CS 386: Project 3 Report **Nitish Shah** (160030005)
October 27, 2018

1 Introduction and Overview

Customer reviews are a wealth of information for the sellers and buyers. Reviews can pinpoint exactly what the customer liked or didn't like about a product. Analysing text reviews of a product generally helps in understanding two things, what qualities of products (here clothing) a person likes or dislikes and what qualities of a cloth the people, collectively, like or dislike. Question 1 to 4 aims to analyze the former and Question 5 attempts to answer the latter.

Chosen problem for question 5: A lot of times, a popular product may have a large number of lengthy reviews, which makes it impossible to read and act upon all the information in all the reviews. We try to get the features for a cloth from the reviews. In other words, we try to extract informative sentences about a cloth from its reviews.

2 Methods

Question 1 and 2 involves processing the reviews to get important keywords for each review by removing punctuation marks, performing a 1 edit distance naive(no use of language models) spell-check, removing stop words and stemming or lemmatizing the words to get the root of each word. Question 3 involves making a vocabulary from all the reviews, building a term frequency (TF) and term frequency-inverse document frequency (TFIDF) matrix for the reviews and performing Latent Semantic Analysis (LSA) on the TFIDF matrix to represent reviews and vocabulary in a new reduced dimension. In Question 4 both of the above TFIDF and LSA are used for information retrieval (IR) and their results are compared. In Question 5, the method used is:

1. Select a Clothing ID, all the operations will be done on the reviews of this one cloth

2. Each review is cleaned as done in question 1 and 2 and each review is furthur seperated into sentences. Part Of Speech(POS) is determined for each word for these sentences and each sentence and the POS of each word in these sentences is stored.

3. The features we are looking for are probable to be nouns, so n (for each n in a predefined list e.g. [1,2,3]) consecutive nouns (now called a noun group) are taken from each sentence, with a maximum number of allowed words between these nouns (called the distance of the noun group) in the original sentence

4. Every other review sentence (target) is then checked to see if the noun group is present in it. The matching is order independent, so the noun group [dress, size] would match both ‘size of dress’ and ‘dress size’. There is a maximum number of extra words allowed in between the terms from the noun group in the target (called the distance). so if that threshold is 1, then [size, dress] will match with ‘dress size’ in a sentence (with distance 0) but ‘size of this pretty dress’ (distance 3) won’t be consisered a match.

5. Thus, each such feature is extracted from each sentence and its frequency is counted and stored

6. These features are then sorted based on their frequencies and the review sentences these features were found in along with its frequency and a sentiment measure for each sentence is stored in a file

3 Analysis of Results

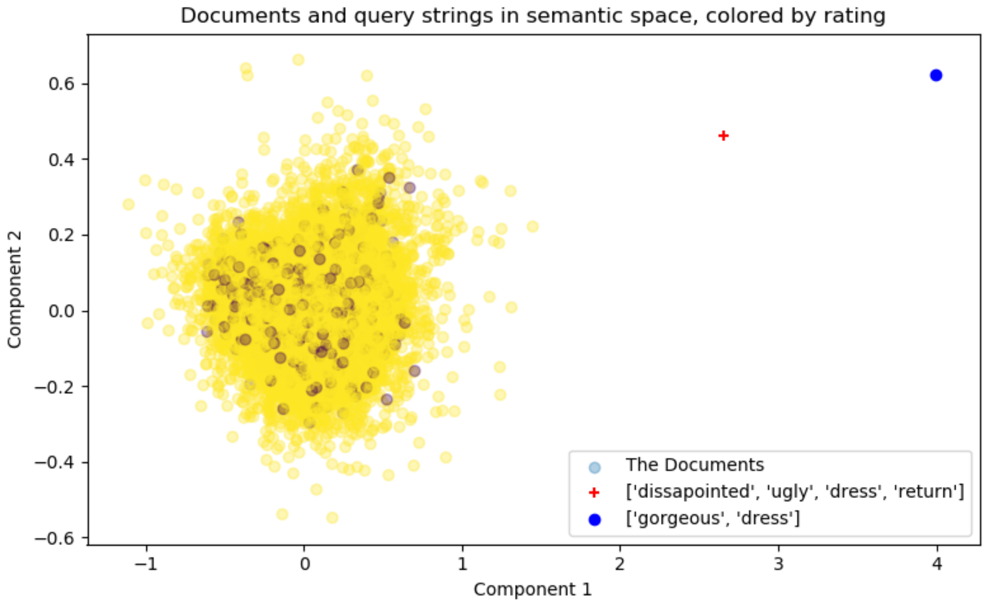

Figure 1: Documents in the Semantic Space, showing first 2 principle components, Yellow: reviews rated less than 3, Red: reviews rated more than 3

Question 4: For query **’dissapointed and ugly dress, i am returning it’**
Top 2 matched documents using **TFIDF** IR:
1. Similarity: 0.42 Review: Ugly! : Too much material everywhere. i will return it asap. this is super-ugly.
2. Similarity: 0.38, Review: Dissapointed:(: The material looked much nicer in the online picture.

And top 2 matched documents using **LSA** IR:
1. Similarity: 0.96, Review: Dress for the younger girls : i sort of looked like a little girl in this dress. it is very pretty but more for the younger girls not for the over 50 group.
2. Similarity: 0.95, Review: : I disagree with the most recent review! they are a jean style with a patterning on them a texture. great colors!

As can be seen from above, The results of TFIDF are reasonable and the reviews are similar to the query. The results of IR by LSA are not very satisfactory. This can also be seen in this figure 1, where reviews with good and bad ratings seem to be spread symetricaly around zero. This is because LSA groups the reviews by some combination of words, and understanding exactly what these different dimentions or concepts represent is often very difficult.

Question 5: Some features for product ID 1060 the program identifies are (out of total 25): well fit, perfect length, super comfy, side pocket, regular size, hip waist, type body. Some of these features are very useful and summarize the product well or give important information (14 out of 25) and some reviews (9 out of 25) give useless information. Manually reading all the reviews and getting all the features isn't possible, but the results seem promising. With some more feature pruning (for example: reject noun groups which appear in sentences with other different noun groups) and with a better POS tagger, this algorithm will give good results.