A Text Mining Approach to Understanding UW-Madison Division of Extension’s Planned Work

Deanna L. Schneider
University of Wisconsin-Eau Claire

Contents

Background and Problem Statement

Each year, UW-Madison Division of Extension faculty and staff submit plans of work, outlining the major programs they intend to work on in the upcoming year. Included in the plan of work is a situation statement. A situation statement is a concise summary of relevant existing data sources (e.g. local, state or national) and new data/findings (e.g., specific information that the employee or partners have collected for developing programming). The situation statements guide the rest of the plan of work, which should be a response to whatever issues are raised in the situation statement. In general, the situation statement outlines the identified need at the local community level that the Extension employee intends to address with their educational efforts.

Extension has recently undergone a restructuring process. One of the major components of that restructuring was the transition from four "program areas" to six "institutes." The previous structure included the following program areas:

- Family Living Programs (FLP)
- 4-H Youth Development (4HYD)
- Agriculture and Natural Resources (ANR)
- Community, Natural Resources, and Economic Development (CNRED)

In the new structure, Family living was broken into two institutes: Health and Well-Being (HWB) and Human Development and Relationships (HDR). 4-H Youth Development was renamed to Positive Youth Development, with the intention that all youth-centered programming would be under this institute. Agriculture was separated from Natural Resources. CNRED became Community Development, with those individuals primarily focusing on natural resource development moving to the Natural Resources institute. This left the following six institutes:

- Agriculture (AG)
- Natural Resources (NR)
- Community Development (CD)
- Health and Well-Being (HWB)
- Human Development and Relationships (HDR)
- Positive Youth Development (PYD)

The purpose of this analysis is to determine whether educator's planned programming topics align with current or historical programmatic structures or with some other undetermined structure.

Data Collection and Preparation

Collection

Educators submitted 413 plans of work to a central data collection site, from which the data was extracted. Each plan of work was tagged by the educator with one or more institutes. Twenty-five (6%) were identified as aligning with multiple institutes. For this analysis, I kept only the 388 plans of work that were identified as belonging to a single institute. The variables used were the situation statement itself and the identified institute.

Cleaning

The data contained multiple non-printable characters. I removed these from the data.

We have a lot of internal, relatively meaningless words. For example, “development” is used in three of our institute titles. We also use the words “Wisconsin” and Extension” excessively, and since these situation statements are describing local conditions in 72 counties, county names become problematic. I added a custom stop word list to the Snowball package stop word list and removed all stop words.

We have some words that are important to our work, but contain numbers, such as 4-H. I replaced 4-H with “fourh” so that I could remove other numbers without losing this important term.

Methodology and Findings

Text mining can be approached from numerous ways. For this analysis, I focused on unsupervised learning techniques, but used an employee-supplied target variable as a reference for the coherency of the identified topics. The following steps were followed:

1. Tokenization, POS Tagging, and Lemmatization.
2. Document Term Matrix Generation
3. Latent Dirichlet Allocation (LDA) Topic Modeling
4. Network Analysis of Shared Words
5. Term Frequency-Inverse Document Frequency Weighting by Topic

The discussion below includes findings from each step of the method.

Tokenization, POS Tagging, and Lemmatization

Jockers (2014) has found that topic modeling with nouns only can generate “highly coherent and highly thematic topics” (p. 156). I used the UDPipe package to tokenize and lemmatize the words, keeping only the lemmatized version of nouns for our analysis.

Document Term Matrix Generation

Using the tm package, I generated a document term matrix for topic modeling, resulting in 3492 unique terms in the 388 documents, distributed in 6 institutes (Figure 1).

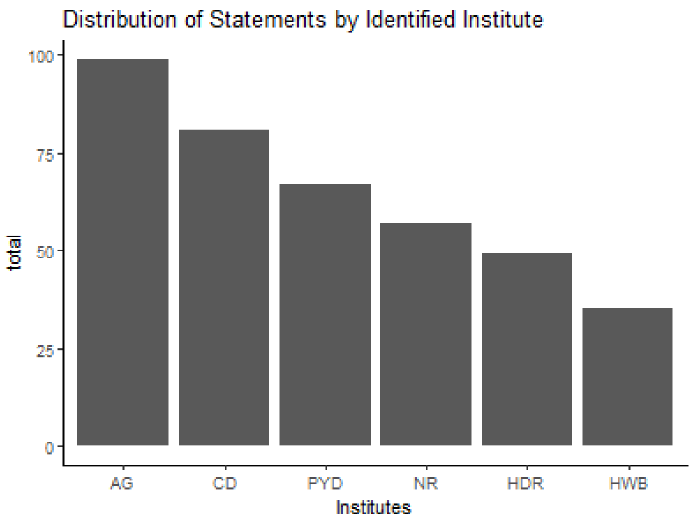


Figure 1 - Distribution of statements by user-tagged institute.

Latent Dirichlet Allocation (LDA) Topic Modeling

I used Latent Dirichlet Allocation topic modeling from the topicmodels package to generate models with four and six topics, replicating our old and new programmatic structure. I repeated the modeling process multiple times with different seeds. In general, the 4-topic model was coherent and consistent from run to run. The 6-topic model changed with each run, sometimes significantly so.

A visualization (Figure 2) of how the modeled topics aligned with institutes with a seed of 5 suggests that in both the 4-topic model and the 6-topic model, the Agriculture Institute makes up the majority of one topic. The HWB and HRD institutes combine in the 4-topic model (equivalent to FLP prior to the reorganization), while a portion of each make up the majority of 2 different topics in the 6-topic model, suggesting that perhaps individuals in these new institutes are still working across institute focus area. Natural Resources and Community Development remain closely aligned with each other in both the 4-topic model and 6-topic model.

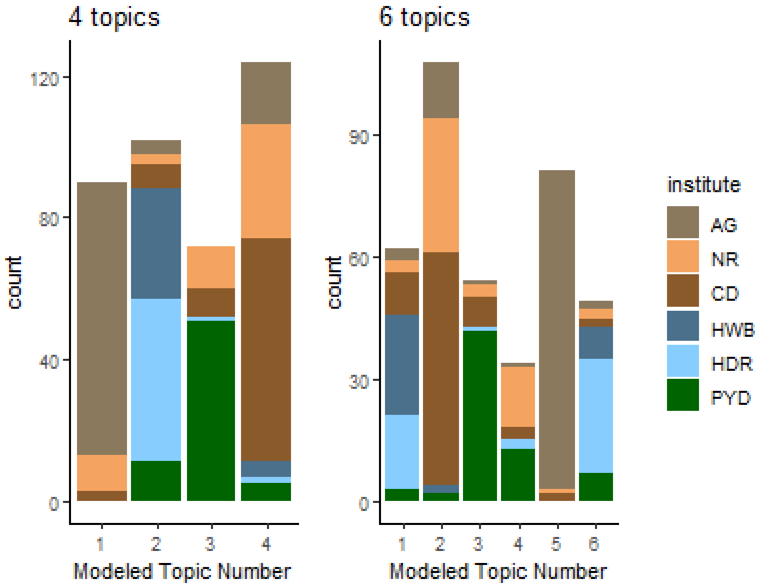


Figure 2 - Distribution of institutes within each modeled topic.

Visualizing the terms in each topic

The R package LDAVis produces interactive visualizations of the modeled topics. I created visualizations for both models and evaluated topics using a lambda of .2 to filter the words to those that are primarily associated with the topic at hand.

The 4-topic model

You can explore the 4-topic model online, or see a static image in Figure 3.

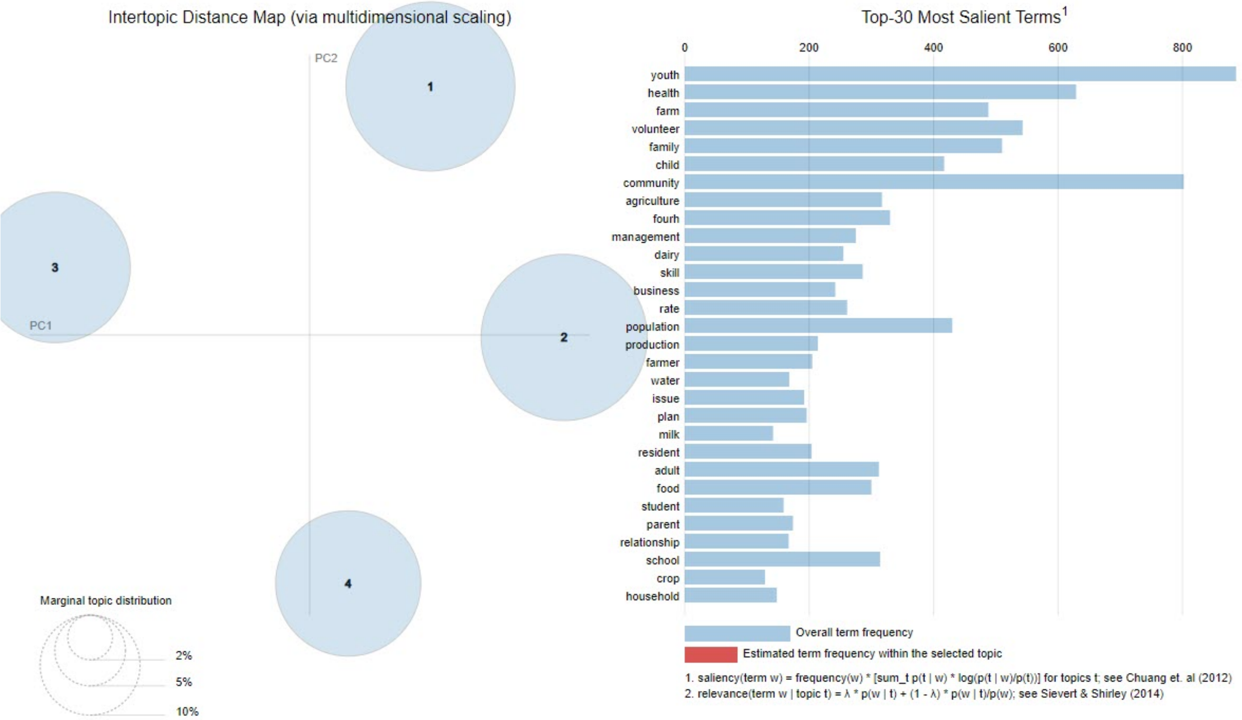


Figure 3 - The 4-topic model's distance and 30 most salient terms

LDAVis orders and numbers the topics by size of the topic, while the topicmodels package does not. The following crosswalk identifies how topics were numbered in each of the 2 visualizations and how I characterized each topic. The discussion following the table refers to the LDAVis topic numbers.

LDAVis (Figure 3)	Topicmodel (Figure 2)	Identified Theme
Topic 1	Topic 2	Families and Health
Topic 2	Topic 3	Youth Development
Topic 3	Topic 1	Agriculture/Agri-business
Topic 4	Topic 4	Community Development and Natural Resources

Topic 1 primarily consists of words related to health and families. These are the 2 most common words at the noted lambda. They are also the two primary topics formally associated with Family Living Programs. Notable shared words in this topic include the words population (which is shared across three topics) and food (which is shared across 2 topics). Food is interesting in that in Topic 1 it is about food consumption (obesity) or food insecurity (poverty), while in Topic 3 it is about food production. While compound nouns and adjectives were not included, the model correctly identified that food belonged in both topics, from different perspectives.

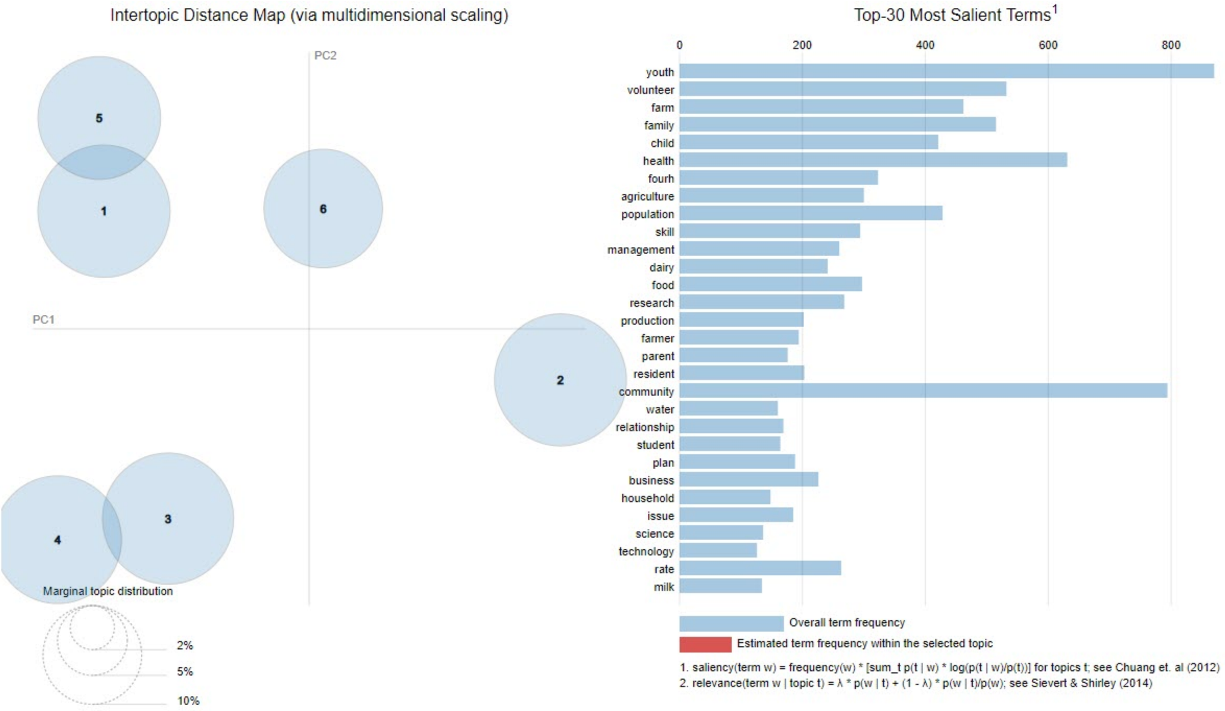
Topic 2’s top words are youth, volunteer, and 4-H (fourh). These words are core to our Youth Development program. Interesting in this topic is the word upham. Upham Woods Environmental Education Center is the location of our statewide 4-H camps. Prior to reorganization, Upham Woods was in the 4-H Youth Development program area. It is currently in the Natural Resources institute. This model, however, put the word entirely in the topic devoted to youth development.

Topic 3’s top words are farm, agriculture, and management. Only 2 words in this topic are shared when evaluated with a lambda of .2 - job and economy. Job is shared with Topic 2, where it most likely relates to teen workforce development. Economy is shared with Topic 4, where it fits with community development.

Finally, Topic 4 is the least coherent of the topics. The top terms are water, issue, and plan. But, Topic 4 nicely aligns with what was formerly called Community, Natural Resources and Economic Development – the program area that dealt with how communities balanced economic and development needs with protecting their natural resources. Still, there are 9 terms in this topic that are shared with other topics at the .2 lambda.

The 6-topic model

You can explore the 6-topic model online, or view a static image of the model in Figure 4.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4 - The 6-topic model's inter-topic distance adn 30 most salient terms.

The following crosswalk identifies how topics were numbered in the 2 visualizations and how I characterized each topic. The topic numbers in the following discussion refer to the topic numbers in the LDAVis column.

LDAVis (Figure 4)	Topicmodel (Figure 2)	Identified Topic
Topic 1	Topic 3	Youth Development
Topic 2	Topic 5	Agriculture/Agri-business
Topic 3	Topic 1	Health and Well-Being
Topic 4	Topic 6	Family Relationships and Risk Factors
Topic 5	Topic 4	Science, Technology, and Workforce Development
Topic 6	Topic 2	Natural Resources and Government

In general, the 6-topic model includes topics that are much more closely aligned. Using this seed, there are 2 sets of topics that overlap, and 2 topics that are well-defined with considerable distance between them and other topics. The size of the topics is also much more similar than in the 4-topic model, and each model contains more shared words.

Topic 1 is primarily about youth development, sharing the same top 3 terms as our youth development topic in the 4-topic model. But, unlike the 4-topic model, in this model, this topic also incorporates some additional terms regarding demographics and underserviced audiences. There are certainly pressing needs in Wisconsin for underserved audiences, but they are typically a fraction of our youth development programming.

Topic 1 overlaps with Topic 5, which I have identified as being about science, technology, and workforce development. The top 3 terms in this topic are skill, science, and technology, but it also includes terms like student, career, and college – topics that are addressed in both the Youth Development and Community Development institutes.

Topic 2 is our most distant topic and covers terms related to agriculture and agri-business. The top 3 terms in this topic are farm, agriculture, and dairy. Only 4 terms are shared with other topics, but that is twice as many shared terms as we saw in our 4-topic model with the same theme.

Topics 3 and 4 are overlapping topics, both in the general area of families. Topic 3 focuses more on health and food security, and closely aligns with our Health and Well-Being Institute, but also incorporates some terms that we would generally consider more closely aligned with community development. Topic 4 focuses more on relationships and risk factors, including such terms as drug, death, alcohol, substance, suicide, incarceration and abuse, despite the top 3 terms being child, family and parent.

Finally, Topic 6, while separated from the other topics spatially, also includes many shared terms. Topic 6 is the only topic where all 3 of the top terms (water, network, city) are shared terms at the .2 lambda. I've titled this topic "Natural Resources and Government" as it appears to be the intersection of government, policy and organizational development, and natural resources.

Network Analysis of Shared Words

In both the 4-topic and 6-topic model, we saw some words that are shared between topics. We can understand these relationships by visualizing them as a bipartite network with words as one type of node and topics as the second. Words are affiliated with topics. By examining which topics are affiliated with the same words, we can get a sense of topic overlap.

Figure 5 shows the affiliation network between the 10 most frequent words in each topic of the 4-topic model. Width of edges correlates to the word frequency. Size of the word corresponds to the degree. Note that all the topics are connected by shared words (1 component).

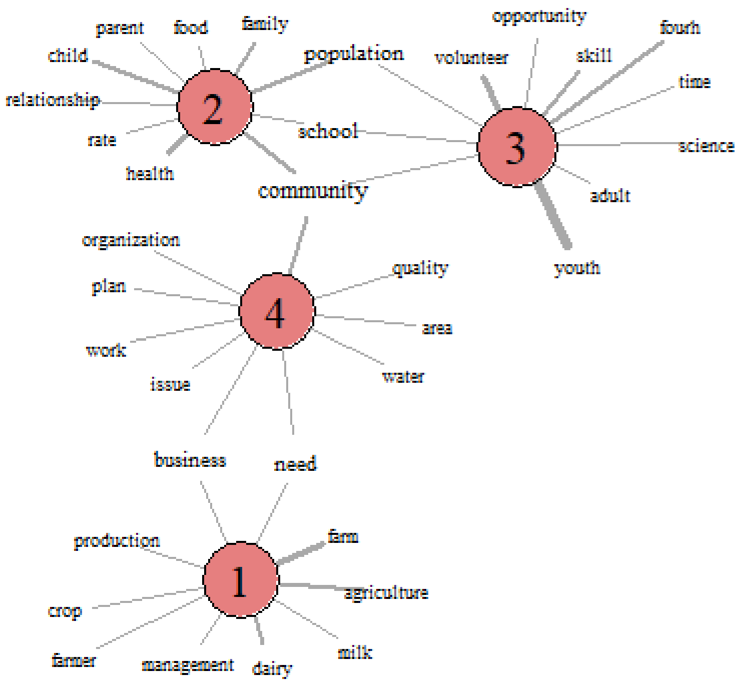


Figure 5 - The relationships among frequent words in the 4-topic model

Figure 6 shows the network of 10 most frequent words in each topic of the 6-topic model. Again, note the single component. But here, notice the increased number of words that share connections with multiple topics. Community, population, youth, and school are all connected to 3 topics. The average degree of the 6-topic network is 2.31, compared to the average degree of the 4-topic network, which is just 2.10.

If we project this bipartite network and look at just the connections among topics (Figure 7), we can more clearly see how much more interconnected the topics are in the 6-topic network. The 4-topic network has an average degree of 2, while the 6-topic network has an average degree of 6.

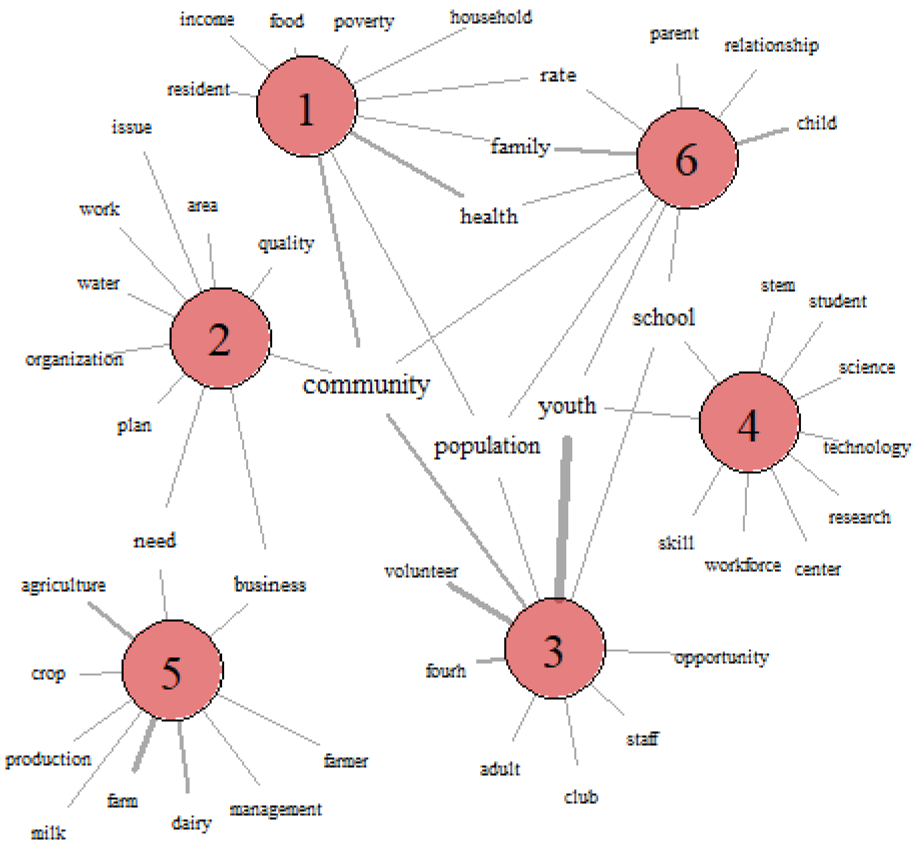


Figure 6 - Relationships of frequent words in the 6-topic model

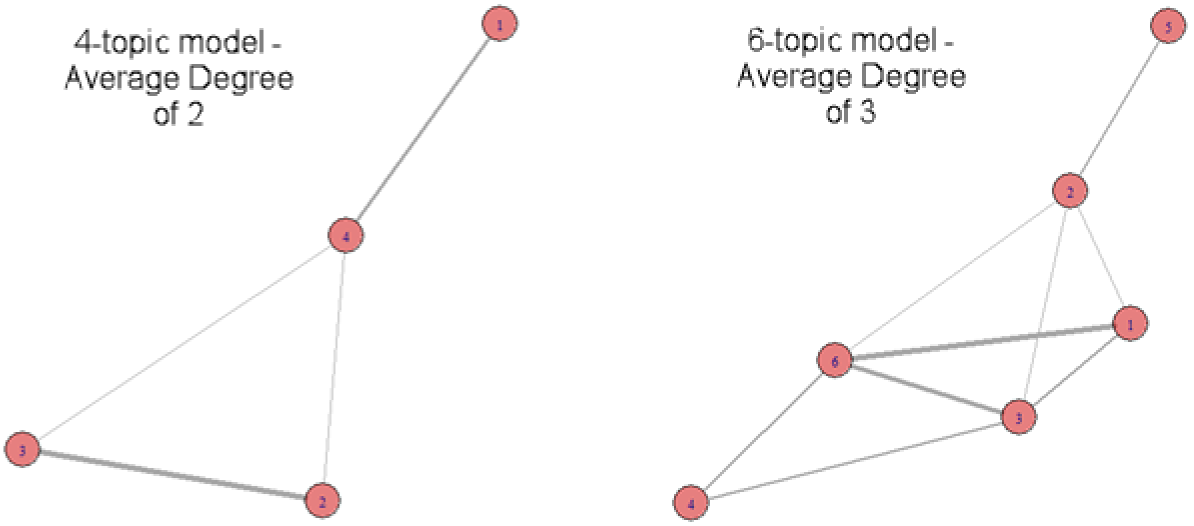

4-topic model - Average Degree of 2
1
4
3
2
6-topic model - Average Degree of 3
5
2
1
6
3
4

Figure 7 - Bipartite projections of the relationships between topics in the 4 and 6-topic models. Line width is the weight of the relationship.

Term Frequency-Inverse Document Frequency Weighting by Topic

Finally, we can look at the words that are frequent while retaining the most uniqueness in each topic. These are the words that have a high term frequency, inverse document frequency (tf-idf). The higher the tf-idf score, the more relevant the term is in identifying the theme or meaning of the topic. Figure 8 displays the top 10 highest-scoring words in each of the topics in our 4-topic model.

Topic modeling itself does not use the tf-idf of terms to assign topics. Using it results in slightly different important terms being identified in each topic compared to those returned by topic modeling. However, using the high scoring tf-idf terms from our topic models can help us to further refine our understanding of the themes of each topic. For instance, in our 4-topic model, while our general understanding of Topics 1 and 4 remain largely the same, we uncover a word we've yet to see in Topic 2 (Families and Health) – alice. This term most likely refers to the ALICE Report, a United Way report detailing the conditions of the working poor in Wisconsin. The high tf-idf terms in Topic 2 suggest that this topic's theme leans more towards poverty and life stressors than we may have originally thought.

New terms appearing in Topic 3 include engineering, latinx and modality, contributing to our understanding of the importance of both science and underserved audiences in Topic 3 (Youth Development).

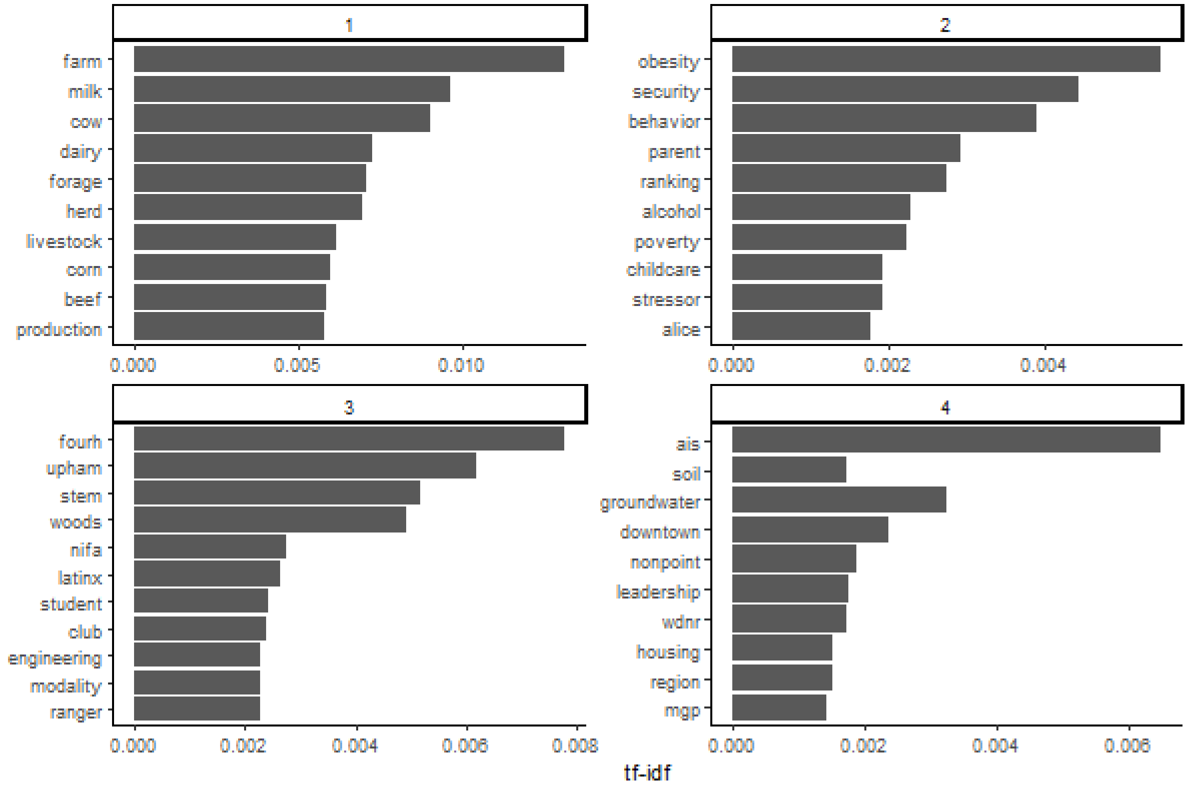


Figure 8 - Most relevant words in the 4-topic model by tf-idf weighting

Likewise, the identified terms in the 6-topic model (Figure 9) reveal some new information. We again see alice appearing, this time in Topic 1 (Health and Well-Being). In Topic 2 (Natural Resources and Government), we also see a new term, wdnr. This refers to the Wisconsin DNR, with whom we partner on many of our natural resource programs. Another new term in this topic is nonpoint, referring to nonpoint source pollution.

Topic 3 (Youth Development) has a new term as well, jone. After investigation, it appears that this is the name of a researcher that is commonly cited. This is an example of a word offering little true meaning without a deep understanding of the topic area.

The remaining topics have good crossover in terms with those identified in the initial topic modeling.

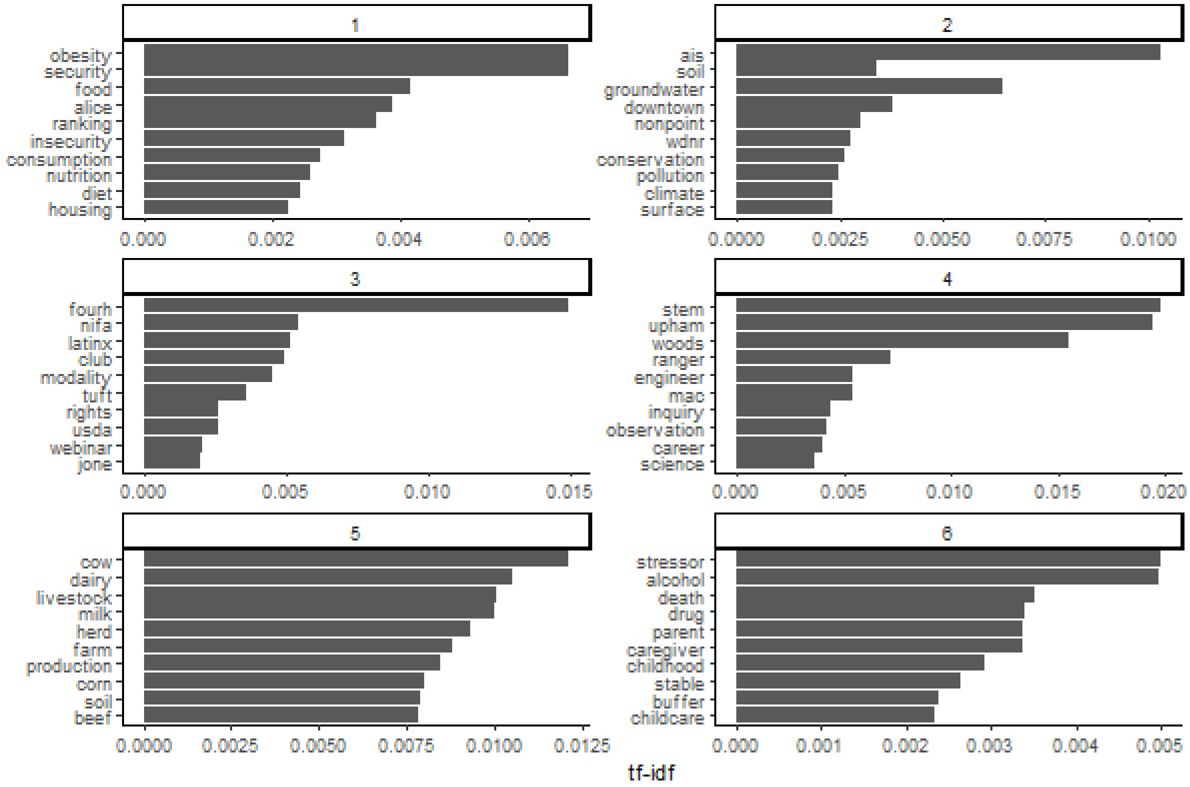


Figure 9 - Most relevant terms in the 6-topic model by tf-idf weighting.

Conclusions and Next Steps

Throughout this exploration we have seen that the themes in Extension Programming do not directly align with our current 6-institute structure. However, they can be summarized as four broad program areas.

The next step could be to look for more granular topics within each of the identified institutes. We might find that a higher number of topics finds more separation and coherence than 6 topics allows. Additionally, while this analysis included planned programs, we could also complete the same analysis with reported outcomes. Reported outcomes should closely align with planned programs. Topic modeling of reported outcomes could be enlightening both in evaluating if planned and reported themes match and if we can identify unplanned emerging issues facing the people of Wisconsin.

References

Auguie, B. (2017, September 09). Arranging multiple grobs on a page. Retrieved April 11, 2019, from https://cran.r-project.org/web/packages/gridExtra/vignettes/arrangeGrob.html

Bail, C. (2018, June 20). Textnets. Retrieved April 5, 2019, from https://github.com/cbail/textnets

Jockers, M. L. (2014). Text analysis with r for students of literature. Cham: Springer.

Maiaini, David, et al. (n.d.). R tm package invalid input in 'utf8towcs'. Retrieved April 2, 2019, from https://stackoverflow.com/questions/9637278/r-tm-package-invalid-input-in-utf8towcs

Sievert, C. (2018, April 25). LDAvis. Retrieved April 10, 2019, from https://github.com/cpsievert/LDAvis

Silge, J., & Robinson, D. (2019, March 23). Text Mining with R. Retrieved April 8, 2019, from https://www.tidytextmining.com/

Statcompute. (2013, May 18). Conversion between Factor and Dummies in R. Retrieved April 5, 2019, from https://www.r-bloggers.com/conversion-between-factor-and-dummies-in-r/

Text Mining example codes (tweets). (n.d.). Retrieved April 2, 2019, from https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html

United Way Wisconsin. (n.d.). United for ALICE. Retrieved April 11, 2019, from https://unitedwaywi.site-ym.com/page/2018ALICE

Wijffels, J. (2019, February 15). UDPipe Natural Language Processing - Text Annotation. Retrieved April 8, 2019, from https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html#udpipe_the_r_package

William.Zheng. (2017, November 21). Topic modeling using Python and pyLDAvis: Part2. Retrieved April 10, 2019, from https://www.youtube.com/watch?v=SF50IK5XgKA