Topic Analysis of Review Data.

Dataset and EDA: The dataset provided is customer reviews for Lenovo K8 Note mobile phone as provided by customers on ecommerce website amazon.com. There are total of 14675 reviews. Along with the reviews, a sentiment tag is also provided in the form of 0 and 1, where 0 stands for negative sentiment with rating 1 or 2 while 1 stands for positive sentiment with rating 4 or 5. (can be verified by taking a sample from each and looking for negative words). The distribution is nearly balanced with 7712 – negative reviews and 6963 - positive reviews, although indicating an inclination towards the negative ones. There are no null values in the dataset. A further analysis into the length of the reviews, it was found that on average people who wrote negative reviews were more elaborate than the positive ones.

Data cleaning and normalization: Here, instead of NLTK, Spacy library was used with the language model 'en_core_web_sm' due to inaccurate pos tagging. All the tokens were converted to lower case and stopwords were removed using default stopwords collection of spacy. Only those tokens were included in the corpus which were noun and had a character length of greater than 2. All the punctuations were removed using regex. The dataset was then vectorized using TfidfVectorizer allowing maximum occurrence in 50% of documents and requiring minimum occurrence in atleast 1 percent of documents. The resulting vectorized corpus had a shape of (14675,72). The vectorized corpus is a scipy sparse matrix which was then converted to a numpy array. Then the array was separated into positive and negative reviews for further analysis. To get feature importance across all documents, the tfidf scores were summed along vertical axis and were paired with the feature names. Now, the top words for positive and negative reviews could be plotted.

Building LDA Model using Gensim: Gensim’s LDA Model was used to build a topic model on the given corpus for 12 topics, initially. For analysis, each document was tagged with the highest topic probability. The TransformedCorpus gave a list of tuples which paired each topic with topic probability for each document, which was then sorted to get the dominant topic for each document. Using show_topic, the keywords for each topic were obtained. All of these were combined to create a new dataframe. For further analysis, 3 documents which had highest topic probability for a given topic were found out, and then names were provided for the topics found out earlier for context setting. This model had coherence score of 0.526. To find the optimum model, the same gensim model was fit on all topic numbers from 2 to 20 and the graph was plotted for their coherence score. The best result was obtained for 5 topics, with a coherence score of 0.59. Again, the most representative topics for each documents were analysed to obtain business friendly topic names.

Analysis: The topics were then assigned back to the documents dataframe. It was seen that most of the people who rated negatively talked about either battery and overheating problem or about camera quality and sound. But a large number of people who rated positively also talked about camera and sound. It seems like, there is some subjectivity in judging the quality of camera and sound, but it remains the most talked about topic, regardless of whether people rated positively or negatively. There are people who faced problems regarding battery backup, charging and overheating, although the problem is not universal in all customers.