# Analyzing Open-Ended Survey Questions Using Unsupervised Learning Methods

Fang Wang[1], Edward Mulrow[2]
[1] NORC at the University of Chicago, 55 East Monroe St., Chicago, IL 60603
[2] NORC at the University of Chicago, 4350 E-W Highway, Bethesda, MD 20814

## Abstract
Unsupervised learning methods such as topic modeling or k-means clustering can provide techniques for organizing, understanding and summarizing text data without using any manually labeled records as training data. It uses annotations to organize text and discover latent themes in documents without target attributes. We explore using unsupervised learning to classify open-ended survey question responses. By grouping similar responses together, we construct a class of "topics" which are described by sets of keywords and reduce the exploration of open ended text information to common categorical analysis. We analyze the open-ended survey answers in the "10 Years After 9/11 Survey" and the "National Cultural Building Survey" using the text clustering and the topic model respectively as two examples.

**Key words:** Unsupervised learning, topic model, k-means clustering, open-ended survey question, text data

## 1. Background

Open-ended survey questions are questions without any prescribed categories of answers such as "yes, no" or "A, B, C" etc. Instead, the answers are in non-categorical, non-numeric text formats, like sentences or documents. For example, the questions "Could you talk about this historical building briefly?" or "What kind of freedoms will you tell a foreigner who is new to US?" are open-ended questions. The answers can be a group of unrelated words, a few sentences or a short speech. These kinds of text answers may give the survey conductors a deeper understanding of the survey respondents. But the shortcomings are obvious; unlike categorical or numeric variables, such text answers may need a large amount of manual, labor hours to interpret and summarize. Additionally, these answers are often not publishable "as is"due to confidential reasons.

Here we use machine learning and text mining methods to automatically search, summarize and group the text data into meaningful categories. These derived categories can be easily analyzed and published like other categorical variables. This method can help reduce the large amount of manual work needed to summarize text answers, and leads to analyzing open-ended survey questions directly.

Text mining consists of methods for deriving useful information from data in text formats. The common steps of text mining are: structuring the text, deriving patterns, and interpreting the output. In more detail, researchers usually conduct the following steps to analyze an electronic archive of text-based data:

1. Start with a set of documents (a corpus)
2. Use text mining to isolate key words
3. Form an input matrix: rows are documents, columns are vocabulary, and cells are count of the number of time the word is in the document

4. Final extracted topics or descriptors of the grouped text clusters are combinations of words. Some words can turn up in more than one topic and importance of the word within a topic will vary. Subject matter experts may be needed to help interpret the topic

Comparing with widely used supervised learning models such as logistic models, the unsupervised learning models require no training data (the dependent variable Y) in the models, only the free text data from the open-ended questions (the predictors X) is needed as input file. The unsupervised models help find the hidden structure and latent topics in the unlabeled data enforcing very few assumptions on the data. We will talk about two kinds unsupervised learning models: the text clustering and topic modeling in this paper.

## 2. The 10 Years After 9/11 Survey

The first example we consider is text data from the survey "Civil Liberties and Security - 10 Years After 9/11". This survey was conducted by NORC and explores public opinion about national security and the rights that define the American way of life. The survey also looked back at the impact of the events of 9/11, and how the events have affected the way Americans live their lives today. See more at: http://www.apnorc.org/projects/Pages/Civil-Liberties-and-Security.aspx#sthash.aT1WCR6A.dpuf

The following open-ended question is analyzed in this example:

*Q5. If someone from another country were to ask you to make a list of the specific rights and freedoms you have as a resident of the United States, what would be the first thing you would put on the list?*

Two examples of the types of answers received are:

1. *Express yourself*
2. *Freedom to vote*

There were 1,087 responses to this survey question. Across all these responses, 289 unique words were used after deleting the words "freedom", "of" and other English stop words such as "and", "the", "to" etc. The distribution of number of words used in each answer is shown in Table 1.

**Table 1:** Distribution Statistics for the Number of Words Used
Answers to Question 5–The Freedom Question
10 Years After 9/11 Survey

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 1.000 | 1.000 | 1.364 | 2.000 | 8.000 |

The median of number of words used in an answer is only 1, with a mean is 1.36. The third quartile is only 2, which means that most people responded with only 1 or 2 words (excluding the words "freedom" and stop words). Thus, most of the answers are short phases, like "Freedom of X" or "Freedom of X and Y".

We tested topic models to see if the responses could be separated into a small set of topics (categories), but the topics were not easy to interpret. However, we found that k-means text clustering successfully

classified these short answers into meaningful categories. The k-means clustering methodology partitions data into 'k' subsets, where each data element is assigned to the closest cluster based on the distance of the data element from the center of the cluster. The descriptors of a cluster are a set of words used frequently in question answers. In order to use k-means clustering with text data, we need to apply a text-to-numeric transformation to our text data. This process is same as the text analytics process described in Section 1. That is, we started with a corpus, built the term matrix and counted the number of times a word used in each answer.

There is no standard rule for choosing the number of clusters. We tested different numbers and selected 6 as the final number. Table 2 lists the frequency of the words used more than twice in each of the 6 word clusters. The R tm package was used to perform the text mining. The software removes stop words and standardizes remaining words. In performing this standardization, some words may be truncated. The software does not automatically correct misspellings of words. For example, you see the word religion in cluster 2, but you also see misspellings of those words in cluster 6. In future work, we hope to correct this type of issue.

**Table 2**: Frequency of Words Used More Than Twice in a Cluster
Question 5–The Freedom Question
10 Years After 9/11 Survey

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cluster 1 | elect | repr | govern | offici | | | | | | |
| | 4 | 3 | 2 | 2 | | | | | | |
| Cluster 2 | religion | speech | arm | bear | express | practic | | | | |
| | 107 | 8 | 3 | 3 | 3 | 3 | | | | |
| Cluster 3 | speech | assembl | | | | | | | | |
| | 469 | 2 | | | | | | | | |
| Cluster 4 | vote | | | | | | | | | |
| | 68 | | | | | | | | | |
| Cluster 5 | arm | bear | choic | life | express | choo | happi | live | travel | plea |
| | 31 | 26 | 25 | 12 | 10 | 9 | 9 | 8 | 8 | 7 |
| | speak | liberti | opinion | properti | pursuit | job | sppech | amend | assembl | bare |
| | 7 | 6 | 6 | 6 | 6 | 5 | 5 | 4 | 4 | 4 |
| | bill | dont | relgion | religi | abil | believ | civil | countri | educ | hard |
| | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 |
| | pay | press | religon | safeti | tax | ammend | busi | constitut | determin | equal |
| | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 |
| | freeli | gun | health | land | law | movement | opportun | pick | pretti | regligon |
| | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | respect | rid | self | stay | talk | voic | whatev | world | | |
| | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | | |
| Cluster 6 | worship | god | plea | | | | | | | |
| | 11 | 2 | 2 | | | | | | | |

Table 3 shows how many responses are grouped into each cluster; the most frequent word is used to label the clusters in the first row.

**Table 3:** Count of Answers Assigned to Clusters
Question 5–The Freedom Question
10 Years After 9/11 Survey

| 1 elect | 2 religion | 3 speech | 4 vote | 5 arms | 6 worship |
|---|---|---|---|---|---|
| 4 | 107 | 469 | 67 | 430 | 10 |

Clusters 1 and 4 appear to indicate the same theme, as do clusters 2 and 6. So we reduces the 6 clusters into 4 categories of "Freedom of Religion", "Freedom of Speech", "Freedom to Vote" and "Freedom to Bear Arms." Table 4 provides a summary of the reduced clusters.

**Table 4:** Count of Answers Assigned to Reduced Clusters
Question 5–The Freedom Question
10 Years After 9/11 Survey

| 1 Religion | 2 Speech | 3 Vote | 4 Bearing arms |
|---|---|---|---|
| 117 (11%) | 469 (43%) | 71 (7%) | 430 (39%) |

To test the effectiveness of using our text cluster approach to automatically categorize the responses, we randomly selected 10 records and manually labeled them. Nine of these matched the text clustering categories. We repeated this check several times, and, on average, the accuracy rate is 9 out of 10.

## 3. The National Cultural Building Survey

In the second example, we analyzed data from the National Cultural Building Survey. This survey is a national study of cultural buildings in the United States. The survey interviews were structured and included both closed-ended and open-ended questions. The open-ended portions was actually an interview of the organization's executive director. The interview lasted one-hour, and were recorded and transcribed into scripts in text format. There are 84 of these open-ended interviews, and the distribution of number of words in each interview is shown in Table 5:

**Table 1:** Distribution Statistics for the Number of Words Used
Executive Director Intervies
National Cultural Building Survey

| Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|
| 30 | 1,936 | 2,912 | 2,983 | 4,072 | 7,194 |

We tested k-mean clustering with these data, but it didn't generate understandable results; however, topic modeling yielded interpretable topics. Topic modeling provides methods for automatically organizing, searching, and summarizing large electronic archives. The features of topic modeling include:

1. The topics are combinations of key words,
2. A document typically concerns multiple topics in different proportions;
3. A word can show up in different documents, importance of the word within a topic will vary.

For this analysis, we fit a Latent Dirichlet Allocation (LDA) model; estimated using Gibbs sampling methods. LDA is a commonly used topic model algorithm. As with text clustering, there is no standard way to choose the number of topics. We tested different number of topics and adopted 3 as the best number for us to understand the documents. The top 10 most frequent words (sorted from most to least frequent) are listed in Table 6. The words of institution names and geographic information are suppressed and replaced by XXX so as not to disclose confidential information.

**Table 6**: Top 10 Most Frequent Words within Each Topic
Executive Director Interviews
National Cultural Building Survey

| Topic 1 | Topic 2 | Topic 3 |
|---|---|---|
| hall | theater | exhibit |
| concert | theatr | expand |
| counti | council | history |
| opera | etc | gallery |
| orchestra | broadway | XXX |
| XXX | gonna | bond |
| ballet | black | record |
| acounstic | sale | gonna |
| campus | tax | XXX |
| resid | grand | XXX |

We can see that first topic's words are mainly about music related performances; the second topic's words are mainly related to live shows; and the third topic's words are mainly related to exhibits or galleries. So we simply refer the three topics as: Topic 1 – Music Performances, Topic 2 – Live Shows; Topic 3 – Exhibits.

The National Cultural Building Survey considered six building types: existing museums, new museums, non-resident performing art centers, resident performing art centers,[1] theaters, and university & government centers. To evaluate the topic modeling topics for categorizing buildings based on interview scripts, we ran a cross-tabulation of the six building types and the three topics. Table 7 shows the results.

[1] In general resident performing art centers (PAC) are home to community art groups. Non-resident PACs host other arts groups as renters

**Table 7:** Building Type by Topic
Executive Director Interviews
National Cultural Building Survey

| Type | Topic 1 | Topic 2 | Topic 3 |
|---|---|---|---|
| 1.Existing museums | 0 | 1 | 14 |
| 2.New museums | 0 | 0 | 9 |
| 3.Non-resident performing art centers | 3 | 18 | 5 |
| 4.Resident performing art centers | 15 | 0 | 0 |
| 5.Theaters | 0 | 12 | 0 |
| 6.University & government | 3 | 1 | 3 |

If we assume museums and university & government host exhibits (Topic 3), non-resident PAC and theaters host live shows (Topic 2) and resident PACs hold musical performances (Topic 1), then the numbers in black are matched cells and the numbers in orange are unmatched. Thirteen out of 84 (15%) cases did not match, but they may not be wrong. For example, universities or government buildings could be used as music performance buildings and so on. So the accurate rate of this model is above 85%.

## 4. Conclusions

- Text clustering and topic models work for the open-ended survey questions; they can help reduce the manual work of interpreting the text answers.
- Text clustering classifies documents into clusters; topic models develop probabilistic distribution to discover latent themes. Both adopt some distance matrix, but text clustering has a simpler theory and simpler distance matrix.
- Text clustering appears to work better for simpler data structure, like short answers or a sentence; topic models appear to work better for more complex data structure like full documents.

## Reference


- Bettina Gruen, Kurt Hornik (2011). "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software*, 40(13), 1-30. URL http://www.jstatsoft.org/v40/i13/.
- Stefan Theussl, Ingo Feinerer, Kurt Hornik (2012). "A tm Plug-In for Distributed Text Mining in R." *Journal of Statistical Software*, 51(5), 1-31. URL http://www.jstatsoft.org/v51/i05/.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.



### Acknowledgement

The authors would like to thank NORC Senior Fellow Norman Bradburn for providing the cultural building survey data, and for his valuable insight about the modeling result.

## Contact Information

Fang Wang
Survey Statistician II
National Opinion Research Center
wang-fang@norc.org

Edward Mulrow
Senior Survey Statistician III
National Opinion Research Center
mulrow-edward@norc.org