Erasmus
School of
Economics

TEXT ANALYTICS FOR MARKETING

Assignment 1

EBD2 – Team 6

Kaan Güner Alkan – 575313

Narmin Qarazada – 428066

Victor de Andrade – 427707

Pham Anh Tuan – 532458

Stage 1: Data description and preparation

E-commerce fashion companies require close attention to consumers’ tastes and feedback on their design. Therefore, text mining is of tremendous importance to gain insight into the consumers’ experience while wearing its products, how to leverage their strengths, and improve their weaknesses through their feedbacks. This report aims to gain more insights from the company’s customer reviews to understand their experience with the products and how to improve not only their marketing strategy but also the quality of the products.

The data set is called Women’s Clothing E-Commerce from Kaggle. It is real commercial data that was anonymized by replacing the company names with the word “retailer”. There are 23486 reviews in this data with 5 variables namely Clothing ID, Age (of the reviewer), Review Text, Rating (on a scale from 1 Worst to 5 Best), and Department Name which contains 6 levels (Dresses, Tops, Intimate, Jacket, Bottoms, Trend).

First, blank reviews and reviews with unknown department names were removed from the data set. This left the Rating level 1,2,3,4,and 5 with the frequency of 821, 1549, 2823, 4908, 12527 respectively. Further investigation showed that there are many reviews per Clothing ID, which restricted the algorithm from running correctly and also created a potential bias toward the Happy reviews with Rating equals to 5 in the later term frequency analysis. Therefore, with each Clothing ID, only the minimum Rating was kept in the data set. This created a data set with 1172 reviews, with the frequency of Rating level 1,2,3,4, and 5 were 236, 150, 191, 233, 362 respectively. During the analysis, reviews with Rating of 1 and 2 were considered as “Unhappy” reviews whereas reviews with Rating of 5 were considered as “Happy” reviews. Then the emoticon symbols used in the reviews were converted into their corresponding texts. This data set was used for sentiment analysis. Another version of this data set was stemmed, removed stop words, and punctuation.

Stage 2: Multi-dimensional Scaling (MDS) – Principle Component Analysis (PCA)

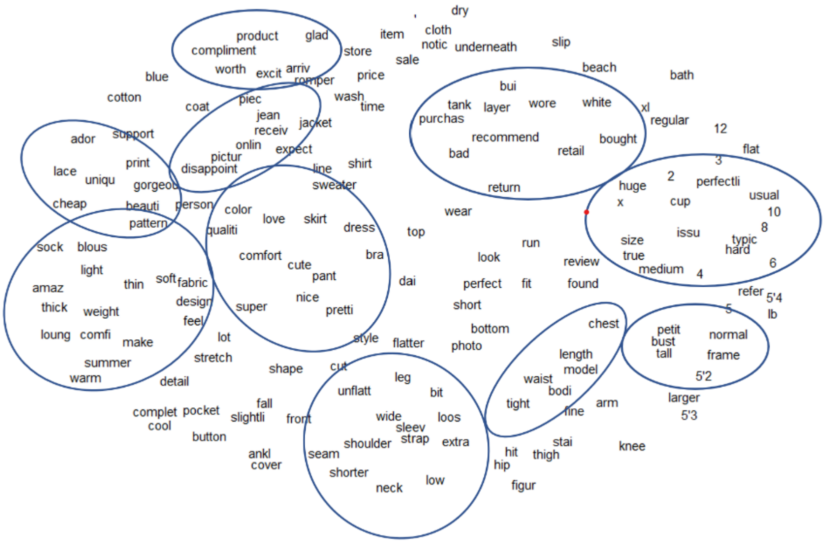


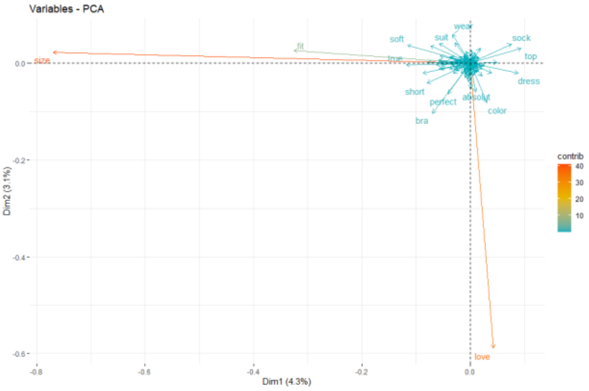
Figure 1: MDS of neighboring words

As can be seen in the figure on the left, the words “sweater”, “dress”, “skirt”, “top”, “bra” and “pants” were often described with the words “cute”, “pretty”, “nice” and “comfort”, which were positive. In addition, ‘fabric’ and ‘design’ were mentioned together with positive adjectives such as “thic”’, “comf”’, “sof”’, “thin”, and “amaz”. The topic “print” and “pattern” were close to favorable adjectives such as “beauty”, “gorgeous”, “uniqu”. The words “excited”, “glad”, “worthy” were close to when ordering or receiving (“arrived”) a product. For a group of products which were related to each other, such as swimming or beach clothing. The features of these types of clothing were then related to the words “white”, “xl”, “regular” and “recommended”. The features of a “cup” “size”. The words “true”, “perfectli”, “huge” and “medium” indicated the cup size. But the words “issue”, “hard” and “usual” could describe the features of how the right cup size was searched and found. In the next circle, we could see body types or measurements which can be related to each other when searching for clothes. These are features of certain body parts.

However, features of certain parts of the clothing, such as “leg”, “sleeve”, “shoulder”, “neck”, “seam”, and “strap” can also be determined as “loose”, “low”, “shorter” and “unflattering” which reflected negative feedback. In addition, the topic “disappoint” was associated with “jacket”, “jean”, “online”, “picture”, “receive”, and “expect”. Besides, “chest”, “waist”, “bodi” were close to “tight” which might be negative feedback.

To interpret our data visually the variables-PCA plot in Figure 2 shows the most important words per dimension. This is a rotated version of the original plot. We have used a varimax rotation to change our axis in the plot in order to improve the interpretability. After the varimax rotations, the same information is summarized in a more meaningful way. There is now a strong connection between the small set of variables. Colors in the plot represent the amount of variation, explained by those two dimensions. Dimension 1 is best represented by “love” and dimension 2 is best represented by “size” and “fit”. Hence, dimension 1 could be labeled as the customers who “love” the purchased products and dimension 2 could be labeled as the customers who review the “size” and “fit” of the purchased products. In Figure 4 can be seen how strongly a review is connected to dimension 1 or dimension 2. Review number 240 is, according to Figure 3, strongly connected with dimension 2 and hence mostly focused on reviewing the size and fit. When viewing the customers review 240, she indeed writes how well the swimsuit fits and makes her confident about her size after her pregnancy.

Finally, MDS allowed groups of words to represent the topics and themes of all the reviews on 2 dimensions, therefore it could provide a more general views of both praises and complaints at once. Whereas, PCA allowed analysts to summarize and investigate specific reviews. However, PCA produced quite many graphs to provide sufficient overall insight. Therefore, in this report, MDS provided better visualization and better served the purpose of this report give the page limit.



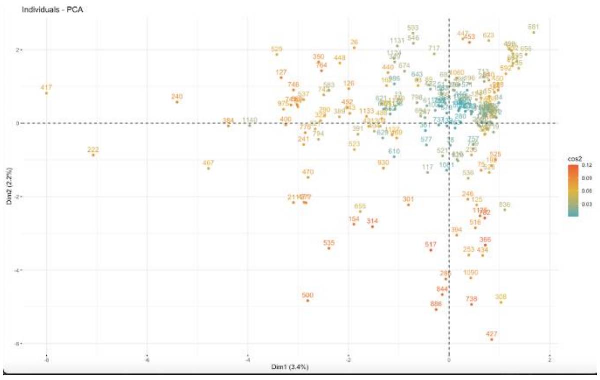
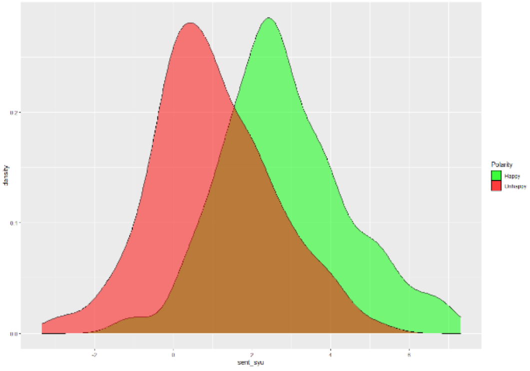


Figure 2: Dim 1-2 of top 200 words for Happy reviews

Figure 3: Dim 1-2 of Individuals

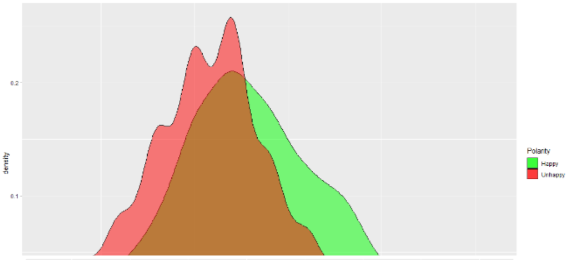
Stage 3: Sentiment analysis

For general reviews, although there are differences in the variance of polarity scores between Bing, Syuzhet, and NRC dictionary, the Happy score has a higher mean therefore its distribution is more on the right of the graph in contrast to the Unhappy score whose distribution of polarity score is on the left due to its lower mean.



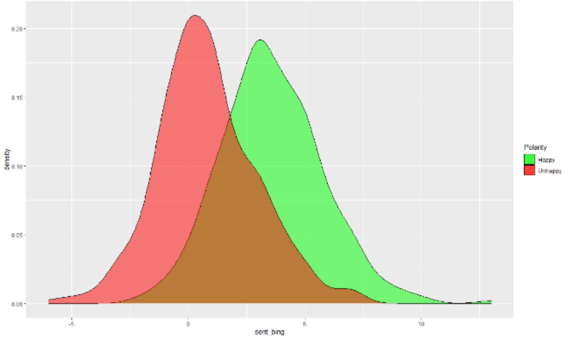
Sentiment analysis from Syuzhet dictionary provided the score of Happy reviews ranging from -1.3 to 7.3 with the mean of 2.92. Whereas, in terms of Unhappy reviews, the polarity score ranged from -3.35 to 5.55 with the mean of 0.984. Therefore, the mean of Happy reviews’ scores was higher than Unhappy reviews’ scores.

Figure 4: Sentiment of Syuzhet dictionary



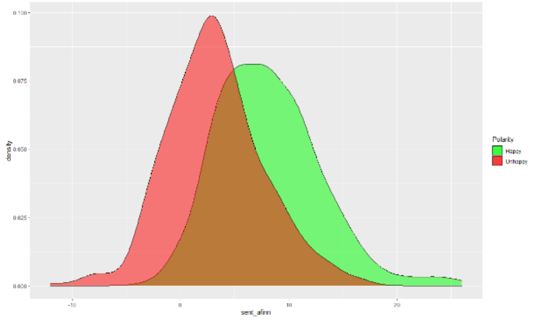
Sentiment analysis from the NRC dictionary provided the score of Happy reviews ranging from -4 to 8 with the mean of around 1.61. Whereas, in terms of Unhappy reviews, the polarity score was between the range of -4 and 7 with the mean of 0.46.

Figure 5: Sentiment of NRC dictionary



Sentiment analysis derived from the Bing dictionary provided the score from Happy reviews ranging from -2 to 13 with the mean of around 3.6. In contrast, for the Unhappy reviews, the polarity score was from -6 to 7 and the mean was 0.8.

Figure 6: Sentiment of Bing Dictionary



The sentiment of reviews based on the AFINN dictionary provided the polarity score of Happy reviews ranging from -3 to 26 with the mean of 8.1. On the contrary, the polarity scores of Unhappy reviews were between -12 to 17 with a mean of 3.2.

Figure 7: Sentiment of AFINN dictionary

In the sentiments derived from all the four dictionaries, there is a large overlap between the Unhappy and Happy polarity scores. Thus, the sentiment analysis per review is not reliable to predict the rating of the reviews to some extent. Therefore, we have to look at keywords in a sentence of happy or unhappy review to find out what the customer complained or praised about. In this way, we can know what bad features of the hotel should improve.

With regards to sentiment analysis per sentence, the sentiment analysis was conducted at the department class level to find more insight into the good and bad features of each department class. In general, the majority of favorable feedbacks associated with the nouns such as “fabric”, “design”, “size”, “material”, “online” and “retailers”. On the contrary, the majority of negative nouns were “itchy”, “scratchy”, “bust”, “mistake”, “shame”, “thread”, “wrinkle” and “snag”.

Compared with no valence shifters, negations and amplifications make the variance of polarity score more volatile and pull the mean of the score distribution toward 0. In another word, the polarity score with negators and amplifications were lower. Unlike the case of polarity score without valence shifter, the Happy sentiment score spread out in both directions (far left and even exceed Unhappy reviews’ minimum polarity score) which might reduce the ability to predict the outcome of the reviews based on polarity score. Negation reverses the intent of negative and positive words, whereas, amplification intensifies the positive or negative meaning of the words. In default, the polarity value of the amplifier is 0.8, the polarized term is 1 or -1, the neutral term is 0. After summing up the total polarity score, it is divided by the square root of the total number of words to account for the polarity term density. The idea behind this is that densely packed polarized words, negators, amplifiers imply stronger polarity meaning.

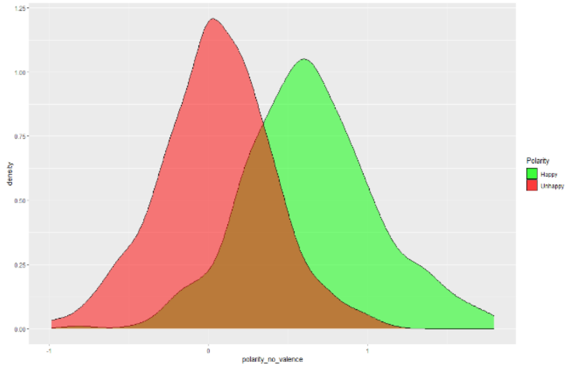


Figure 8: Polarity score without valence shifter

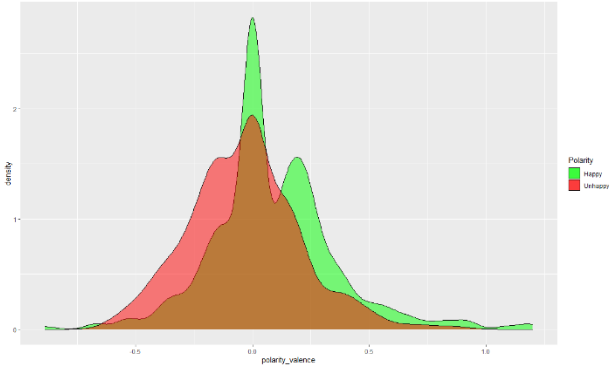


Figure 9: Polarity score with valence shifter

Stage 4: Topic Modelling

Since LDA is a soft clustering method terms can belong to multiple topics. That is why the number of topics K must be fixed. As a measure of fit we use the Perplexity. The formula for perplexity is defined as:

$$Perplexity(Dtest) = \exp\left(-\frac{\sum_{d=1}^{M} \log p\,(wd)}{\sum_{d=1}^{M} Nd}\right)$$

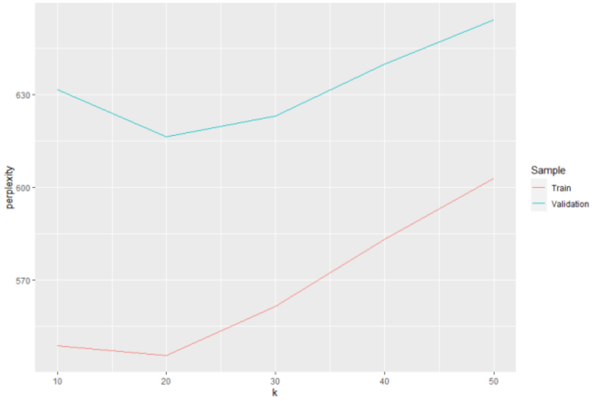


Figure 10: Perplexity to Number of Topics

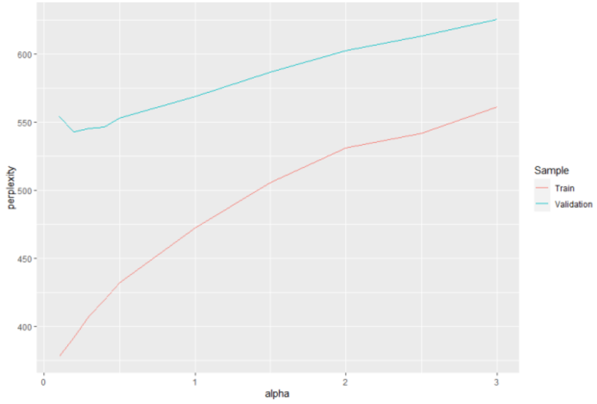


Figure 11: Perplexity to Alpha

Following the formula for Perplexity, Blei, Ng and Jordan (2003)[1] establish that a lower perplexity score indicates better generalization performance. Figure 10 shows that the Perplexity of the validation set is lowest at K = 20. Thus, the amount of topics is set to be 20. After finding K, we again use the Perplexity to find the lowest alpha which is topic sparsity in Figure 11 (a=0.2), since The Gibbs method cannot estimate alpha.

Topic 2: Sweaters, Coats, Fall, Warm, Winter, Layer. This topic relates to clothes worn in colder weather during the fall and winter seasons. As the temperatures decline in these seasons, warmer clothes such as sweaters and coats will be purchased more. Coats and sweaters are also used as layers during this season, which is also reflected in topic 2.

Topic 3: Size, Fit, Waist, Skirt, Pant, Hip. This topic relates to the sizes and fits of the bottom parts of the clothing. Size, fit and waist relate to the measurements of the clothing. Furthermore, the garments mentioned relate to the lower part of the body, the legs. Often these garments are measured in both waist and length (w/l).

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Topic 10: Size, small, fit, unusual, medium, large. This topic relates to the size of the garments. We find that size small has the largest probability, and the sizes medium and large come second and third. Also, unusual is mentioned. An interpretation could be that the size small has an unusual fit, resulting in customers leaving a comment on the size of their garments.

Topic 13: Shirt, color, white, blue, black, red. For topic 13, shirt has the highest probability. In addition, the colors of the shirt have a high probability: white, blue, black and red. This topic discusses the shirts bought and their colors.

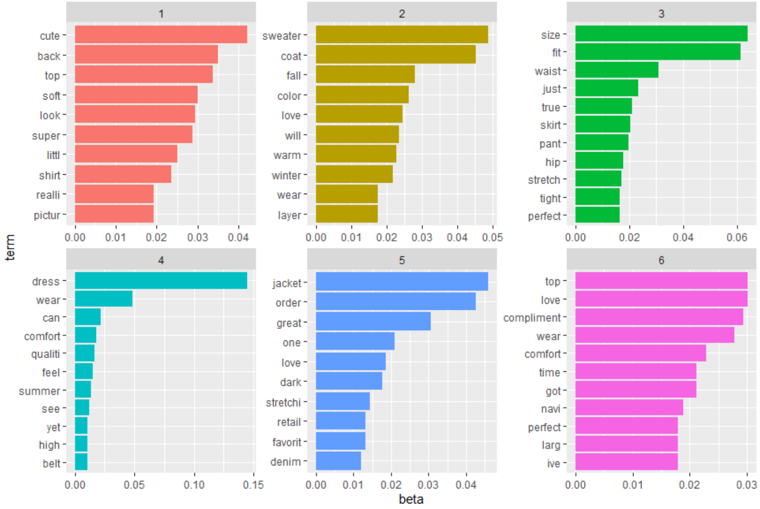
Topic 17: Dress, fabric, fall, knee, skirt, pattern, back, fall. For topic 17 we find a remarkably high probability for dress. We ascertain this topic relates to dresses, as all other words directly relate to dresses. The fabric, knee, skirt, pattern and back of the dress have a high probability for topic 17. Fall also has a high probability for this topic, which is sensible as dresses are often worn in that season.

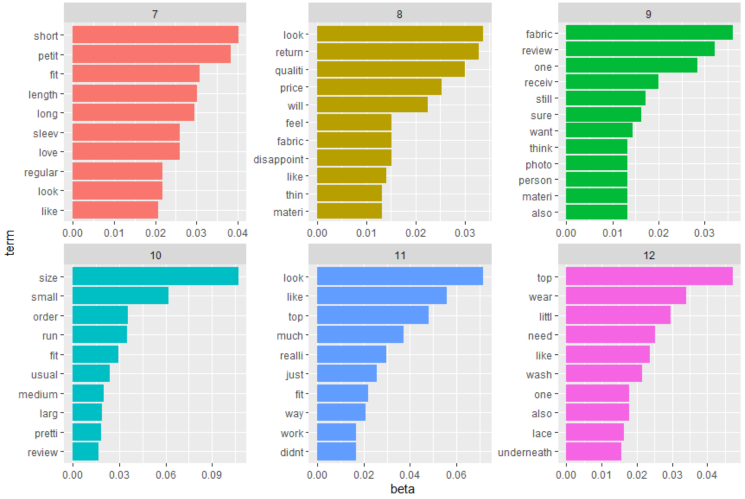
Conclusion

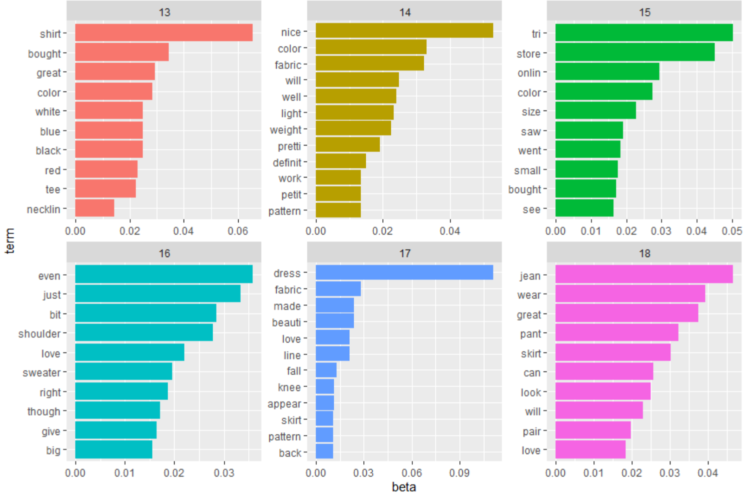
The company receives favorable feedback about the material, design, size, and fabric of the apparel products. These are features that the company should cement as its strengths and emphasize them in their marketing campaign. Therefore, the message in the campaigns should stress on good material, trending designs, soft fabric. In addition, customer also had good experience of online shopping and retailing. On the contrary, there were feedback that the clothes made customers itchy, scratchy, and wrinkle which might be due to the clothes were tight or allergy with the material. Therefore, the manufacturer should use safer, less allergic cloth. In addition, the product also bust, and thread was also mentioned in negative reviews. It might be that the thread was not strong enough, which made the clothes bust. Thus, the company should use more durable thread to improve the overall structure of the clothes. Moreover, the company should also provide tailor service to adjust the size of the shoulder, strap, sleeves to fit better fit the customers. The reviews were mostly made by customers from 18 to 45 years old, therefore, the recommendations and insights from this report might only be applied to these customer segments. In addition, the information about time and market region of the review should be provided to improve the external validity of the report.

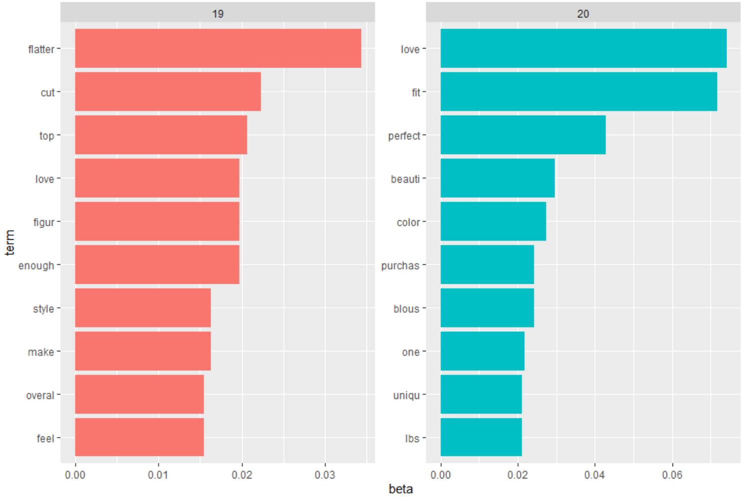
Appendix

Figure 12: 20 Topics Found in LDA




7
short
petit
fit
length
long
sleev
love
regular
look
like
0.00 0.01 0.02 0.03 0.04
8
look
return
qualiti
price
will
feel
fabric
disappoint
like
thin
materi
0.00 0.01 0.02 0.03
9
fabric
review
one
receiv
still
sure
want
think
photo
person
materi
also
0.00 0.01 0.02 0.03
10
size
small
order
run
fit
usual
medium
larg
pretti
review
0.00 0.03 0.06 0.09
11
look
like
top
much
realli
just
fit
way
work
didnt
0.00 0.02 0.04 0.06
12
top
wear
litti
need
like
wash
one
also
lace
underneath
0.00 0.01 0.02 0.03 0.04
term
beta


13
shirt
bought
great
color
white
blue
black
red
tee
necklin
0.00 0.02 0.04 0.06
14
nice
color
fabric
will
well
light
weight
pretti
definit
work
petit
pattern
0.00 0.02 0.04
15
tri
store
onlin
color
size
saw
went
small
bought
see
0.00 0.01 0.02 0.03 0.04 0.05
16
even
just
bit
shoulder
love
sweater
right
though
give
big
0.00 0.01 0.02 0.03
17
dress
fabric
made
beauti
love
line
fall
knee
appear
skirt
pattern
back
0.00 0.03 0.06 0.09
18
jean
wear
great
pant
skirt
can
look
will
pair
love
0.00 0.01 0.02 0.03 0.04
term
beta


19
flatter
cut
top
love
figur
enough
style
make
overal
feel
0.00 0.01 0.02 0.03
20
love
fit
perfect
beauti
color
purchas
blous
one
uniqu
lbs
0.00 0.02 0.04 0.06
term
beta