Topic Modeling Text Analysis Response

For this assignment, I took the JSTOR sample data set entitled “marketingtxt”, and topic modeled it with first 10, then 30, and finally 60 topics. After removing stop words and reading through each topic model, I chose to focus on the 30 topic one. In this response, I will first explain the contents of the data set and what is significant about them, why I chose the number of topics to be 60, and explain the limitations and possibilities of topic modeling.

The contents of the 60 topic topic model of the marketingtxt data set shows that these readings were highly focused on brands, consumer activity and choice, marketing through electronic media, and data analysis. I was able to infer this from inspecting the topic model and finding most of the topics to include words such as “market”, “brands”, “consumer”, “choice”, “television”, “media”, “data”, “analysis”, and “models”. This information makes me realize multiple things. One being that these publications are most likely from the late 1900s to early 2000s, since they refer to electronics such as television and the internet, which was not invented until around that time. Another thing the data shows me is that this corpus is focused heavily on the interaction between consumers and the market. More specifically, what the consumers choices are, what influences them, and how they affect the conditions of the brands in certain markets. If I would wish to find out more from this data set, one thing I would do is turn the data into some sort of visualization (Brett, JDH, 2012). Past topic models that we have studied have almost always shown the data found in some sort of graph or visualization. By doing this, it can be a lot easier to spot certain patterns or significant information that I may not have found before by just looking at the words of the topics themselves. Another way I could potentially discover more information is by increasing the number of topics I run through the topic model. This could lead to a discovery of new terms or topics that shed light on the context and significance of this corpus.

This is why I chose the number of topics to be 60, instead of 30 or 10. When I increased the number of topics to 60, the findings highlighted more words and topics that showed additional significance to the data. For instance, it shows much more emphasis on marketing experiments, simulations, and their findings. The increase of topics also highlights certain geographical areas, such as certain universities and states around the United States. This tells me that this corpus was focused on marketing and consumers in the United states primarily, and not abroad elsewhere. Both of these discoveries also points out the fact that this corpus is most likely concerned with marketing on a broad, national level, not just for individual areas or consumers. Even though I found increasing the number of topics to be helpful for me, this result will not be the same for all other corpora. I will explain why this is the case, along with other limitations and advantages to topic modeling.

When the number of topics is increased in topic modeling, you can find that there is a possibility of an increase in outlier words that do not really relate to the topic. This can be disadvantageous because these added outlier words do not contribute to the context of the main topic of the corpus. This in turn can make it more confusing to understanding the significance and meaning behind the data. Another limitation of topic modeling is because there are so many pieces usually in a corpus, it can be difficult sometimes to find certain specifics of some of those pieces. Since the topic model just pairs individual words that occur near each other often together, it can be challenging to decipher the connection between these pairing of words and the context or meaning of the data. However, the abundance of pieces used in topic modeling can also be an advantage.

It can be an advantage because the more pieces there are in a certain corpus, the more information you acquire to help you with finding what you are looking for. Through distant reading, major themes and correlations are easier to spot since the amount of information is so great. When only a few pieces are used, it an be difficult to discover a connection or pattern over a broad time period or scope of readings. Another advantage of topic modeling, is the patterns and connections between different areas of information that can be found. For instance, past topic models we have looked at have found connections between the publishing of certain novels and critical historical periods in major countries (Moretti, Verso, 2007). This showed why certain books or more/less books were published during times of war, revolution, and high economic prosperity. In general, topic modeling is very advantageous tool to anyone studying the history of literature publication and studies, and can help one discover critical patterns and connections in those areas.