How We Use Machine Learning to Create Our Practitioner Guides:

A methods explanation for a general audience

Prepared by
Austin Hocker and Grant Crider-Phillips

CONTENTS

WHAT'S IN THIS PAPER

This paper gives a general audience a surface-level understanding of topic modeling and how we use it to produce our practitioner guides. After reading this paper, a reader should have a basic understanding of how we prepared text, what a topic model such as Latent Dirichlet Allocation (LDA) produces, and how we interpret the results of LDA to identify themes in student comments. The reader will be able to read through one of the guides at https://provost.uoregon.edu/ada/practitioner-guides and understand how each section is produced. The reader will also be prepared to relate these processes to code notebooks from the supporting repository at https://github.com/Grant-CP/practitioner-guides.

WHAT'S NOT IN THIS PAPER

This paper does not contain a detailed technical explanation of how LDA or its fitting process works, nor does it explain how to program. A full description of some parts of our process will require the reader to visit the above repository, where effort has been made to supply explanations and readable code notebooks. For a complete and readable explanation of LDA and the fitting process used, the reader is directed to the paper by Blei et al. (link).

PURPOSE OF PRACTITIONER GUIDES AND THIS PROCESS

Practitioner guides (link) summarize student perspectives to specific questions on the Student Experience Survey (SES), the largest student feedback survey on campus, collecting over 100,000 student comments each year. The SES was developed in 2018 and replaced the previous student surveys to help collect more targeted feedback for instructors and supervisors about students' experiences of specific teaching practices. However, the surveys are also useful for understanding student feedback across the university. Categorizing student responses across campus is useful for identifying areas of teaching improvement, targeting professional development programming, and informing instructors about common types of student feedback that may inform their pedagogical choices. Furthermore, analyzing student responses helps us understand how students interpret the survey questions and how they themselves define different teaching practices like "inclusive" or "accessible" teaching.

To analyze and synthesize practitioner guides from SES responses, we use a hybrid analysis approach, pairing topic modeling, an objective and efficient machine learning approach, with close reading by humans to make meaning of the computer-generated topics. While the topic modelling approach rapidly increases the speed and objectivity of our analysis, human interpretation adds an understanding of natural language and expertise in pedagogy to identify themes in student responses that have practical importance for instructors.

DATA COLLECTION

SURVEY PROCESS

The student experience survey asks for student comments about how 13 specific teaching practices are either beneficial or in need of improvement for their learning. Therefore, the survey structure itself effectively subdivides student responses based on specific themes and based on positive or negative feedback. Each practitioner guide summarizes student responses about an individual teaching practice. For example, in the accessibility practitioner guide (link), we analyze all responses to how accessibility is beneficial for student learning, and all responses to how accessibility could be improved to support student learning. Researchers who are presented with a less structured textual dataset may need a more complex extension of LDA or will find much more general topics. Repeating our process just with a large number of topics chosen will not work well to accomplish what we have done here.

More information about the SES can be found by visiting: https://provost.uoregon.edu/revising-uos-teaching-evaluations.

COLUMNS USED/ ANONYMIZATION

We reduced each document to a unique document identifier, question code, and response text, then removed all other data associated with the responses. Instructor names were also replaced with FIRSTNAME and LASTNAME.

SUMMARY OF PROCESS

The following six steps are involved in creating a practitioner guide. Each is given a short explanation here and a fuller explanation elsewhere in the paper.

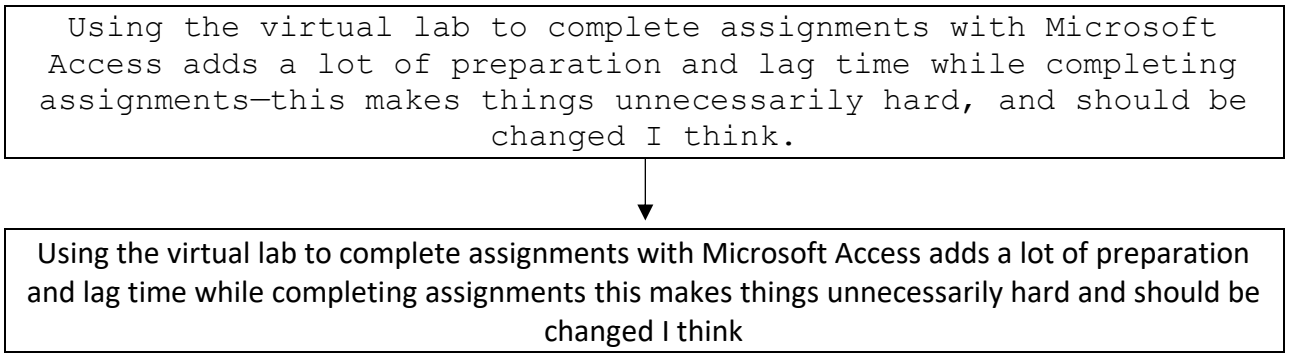
1. Survey collection
    - Students respond to surveys about classes and instructors
2. Anonymization
    - Personally identifying information is stripped from comments and metadata
3. Text preprocessing
    - Text is prepared for computer consumption
4. Topic modeling (LDA)
    - Features for each comment and word are generated using machine learning
5. Human inspection
    - Features from LDA are compared and validated by a human researcher
6. Summary of themes and choosing examples
    - Valuable features are chosen and summarized in a written document

PREPROCESSING

This section describes what steps were taken after surveying students and before sending the textual data to our topic model. Examples and justification are given for the most valuable steps here. A full step-by-step can be found in the preprocessing notebook (link).

The first step is text cleaning, which mostly deals with punctuation and changing special characters indistinguishable to a human eye into standard forms or to a single space.[1] It was especially important to replace various word separators with spaces so that the lemmatizer from Spacy could tokenize and lemmatize the text correctly.

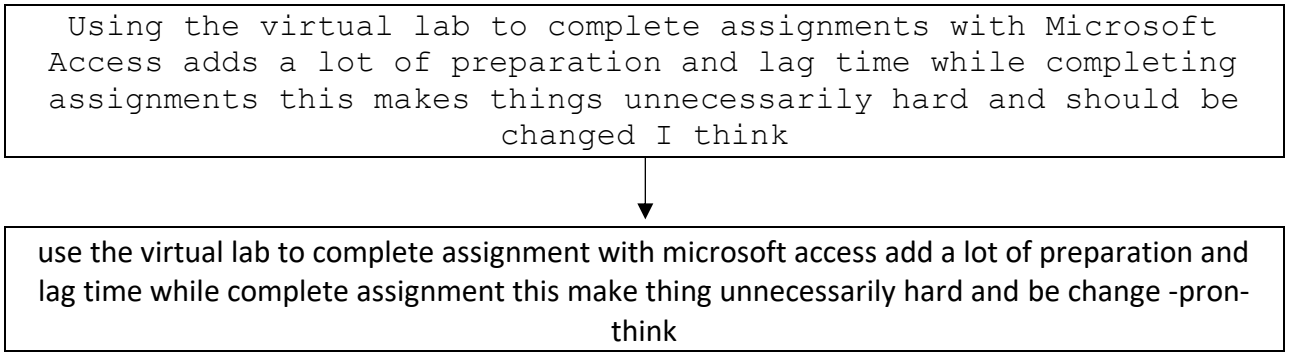
Figure 1 – Example Text Cleaning



The next step is lemmatization. Lemmatization is the process of converting each word into a standard conjugation or declination. For example, "is" becomes "be," or "classes" becomes "class."

We found lemmatization to be extremely valuable for deeper inspection of our LDA models and for other simpler models, but we did not find it to affect how responses were grouped by LDA. Lemmatization was done via default behavior of the Spacy pipeline "en-core-sm," which is the small pipeline trained for English processing tasks.[2]

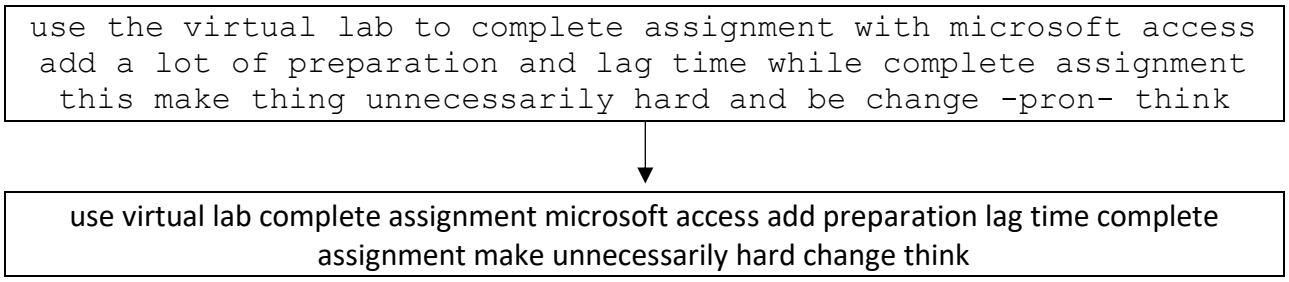
Figure 2 – Example Lemmatization



[1] A fuller description of the exact details can be found in the preprocessing notebook (link)
[2] Documentation for the spacy lemmatization pipeline can be found at https://spacy.io/api/lemmatizer.

The final preprocessing step is to remove stop words. Words that are extremely common or considered to be unimportant for meaning are often removed as "stop words." In our case, we used the default Spacy list of stop words with a few modifications. Our main modification was that we removed references to the instructor such as -pron- (pronouns after lemmatization), firstname, lastname, dr., professor, instructor, and teacher. We found these words to be extremely common, and felt they were unrelated to our task of identifying teaching practices that were salient to students. An alternative analysis, for example one that attempts to identify the sentiment of comments, might find much use from these words.

Figure 3 – Example Stop Word Removal



LATENT DIRICHLET ALLOCATION (LDA)

Latent Dirichlet Allocation, or LDA, finds patterns of word use common across student comments and presents topics based on those patterns.[3] These topics can be used to compare documents, often for the task of clustering together similar documents. LDA is called a topic model instead of a clustering model because it provides enough additional information about each clustering feature that they can be interpreted as topics, like "physics" or "student descriptions of how remote class over Zoom didn't work well." While it is not crucial to understand what type of model LDA is for this context, it is valuable to note that it is not a linguistic model. That is, any understanding of natural language enters this process in the preprocessing and model interpretation steps and does not exist in the design of LDA itself. This is important because we do not justify the use of LDA any further than saying that it worked well for our task.

The easiest way to understand what LDA does to assist a human in reading student comments is to see a few of the things that LDA produces. The reader is once again directed to the paper by Blei et al. (Blei, Ng, & Jordan, 2003) to obtain an understanding of how LDA and the fitting process works, as a full explanation is far beyond the scope of this paper and ultimately not required to understand or even recreate our process.

[3] In the context of math, LDA is a hierarchical Bayesian model, and in the context of machine learning, LDA performs a dimensionality reduction task and has similarities in use to some clustering models.

WHAT LDA PRODUCES

For the topic model, each "topic" is a probability distribution describing how often each word appears when that topic is selected. We explain how we interpret these distributions in the "Topic Synthesis" section. The following gives an example for one of the topics, where the distribution has been altered slightly to down weight words that are common across all topics:[4]

Figure 4 – Term Scores for a single topic in LDA

```
                    Topic 3:
  'office': 0.069, 'expensive': 0.063, 'complete': 0.057,
 'computer': 0.054, 'hour': 0.050, 'student': 0.046, 'book':
0.044, 'access': 0.038, 'class': 0.037, 'textbook': 0.035...
```

For each student comment, LDA breaks it down into a mixture of the topics. For example, with six topics we might see the following probability distribution over topics for a student response:[5]

Figure 5 – The topic mixture of a single student comment

```
Topic 0: 0.223552, Topic 1: 0.000000, Topic 2: 0.010765, Topic
      3: 0.734768, Topic 4: 0.010368, Topic 5: 0.011783
```

Given these pieces of information, LDA models an individual student comment as being generated as follows: for each place where a word will be, a topic is chosen according to the mixture of topics for that comment. Then, a word is selected according to the distribution of words for that topic. That word is put into the comment, and the process is repeated for each potential word in the comment.

HYPERPARAMETER CHOICES

A number of researcher choices in hyperparameter settings are required in producing a topic model. In selecting specific hyperparameters for LDA, we assume topics are not equally represented; for example, that some types of student responses may be more common in the data set. Additionally, we are generous with computational resources to produce models that fit as completely as possible because overfitting[6] is not a concern with LDA as used here. Finally, producing an LDA model requires selecting a set number of topics to be used in the modeling process. We fit models between three and 18 topics and select the models with the highest coherence scores for further analysis. The basic idea of the coherence score is that it looks at how likely the top words in a topic are to be used together based on how words are used together in

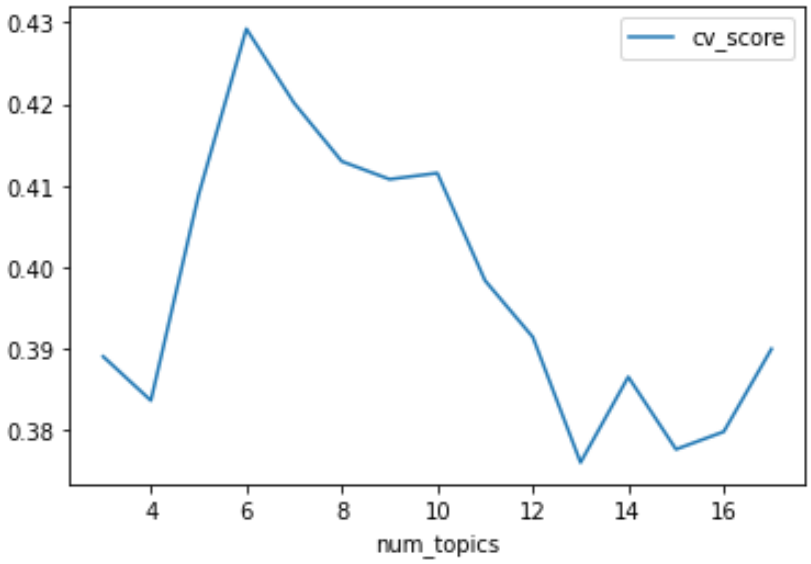
[4] This is done to highlight word words that are specific to topic meaning as opposed to words that are simply very common. The exact formula used comes from (Blei & Lafferty, 2009) and can also be found in the LDA introspection notebook (link).

[5] It is also correct to think of these numbers as the percentages of each topic in the student comment.

[6] Overfitting refers to a common problem in machine learning where a model fits training data very well but cannot generalize to unseen data. We use LDA as a descriptive aid as opposed to a predictive model.

Wikipedia articles.[7] In the example below, we would start our analysis by looking at a six-topic model and move up from there if needed.

Figure 6 – Example coherence scores

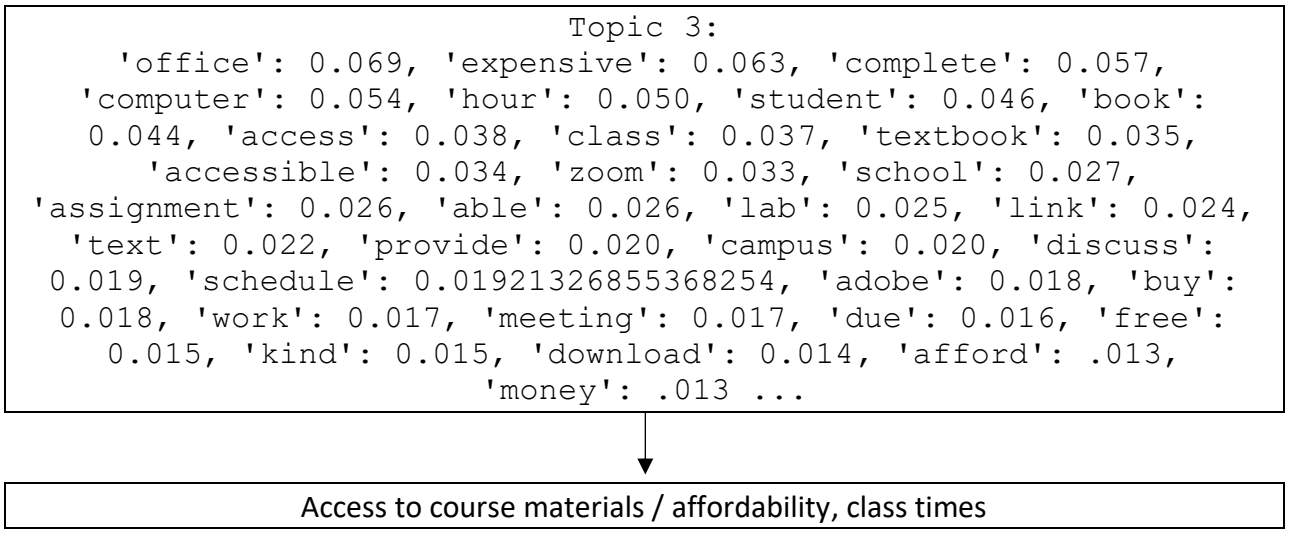


TOPIC SYNTHESIS

Topic modeling can group together documents with any sort of topics, but as the term "topic" suggests, it is also desirable that these computer topics correspond to what a human reader might think of as a common theme. When we say computer topics, we refer to the fact that LDA considers a topic to be a probability distribution over all words that appear anywhere in the responses. If most of the words that are likely to come from that topic (i.e., words with high probability according to LDA) share similar meaning, then we say that the topic is "about" that shared meaning. Below is an example of the first part of the process, which is to determine what, if anything, the topic is about.

[7] See documentation (https://radimrehurek.com/gensim/models/coherencemodel.html) and the supporting paper (Röder, Both, & Hinneburg, 2015).

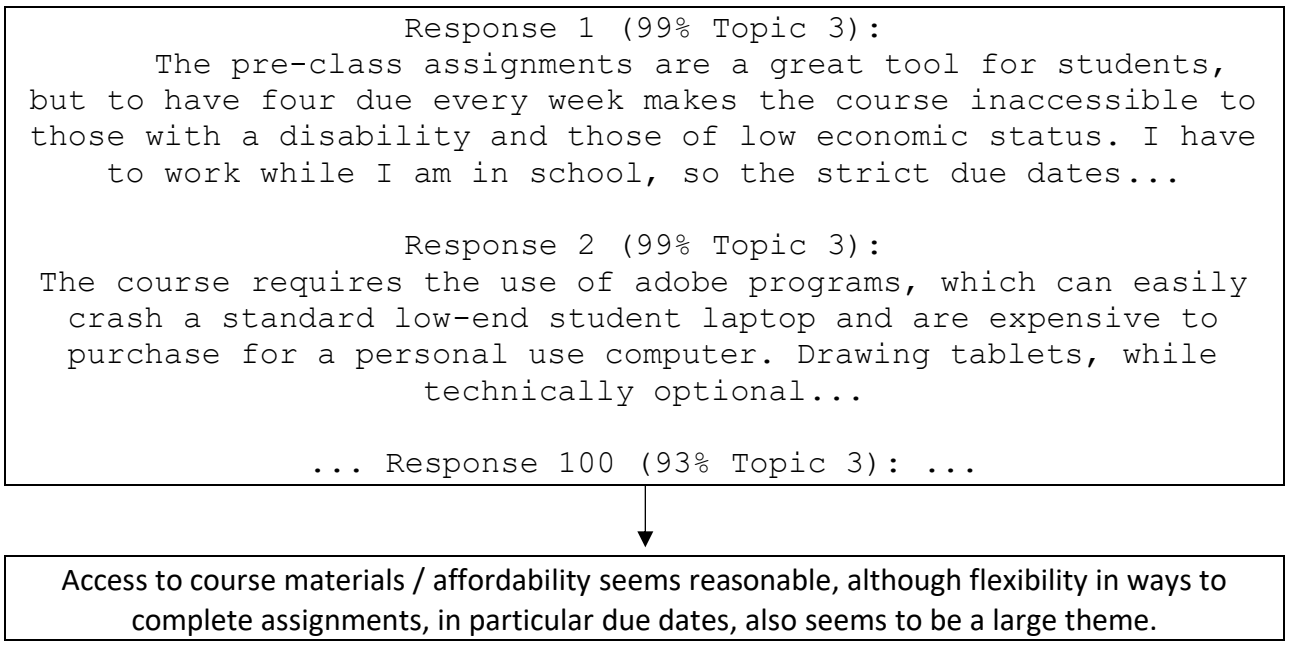
Figure 7 – Interpreting an LDA topic



If the words do not appear to have a shared theme, we move on to a topic model with more topics. If necessary, we reconsider our preprocessing steps, and in particular the list of stop words.

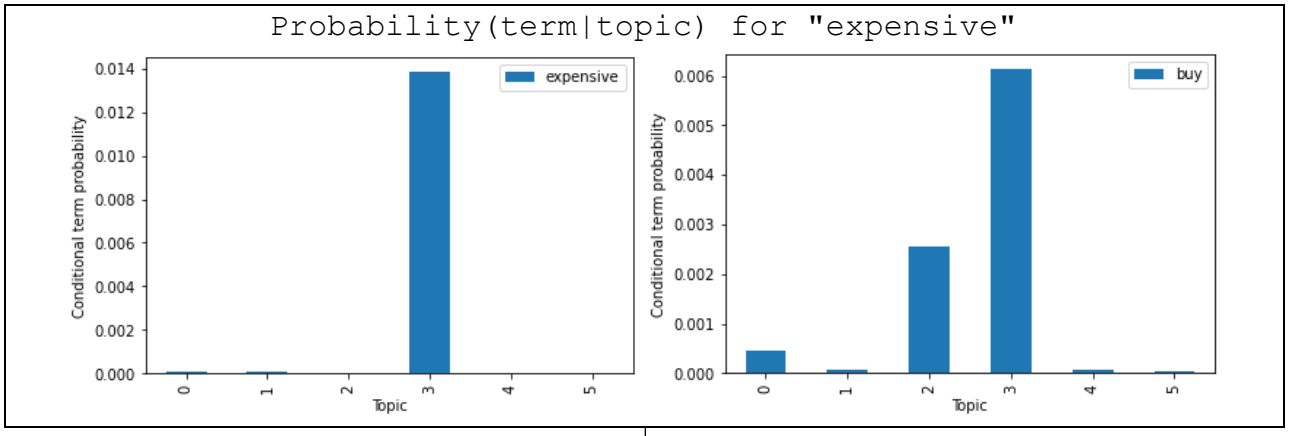
After creating a hypothesis for what the topic is about, the researcher checks that idea against responses that come primarily from that topic. The researcher might look at hundreds of comments sorted this way, which can be done quickly and reliably since it is just a yes/no question to answer whether the human interpretation of the topic that was created earlier actually matches each comment.

Figure 8 – Verifying topic interpretation



As an additional tool to inspect the differences between topics, the researcher can look at the distribution of individual words. If words that are important for a particular topic are shared with another topic, it may be the case that those topics need to be combined, or that the human interpretation of computer topics has failed for that model.

Figure 9 – Examining the distribution of individual words



The word “expensive” appears to be concentrated mostly in topic 3. After checking other common words regarding financial costs such as “buy” and “afford” etc., we confirm that responses regarding financial costs are being captured well by topic 3.

The objectivity of the machine learning pipeline is great for reproducible results with lowered bias from the researcher, but it is also so objective that it models patterns we do not care about. One of the jobs of the researcher is to assign value to some topics and ignore others. For example, for some datasets our approach identified “I feel” statements together in a topic. “I feel” statements had patterns of word use that LDA identified, but a human researcher knows that those comments shared a particular tone or register, instead of sharing a theme relating to a particular teaching practice. In this case the human researcher can ignore the topic put forth by LDA as being irrelevant to the current task and identify if those comments are related to a different common topic. Another example of adding human judgement is when open ended responses frequently mirror the wording of the question. For example, in response to the question “What specifically about the support from the instructor helped your learning?” respondents may often write “The support was helpful because...” These common patterns are identified as a topic but may not be useful for understanding student responses. Therefore, topics like these can be ignored, combined with other topics to identify common themes, or key words from the question can be removed in preprocessing steps. One of the advantages of LDA over a simpler model is that it models documents as being composed of multiple topics, and so it is possible to disregard large parts of topics if needed.

As mentioned earlier in the topic modeling section, one of the most sensitive parts of this process is the transition from distributions over words (computer topics) to themes that have understandable meaning (human themes). It is the job of the researcher to perform this

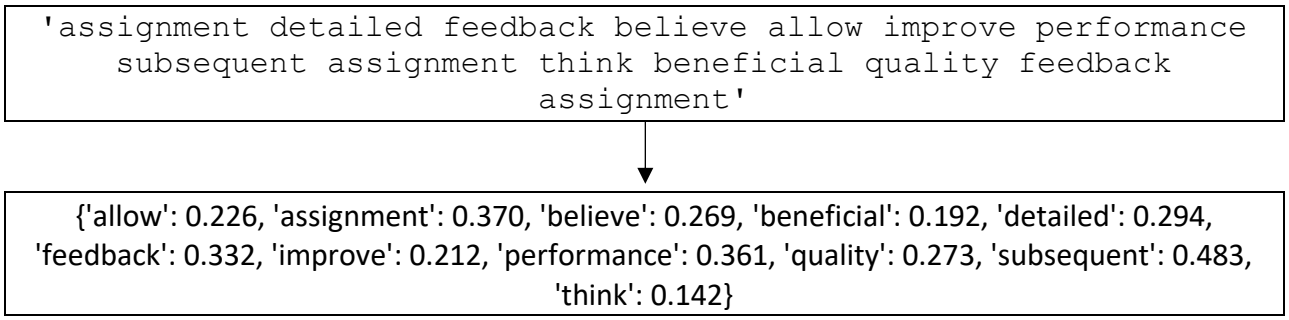
conversion carefully and responsibly, and the need for care is the main reason the practitioner guides are presented without any raw results from the topic modeling process. A naïve reader, not realizing how little a computer understands about human meaning, might try to directly interpret the results of LDA, and it is the researcher's job to protect the reader from misleading themself.

RELEVANT WORDS USING TF-IDF

As a supplement to the more complex theme extraction using human interpretation, we used a term frequency–inverse document frequency (tf-idf) calculation to identify relevant and common words unique to a particular category of student responses. The purpose of this is two-fold. First, we as researchers use the easy-to-understand calculation to check that we are not missing anything obvious with LDA. Second, it allows us to provide the reader with something that is calculated and understandable, albeit with much more limited use than the results of LDA.

The tf-idf calculation is simply a way of counting word frequency while accounting for document length and the tendency of unimportant words to be used in a variety of contexts. As with LDA, the easiest way to understand how tf-idf transforms text is with an example:

Figure 10 – An example tf-idf transformation



In this example, even though the words "think" and "feedback" both appear once in the preprocessed text, "think" is assigned a lower tf-idf score because it appears at least once in more student responses. Scoring words in this way is useful under the assumption that irrelevant general-purpose words like "think" appear in more contexts than specific content words like "feedback."

To produce the tables of relevant words in a practitioner guide such as the one for accessibility, we average the tf-idf scores of words in accessibility responses, then remove the top 200 words in tf-idf scores from non-accessibility responses. Tf-idf does an acceptable job of identifying which words in a document are meaningful, and this process supplements that by identifying which words are meaningful only to the student responses on accessibility.

A better understanding of tf-idf methods and examples of widespread use can be found in a variety of places, including Wikipedia (link). Additionally, the readable code notebook used for our calculations is available in the repository (link).

CONCLUSION

This approach increases the speed and objectivity of our analysis of written comments. First, it significantly reduces the amount of time it takes a researcher to analyze a large body of textual data into common themes compared to traditional approaches like qualitative coding. For example, with large data sets, researchers might randomly sample a subset of comments for analysis due to time and funding constraints. Our approach allows for inclusion of the entire dataset for analysis while still reducing the time it takes to identify themes, and it is done in a completely reproducible manner. Additionally, our approach is inductive and takes advantage of the objectivity of computer-generated topics to identify potential topics of interest that may be missed due to prior beliefs of a human reader. However, we do not wholly rely on topic models to make sense of comments, but rather pair this machine learning approach with human reading to ensure themes are representative of the meaning behind written comments.

The next step for a reader interested in trying it out for themselves is to clone the GitHub repository (link) and follow the tutorials there. To make changes to the process and fully understand the model, reading the LDA paper by Blei et al. (Blei, Ng, & Jordan, 2003) is vital.

BIBLIOGRAPHY

Blei, D. M., & Lafferty, J. D. (2009). Topic Models. In A. Srivastava, & M. Sahami, *Text Mining: Classification, Clustering, and Applications* (p. 75). Boca Raton, FL: Taylor and Fracis Group.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1002.

Crider-Phillips, G. M. (2022, January 27). *practitioner-guides*. Retrieved from Github: https://github.com/Grant-CP/practitioner-guides

Honnibal, M., & Montani, I. (2022, January 27). *Lemmatizer - spaCy API Documentation*. Retrieved from spaCy: https://spacy.io/api/lemmatizer

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eigth ACM International Conference on Web Search and Data Mining* (pp. 399-408). Newy York, NY: Association for Computing Machinery.

Rehurek, R., & Sojka, P. (2022, January 27). *models.ldamodel*. Retrieved from Gensim: Topic Modeling for Humans: https://radimrehurek.com/gensim/index.html