. For example, one of the methods to improve low rating printers is "optimize the appearance of printers and reduce printing waiting time". Text topics contribute to conclusions about what products to improve and how to improve them.

There are many E-Commerce platform researches. In 2012, Mudambi & Schuff dig into the rating of 1,587 Amazon product reviews. They explored what makes customer reviews helpful to a consumer in the process of making a purchase decision. Results indicated that review extremity, review depth, and product type affect the perceived helpfulness of the review(Susan M. Mudambi, 2012).

There are also many NLP analysis on review text of E-Commerce. In 2017, to avoid hand-crafted rules and predefined templates in traditional natural language generation methods, Dong et al. present an attention-enhanced attribute-to-sequence model to generate product reviews for given attribute information, such as user, product, and rating(Li Dong, 2017). In 2018, Haque et al. use supervised learning method on a large scale amazon dataset to polarize reviews and get satisfactory accuracy(Tanjim Ul Haque, 2018). In 2018, Ni et al. build assistive systems that can help users to write reviews(Jianmo Ni, 2018).

Existing researches reveal the importance to do E-Commerce review analysis. Meanwhile NLP techniques makes the reviews very valuable, since reviews can not only help business to get profit, but improve the feeling of users. In this research, review analysis on Office Product will help platform to improve service.

2. Data Processing and Exploration

2.1 Data Processing

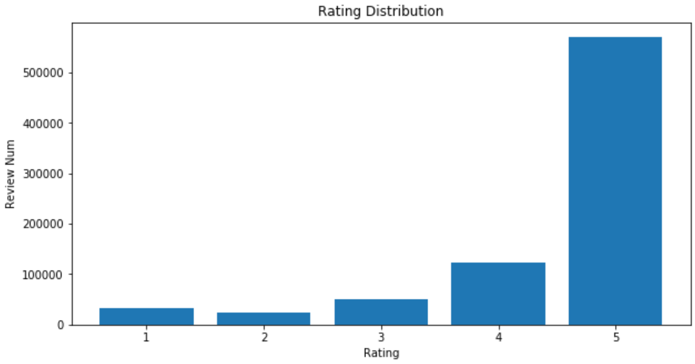
Clean Data and Transform Text to Tokenized Form: This research cleaned reviews with null text, tokenizing reviews text (including removing punctuations, tokenization, removing stopwords and lammitization). With the processed review dataset, text analysis and topic modelling can be conducted.

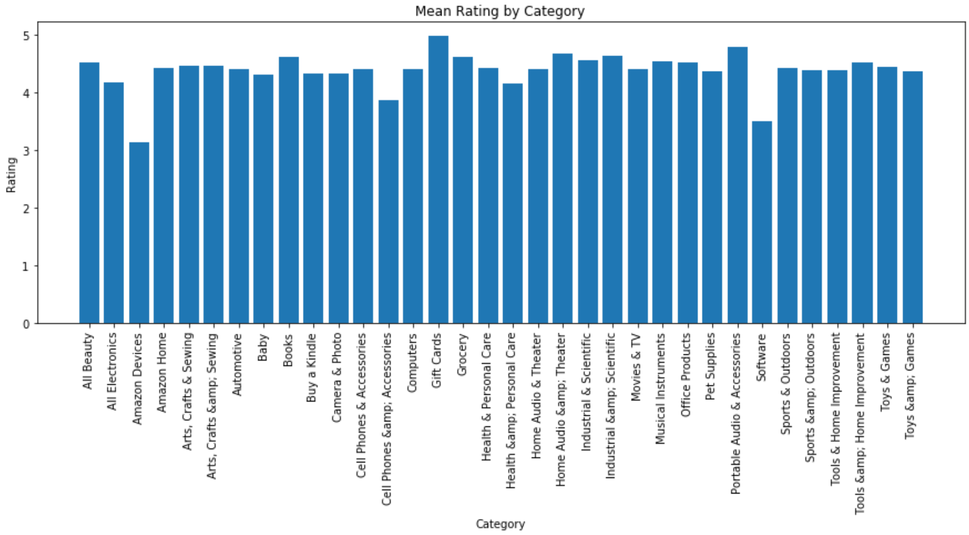
Integrated Dataset: Linked two dataset on the key asin to get a merged dataset with both product information and review performance. Base on the integrated dataset, research on review rating performance of different products can be conducted.

2.2 Data Exploration

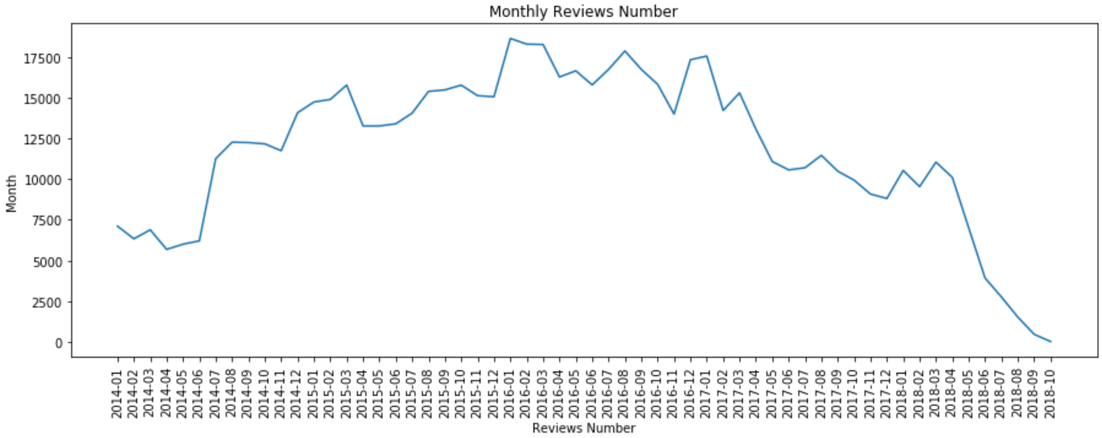
For rating of the products, the mean rating is 4.47. Rating Distribution shows that most reviews are high rating. Both the mean rating and the rating distribution reflect that consumers have good experience when buy and using products form Amazon.

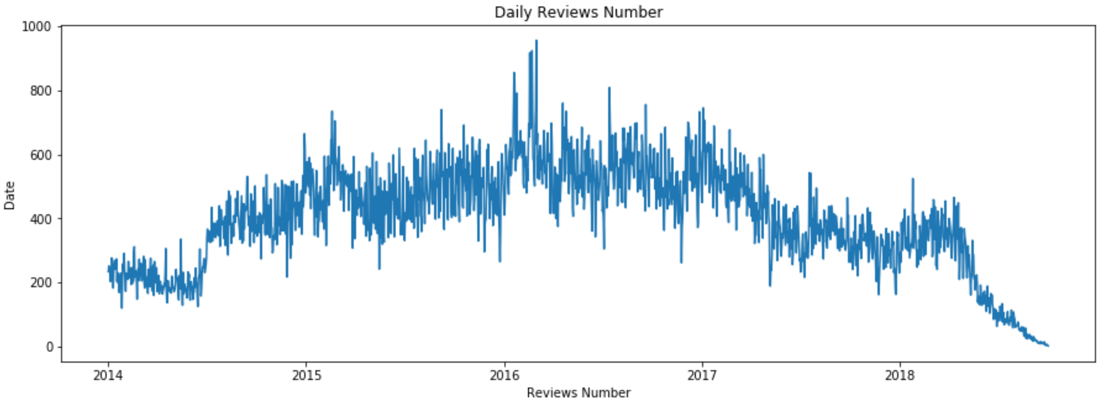
Mean Rating by Categories shows that Amazon Devices, Software and Cellphones & Accessories have low mean rating (In Amazon, products can be divided into several categories, so Office Product can be other categories at the same time), so **products in these categories are ones that Amazon should contribute to improve**.



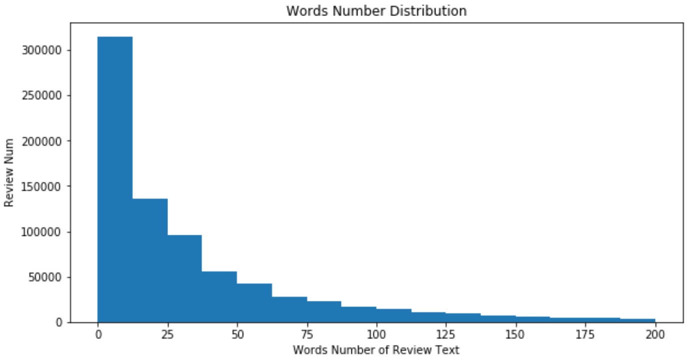


For number of reviews, the total number of reviews 800,144. Reviews Number Monthly Trend shows that from December to January next year, the number of reviews surges, and later decreases. May be this is for the reason that people rework and buy office product to prepare for the new work after Christmas Holiday.

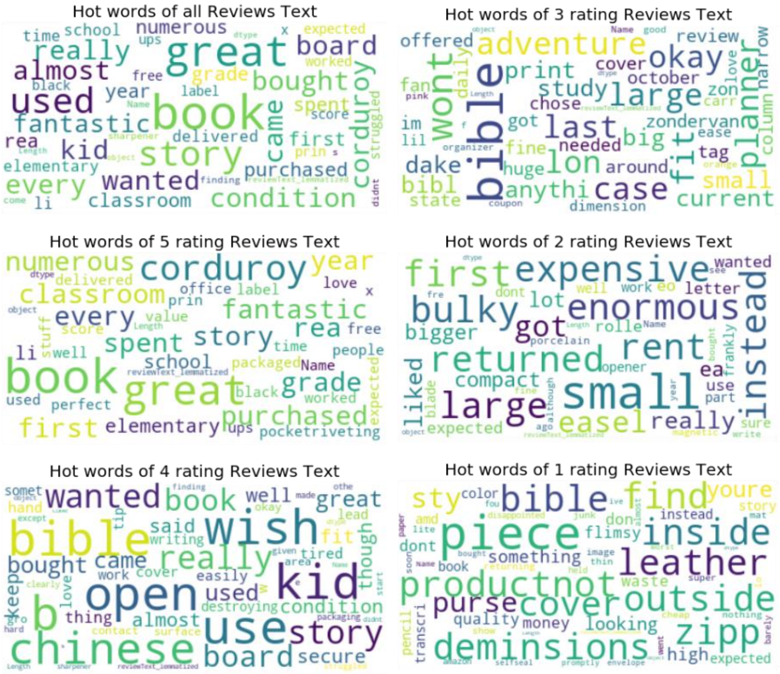




For number of text words, the total number of review text words is 35055381, the mean num of words per review text is 43.81. Words Number Distribution Shows that most of the people writes short review text.



As for Hot words of reviews, reviews with different rating are also not the same. **Overall, words are common and positive**, such as book, story and great. For high-rating products, words are positive, such as fantastic, great and love. **For words with only rating 1, there is little positive word. Instead, negative words such as waste, disappointed appear.**



3. Topic Modelling Visualization and Interpretation

3.1 Topic Modelling Instruction

This research applies LDA topic modelling method to extract topics of reviews text. In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Latent Dirichlet Allocation, LDA is yet another transformation from bag-of-words counts into a topic space of lower dimensionality. LDA is a probabilistic extension of LSA (also called multinomial PCA), so LDA's topics can be interpreted as probability distributions over words. These distributions are, just like with LSA, inferred automatically from a training corpus. Documents are in turn interpreted as a (soft) mixture of these topics (again, just like with LSA)(Blei, 2003).

3.2 Topic Modelling on All Reviews Text

This research conducts LDA Topic modeling. Text topics show marketing trend, which can help Amazon to develop related sellers and make better brand advertisements. After testing, it performs the best when topic number is 10. The 10 topics and their key words are shown in Table 3. Topic clusters and their relationship with key words are shown in the picture Topic Clusters and their Relationship with Key Words. Conclusions about marketing trend can be drawn as below:

3.2.1 Cost Performance of Products is an Important Aspect for Users to Consider

From the topic 1: great and cheap, customers consider balance between quality and price of products. Cost performance play a very import aspect for users. Operation team should label products that have good cost performance, and advertise these products to users that consider it. By this marketing strategy, sales would improve.

3.2.2 Staples & Shredders Provided by HP are Very Popular

From the topic 3: staple & shredder, devices provided by HP get good reviews text including loving, highly recommending. Amazon can cooperate tightly with HP and try to keep the product quality and provide more. This Official cooperation is a very good support both for Amazon and HP. As a result, both product supply and brand value will be better.

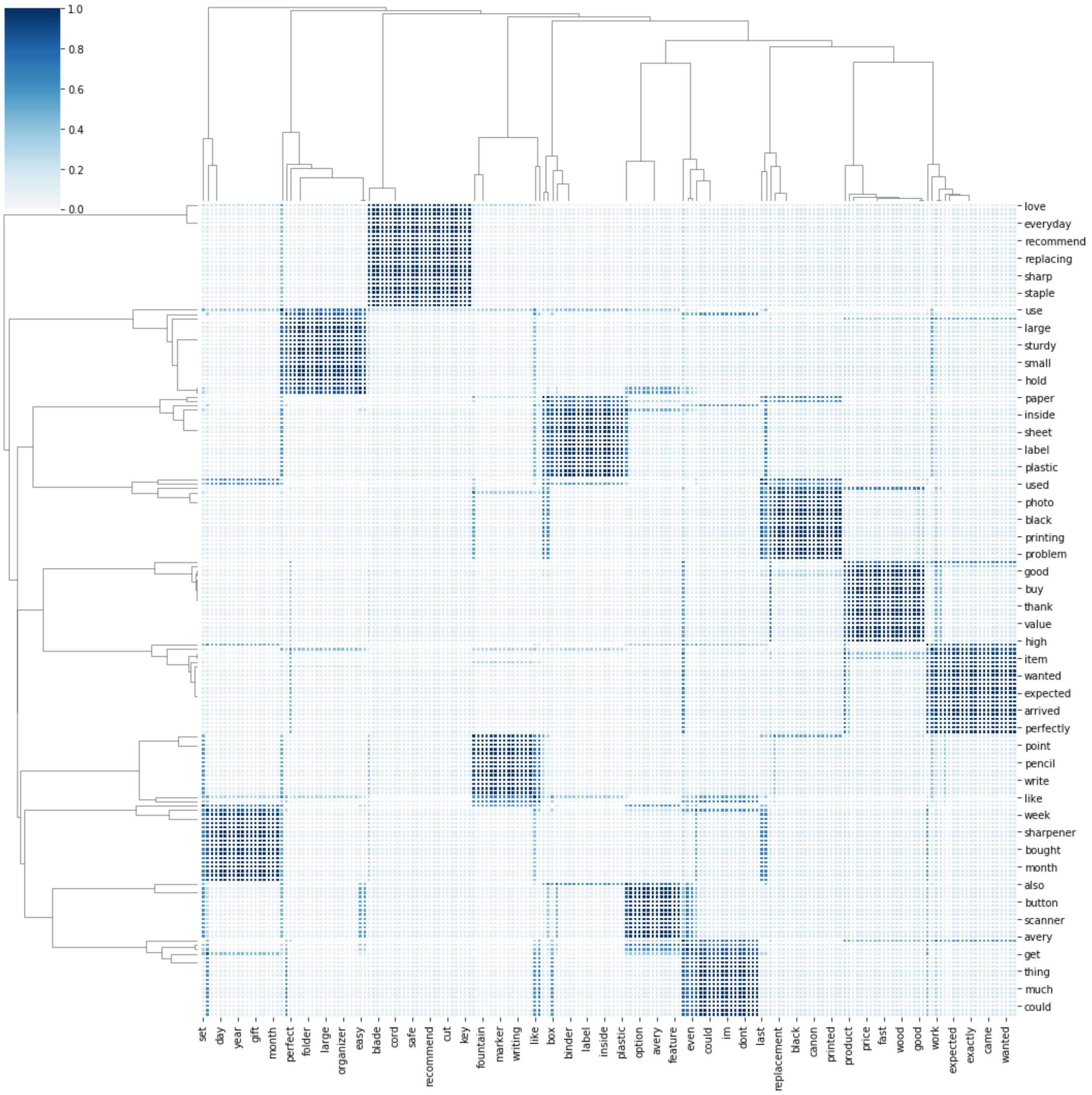
3.2.3 Reminding Tips such as Cards, Paper and Labels are Popular

From the topic 9: reminding tips are popular. When doing marketing campaign, it might be a good idea to give some small tips products as present. Because of the low price and the popularity of them, the campaign would not cost a lot but might get good feedback.

Table 3: Topics and Key Words of Reviews Text

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8
great and cheap	great	good	price	product	quality	work	excellent	buy
fit	fit	size	perfect	use	great	small	easy	hold
staple & shredder	love	HP	recommend	highly	key	staple	shredder	recommended
work	work	well	expected	exactly	nice	needed	described	worked
don’t like	one	would	get	like	don’t	I’m	time	back
printer	printer	ink	cartridge	print	paper	printing	toner	color
phone	one	year	day	love	month	planner	phone	use
pen ink color	pen	color	pencil	ink	write	love	like	writing
reminding tips	card	paper	label	page	tape	cover	use	binder
easily use	use	easy	need	time	also	feature	set	very

Topic Clusters and their Relationship with Key Words



3.3 Topic Modelling on Low-rating Reviews Text

The previous analysis is more about marketing strategy & trend, which can help amazon to develop related sellers. **In this session, concise improvement suggestions for specific products will be made.**

The research conducts LDA Topic modeling on low-rating reviews text (rating = 1) to get product optimization direction. After testing, it performs the best when topic number is 8. The 8 topics and their key words are shown in Table 4. Topic clusters and their relationship with key words are shown in the picture Low Rating Topic Clusters and their Relationship with Key Words. Conclusions about product optimization can be drawn as below:

3.3.1 Connect with the HP printer seller or enterprise for improving its cartridge & ink

From the topic 1: HP printer cartridge & ink, it is possible that HP printers' cartridge or ink is defective, so it's better to connect with the sellers or the enterprise to improve their products.

3.3.2 Try to Lower the Scanner Price and to Improve the Product Operation

From the topic 3: scanner waste money and Topic Clusters and their Relationship with Key Words, it is possible that consumers think scanners are too expensive; and scanners may waste time because of their operation trouble. Thus, trying to lower the price and to improve the product operation might be necessary.

3.3.3 Test if there is service call mistake, if there is any, deal with it

From the topic 4: service call, it is possible that the service calls are not smooth or the service is poor. Communicating with related departments, or doing survey to consumers, or dig into the reviews text related to call service, can be used to make sure whether the call service is good or not. If there is any service call mistake, dealing with it would help Amazon to avoid low rating.

3.3.4 Test if pen ink color is right and try to make it better

From the topic 5: pen ink color, it is possible that many people think the ink color of pen is not right. Amazon product operation team can do a research on whether ink has right color and why some of them has no right color.

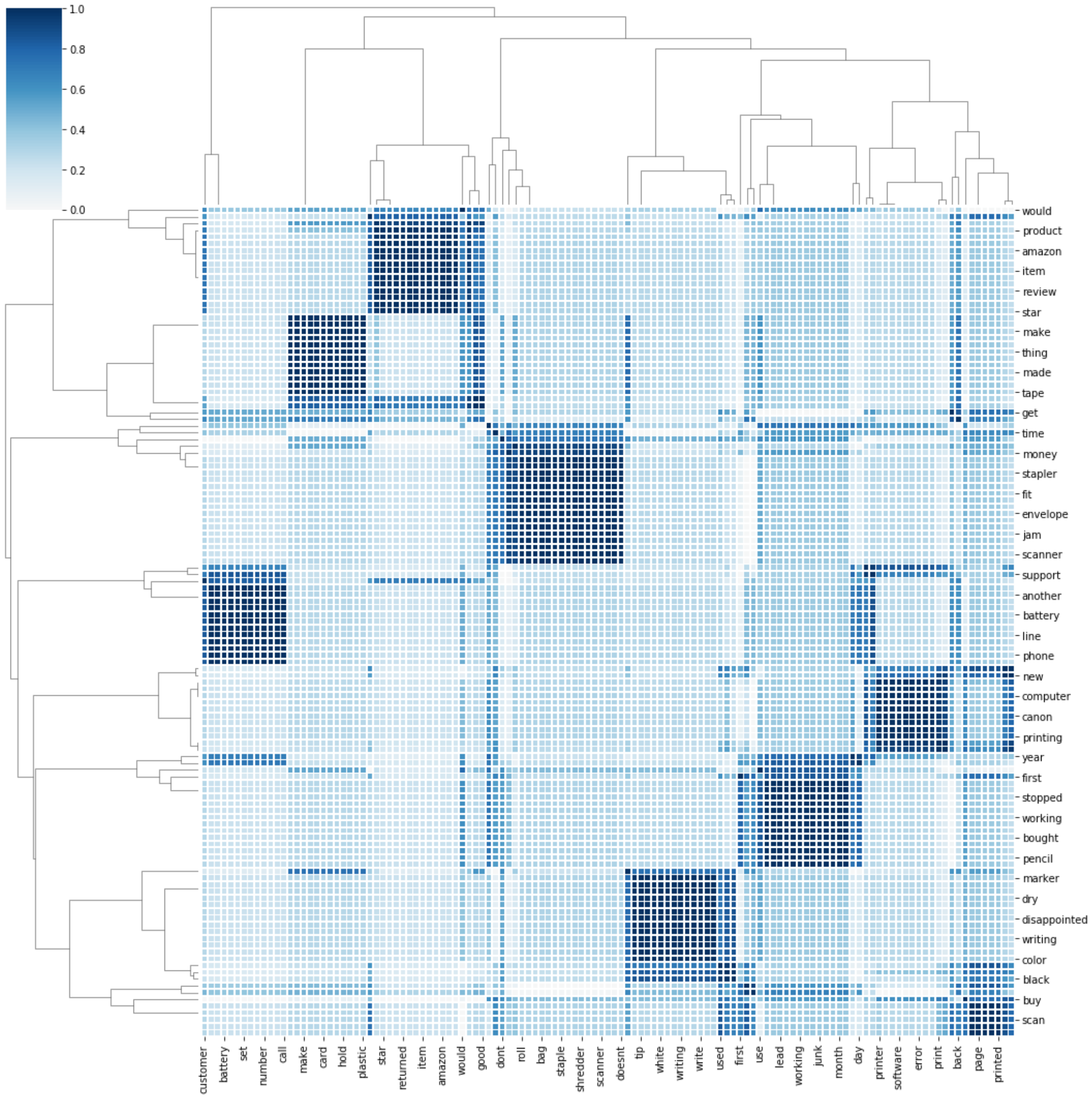
3.3.5 Connect with pencil producer to improve pencil quality, meanwhile improve the quality inspection of pencil products

From the topic 6: pencil broke quickly, it is possible that pencil quality is bad. So connecting with pencil producer to improve pencil quality is good for pencil selling. Also, quality inspection towards pencil merchandises can make sure less junk pencil not be sent to customers.

Table 4: Topics and Key Words of Low-rating Reviews Text

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
HP printer cartridge & ink	printer	print	cartridge	HP	work	ink
cartridge & ink	cartridge	ink	one	page	time	toner
scanner waste money	scanner	paper	money	work	waste	sheet
service call	phone	call	one	service	work	number
pen ink color	pen	ink	color	black	write	like
pencil broke quickly	pencil	month	work	junk	one	use
paper	one	like	paper	would	use	back
product quality	product	label	would	quality	review	amazon

Low Rating Topic Clusters and their Relationship with Key Words



4. Data Insights and Conclusions

The research question is **what rating performance, text topics are of products**. The research made exploratory analysis and conducts LDA topic modelling to determine the review text topics of each product. All these are useful for Amazon to apply in the real business position.

By rating performance analysis, products to improve is as follow: products in categories of Amazon Devices, Software and Cellphones & Accessories are ones that Amazon should contribute to improve.

By hot words analysis, review words are as follow: 1. Overall, words are common and positive; 2. For words with only rating 1, there is little positive word. Instead, negative words such as waste, disappointed appear.

By text topic analysis, the following suggestions is about **marketing growth strategy**: 1. Cost Performance of Products is an Important Aspect for Users to Consider; 2. Staples & Shredders Provided by HP are Very Popular; 3. Reminding Tips such as Cards, Paper and Labels are Popular

The following **suggestions for specific products** might be helpful to avoid low rating: 1. Connect with the HP printer seller or enterprise for improving its cartridge & ink; 2. Try to Lower the Scanner Price and to Improve the Product Operation; 3. Test if there is service call mistake, if there is any, deal with it; 4. Test if pen ink color is right and try to make it better; 5. Connect with pencil producer to improve pencil quality, meanwhile improve the quality inspection of pencil products