Thorsteinn Nielsen
N10717700

Assignment 8

LDA topic analysis using LDAvis in R

Visualization link: goo.gl/1xMn53

GitHub repository link: https://github.com/klakinielsen/Projects

Database

Amazon product reviews for the category of Patio, Lawn and Garden with 13,272 reviews
Taken from: http://jmcauley.ucsd.edu/data/amazon/ (5-core)

By fitting a topic model with 15 topics and using the visualization tool LDAvis, we get a very comprehensive overview of the different topics within the data. The purpose of this case is to explore some of the topics and analyze their relationships.

Visualization

This histogram essentially shows the frequency of each word in the data given a topic, specifically the 30 highest occurring words for each topic. The topic occurrence (red) can also be compared to the overall occurrence (grey).

The bubble graph shows us two things. First, the size of the bubble represents the topic's proportion of the entire dataset. That is, how much of the text is dedicated to certain topics. Second, the distance between the bubbles represent the likeliness of the topics being discussed at the same time and their words being shared.

The lambda value is a very important element. It determines the weight put on the ratio of topic occurrence to overall occurrence. That is useful when the trying to understand the meaning of the topics. By altering the lambda, we can increase or decrease this weight. The lower the lambda, the higher the weight and as a result, the words with the higher ratio are ranked higher. For this model we are working with the default lambda of 0.5.

Topic Interpretation

The following conclusion is an attempt to give further understanding of certain topics.

<u>Topic 1</u>

The most occurring words, amazon and review suggest that this topic is about amazon reviews. I find this not really definitive enough and when you look at the other words on the top of that list, I'm, time., upgrade etc. I'm not able to figure out the contextual relationship. This is where changing the value of lambda might help us understand this topic. By lowering the lambda from 0.5 to 0.2 we can see that words like service and customer are no close to the top of the list. No I can confidently determine that topic 1 is about reviews on amazon's customer service.

Thorsteinn Nielsen
N10717700

Topic 2

Pulling lambda back to 0.5 we see the top words are deer, spray, product, ants and repellent, which tells us that this topic is about animal repellent spray.

Topic 3

The top words of mouse and trap are very obviously suggesting this topic is about mouse traps

Topic 4 and 5

The top words for topic 4 are trimmer, battery, cut, string and branches suggest the topic is a battery powered hedge trimmer. The top words for topic 5 are mower, grass, lawn, blower and leaves which interestingly suggests two topics, lawn mowers and leaf/lawn blowers. Now when you look at the bubbles for topic 4 and 5, they are very close to one another on the graph, suggesting these topics are commonly intertwined and they have words that are shared between them. If you hover over some of the words in topic 4 (battery, cut and branches) you see that they are also being used in topic 5. Similarly, you can hover over the words in topic 5 (grass, lawn and blower) and you see that they also appear in topic 4. This all makes sense because both topics are about certain big equipment that you use to take care of your lawn/garden.

Note that overlap of bubbles doesn't have to occur for the topics to have shared words. The distance from the center of one bubble to the center of another represents the commonality.

Further exploration of the topics is encouraged.