**Let the Machine Help Understand the Text**

**Bingqing Wang**

**STA 208 Final Project**

**Background**

Social scientific studies often involve human coders to read open-ended responses and code each response with numbers, so that follow-up statistical analysis can be employed. This type of method usually suffers several disadvantages.

1. Human coders can suffer fatigue and unreliability issues, which may result in unreliable coding for further analysis.
2. Human coders are usually cost inefficient and are harder to find compared to computers.

The present project aims at using machine learning technique to code textual responses. With minimal input and guidance, a machine can code the textual responses within seconds and with no fatigue and high reliability. The entire project consists of three part and two datasets I have collected in previous studies.

First, I will employ the first dataset to test the LDA (Latent Dirichlet Allocation) method in abstracting topics from open-ended responses. The first step is to ensure that meaningful different topics can be recognized by the machine and there are distinctive different topics in the dataset. This is an insurance policy to make sure that scholars are making meaningful inference and usage with the programming.

Second, I will use the same dataset to improve the procedure. In particular, I will use graphical illustrations to determine the number of salient topics in the corpus. And the significant words in each of the topics. In this part, meaningful interpretations of the results will also be provided, based on social scientific theories and research.

Third, I will improve the program so that it can fit various demands in social scientific studies. Due to the complexity of human behavior, social scientific studies usually have different specific demands in different areas. For example, people may declare several meanings in one responses. The last step is to improve the program to adapt to different environment.

*Research Interest*

Entertainment is a big part of human life and consumes significant amount of everyone’s time. The concept of enjoyment has always been a focus of research since people want to understand what we enjoy and why we enjoy some content over the other. The two-factor model developed by Oliver and Raney (2011) suggest that people in general enjoy content that has either hedonic or eudaimonic values. People not only like content with hedonic values to seek fun, laugh and “shallow” enjoyment, but also enjoy content with eudaimonic values to find life meanings, personal growth, or intellectual challenge. Other than the two-factor model, tremendous amount of research suggests the social function of entertainment. People enjoy entertainment with other people and the relatedness it provides. The present projects use open-ended responses collected in two studies to understand the enjoyment experiences of video games and movies: What are the functions of these contents in their point of view?

**Part I. Topic Modeling with LDA**

LDA is a generative statistical model that allows observations to be grouped into unobserved groups. When using LDA to understand documents, each document is viewed as a mixture of different topics, and the topics distribution is assumed to have a sparse Dirichlet prior, which is usually the case in terms of textual data (Blei, Ng, & Jordan, 2003). LDA method has been widely used ever since it was published.

In the first part, LDA method is employed to abstract topics from participants' responses. The dataset is from a study I have done about video game playing in families (Wang, Taylor, & Sun, 2018) and the open-ended question asked the participants:" Based on your experiences, what changes does video game bring to your families." I employed human coders to code these responses into three categories: social, hedonic, and eudaimonic functions. Part I uses LDA method to extract three topics from the responses without any guidance. The result of three topics with showing three words in each topic is as follows:  
[(0, '0.056\*"u" + 0.046\*"it" + 0.042\*"together"'), (1, '0.031\*"game" + 0.018\*"video" + 0.015\*"time"'), (2, '0.034\*"game" + 0.022\*"video" + 0.018\*"play"')]

It is not clear how topic 2 and topic 3 are different, but it is somewhat obvious that topic 1 is about people connection while the other two are about playing and games.

### Part II. Improve the topic modeling

In this part, I will try to improve the topic modeling results using various strategies. Mainly, the goal is to find an optimal number of topics in a specific corpus. In the previous part, I suspect there are three topics in the corpus, but the number is by no means determined. In traditional social science studies, there are usually two ways of determining the number of topics researchers would like to code. First, researchers can look up the theory and see how many topics theories suggest. Second, people can look at first few responses and have a brief understanding about how many topics there are in the corpus. Both techniques have serious disadvantages. Theories and practices can be far away from each other and using theory to guide the coding does little in improving the theory. The first few responses can be varied biased and far from accuracy. Thus, using machining learning to help determine the number of topics can be a very beneficial tool in social science studies. I will use the same dataset from the previous part.

In this part, I first employed LDA method to extract three topics in the dataset. Using graphical demonstration, the three topics are depicted as follows:

A screenshot of a cell phone

Description generated with high confidence

From the graph above, topic one, whatever that topic is, is insignificant in the corpus. This suggests that three topics do not fit the responses, even though in theory, there should be three topics in the responses. Next, I tried two topics and see the results as follows:

A screenshot of a social media post

Description generated with very high confidence

From the graph above, the two topics are about the same size, and although there are a lot of word overlaps between the two topics, some distinctions can be made to distinguish the two topics.

Analysis of the first corpus.

From the results above, there are two main topics in the responses. First, people enjoy playing video games and believe it is "something" in the families. Second, from the responses, video games have the bonding effect in family settings, family members share activities and feel something in "common". Therefore, we can interpret the responses and answer the question "what do video games bring to families" like the following:

Video games mainly have two functions in the family settings. First, it provides hedonism/fun to family members so that they have something to enjoy. Second, video games have social/relatedness fucntion to families, so that family members have something in common and they are bonded together.

**Part III. Applying the same technique to another dataset**

In part III, I will apply the same technique to another dataset that is newly collected. The new corpus comes from another study that investigates people's experience in consuming movies. The question asks the participants: Based on your experience, what does watching movies bring to your life. The question is open-ended, and there is no previous human coding on this corpus. In this part, I will apply the same technique and see how many topics I can extract from the responses.

Using LDA method and apply three topics, the results is shown as follows:

A screenshot of a cell phone

Description generated with high confidence

From the results above, the three topics extracted from the responses are about the equal size. The first topic can be recognized as hedonic experience, as the top words in the topic are "reduce stress" and "improve happiness". The second topic can be seen as social experience, as there are words like "people", "discuss idea" in the second topic. The third topic can be treated as eudaimonic experience, as there are words like "insight", and "problem" in the topic. The differences and distinction between these three topics are not obvious, probably due to the relative small sample size.

**Discussion and Summary**

The present project utilizes the LDA method to solicit topics from two corpuses. The first part uses LDA method to elicit topics from textual responses. The second and third part uses LDA method and improves the topic modeling on two different textual responses. The results showed that LDA has great potential in recognizing and differentiating different topics. This technique can be applied in various social scientific studies in the future and help scholars understand textual responses better.

*Results*

The topic modeling results have significant implications for entertainment studies. The first topic modeling shows that two topics are prominent when people answering the question about video game function in families. People generally believe that video games help families unite and they get a lot of fun through video games. However, when people answer the same question regarding movies, they also mention insights, inspirations, and other eudaimonic functions of movies. It seems movies have more functions than video games but when we looked deeper at the responses about video games, we also see responses like "learning" and "cooperation" in the corpus. The eudaimonic topic is not prominent but still exists. So far, it is probably safe to say that, in general, entertainment has roughly three functions in people's life:  
1) Hedonic, people look for fun in entertainment, and seek escapism, diversion, and relaxation.  
2) Eudaimonic, people seek meanings and personal growth in entertainment. They learn to grow and look for inspirations and insights in entertainment contents (Oliver & Raney, 2011).  
3) Social, people see entertainment as a social venue and look for relatedness with other people when consuming entertainment content.

*Limitations*

The present project is not without limitations. First, the results of the topic modeling are not stable. Different iterations usually provide different results, and an ideal result needs multiple trials. It is probably due to the small discrepancies between different topics and responses.  
Second, the differences between topics are not obvious. It still needs very careful human inspection, or even expert inspection to understand and label the different topics. It is probably because it is unguided machine learning, there are no guidelines on how the machine can learn the topics, but also due to the small sample size. If we can increase the sample size to about 1000, I believe the results can be a lot more stable and differentiable.

*Summary*

Overall, the present project utilizes machine learning strategy to traditional social scientific studies. Using natural human textual responses, with the help of LDA machine learning, researchers can achieve a lot more than traditional human coding. The result of three topics can inspire a lot more other studies in this area and future research can also utilize this machine learning technique to understand other textual responses in other social scientific areas of interest.

**References**

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

Oliver, M. B., & Raney, A. A. (2011). Entertainment as pleasurable and meaningful: Identifying hedonic and eudaimonic motivations for entertainment consumption. *Journal of Communication, 61*(5), 984-1004. doi:10.1111/j.1460-2466.2011.01585.x

Wang, B., Taylor, L., & Sun, Q. (2018). Families that play together stay together: Investigating family bonding through video games. *New media & society*, 1461444818767667.