IST 736 – Text Mining Michael Phillips

Final Report 4/24/18

Topic Modeling on Syracuse University ETDs

1. Introduction

Topic modeling is a popular technique for evaluating a large collection of unstructured text data; it provides us with a tool to explore and classify the data without needing to know what the data is "about" beforehand – a kind of *unsupervised* analysis. Topic modeling has been performed on many kinds of collections with different goals, such as recommending scientific articles (Wang & Blei, 2011), evaluating student journal writings (Chen, Yu, Zhang, & Yu, 2016), and predicting popular Twitter messages (Hong & Davison, 2010) to name a few. For the purposes of this report, a dataset was collected consisting of all 4,032 abstracts from Syracuse University ETDs (electronic theses and dissertations). The goal will be to use topic modeling to reveal the main topics of research that are output by the graduate student body at Syracuse University. This would provide current and prospective students at SU with an idea of what the main subject areas of interest are at the university.

In what follows, the analysis is two-fold: (1) first, an evaluation of the topic modeling algorithms (LDA and NMF) is provided, including an analysis of the prolificity of the different topics, (2) then, a further analysis is conducted which evaluates the distribution of topics by each department at the university. This follows a discussion of how the data was collected and preprocessed, and what parameters were used to create the models.

2. Method

2.1 Data collection

The data was collected by web scraping the ETD titles, departments, keywords, and abstracts from the university's institutional repository, SURFACE. As mentioned, there were 4,032 abstracts in total. The dataset was then transformed into another dataset which amalgamated keywords[1] with abstracts, leaving other metadata like department names behind. This is because we don't want the names of the departments or the titles of the ETDs to be evaluated in the topic modeling, however, we still want to be able to retrieve department names later, so we can see which keywords and abstracts (henceforth "documents") correlated with which departments. As a preprocessing measure, all words in the documents were lowercased, stripped of punctuation, and lemmatized. The final dataset consisted of a JSON file, which contained a list of strings, corresponding to the 4,032 preprocessed documents.

2.2 Models

Two popular topic modeling algorithms, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) were used to search for topics in the ETDs. The algorithms

[1] Keywords were simply prepended to the beginning of the abstract. Note that about 10% of ETDs lacked any keywords, and so only the abstracts are evaluated in such cases.

were implemented using the Python machine learning library sci-kit learn. The purpose of evaluating two algorithms was to perform a comparative analysis on their results, in case one algorithm clearly out-performed the other. Each algorithm required the data to be vectorized, and so a TF-IDF matrix was used for the NMF model, and a raw frequency vectorizer was used for the LDA model (which requires raw frequency). In each case, the following parameters documented in (1) were used:

(1) Parameters for TF-IDF and Count vectorizers

Parameter	Setting	Justification
stop words	built-in English stopwords + ['dissertation', 'problem', 'approach', 'method', 'research', 'thesis', 'problems', 'report', 'project', 'results', 'using', 'use', 'described', 'designed', 'chapter', 'chapters', 'study', 'analysis', 'proposed', 'models']	A list of common yet uninformative English words were removed, including a custom list of "dissertation-related" words which appeared prolifically in each topic on the initial run. Upon removing the words, the models' results improved by becoming less-similar.
minimum document frequency	3 (A word had to minimally appear in 3 of 4,032 documents to be considered)	If a word only appears in one abstract, it is not likely not a very identifiable or telling word. Such words are most likely removed via the max features parameter anyway.
maximum document frequency	0.95 (A word can appear in no more than 95% of the data).	If a word appears in nearly all of the documents, it is most likely not a word discernable to any 1 topic. Such words might be eliminated by the stop words filter.
max features	1000 (Only the top 1000 words (by highest raw frequency or TF-IDF score) were considered).	The decision came about by iterations of trial-and-error with lower and higher numbers, and 1000 seemed to best-help the topic frequency to match the department frequency (to be discussed later).

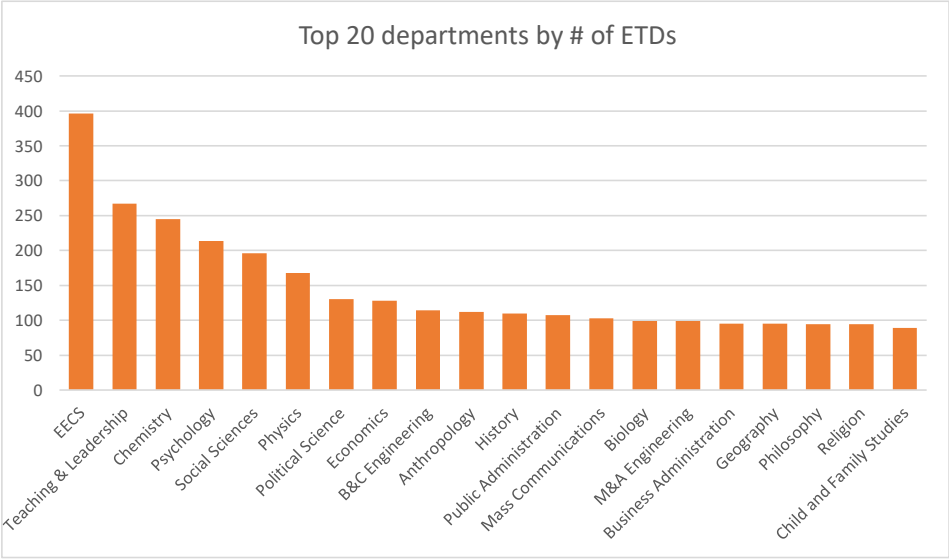
Commented [BY1]: Great!

2.3 Number of topics (K)

The parameter settings for the topic modeling algorithms themselves were set to defaults, with the random seed (random_state) randomly set to 44 for each. The primary parameter of concern was the number of topics, K, to return for each model. One of the difficulties in topic

modeling is gauging how many topics one should search for despite the fact that the data is assumed to be unstructured and unevaluated. However, the dataset in this case is not completely unstructured; there is certain metadata associated with the ETDs that we can use to our advantage. In particular, we can evaluate that among the 4,032 ETDs, there are 60 represented departments. The graph in (2) shows the frequency of ETD documents by department for the top 20 most frequent departments.

(2) Top departments by ETD count[2]



EECS = Electrical Engineering and Computer Science

Thus, if we know that there are roughly 60 different departments which the ETDs represent, one metric for choosing K might be to approximate based off this number. However, since the smallest departments by ETDs have <10 documents which represent them, a number slightly less than 60 is picked, namely K = 50.

Commented [BY2]: Good estimation strategy

3. Results

3.1 Topic results by model

Using the parameters outlined in section 2, both models produced 50 topics for the ETDs, displaying the top 8 words for each topic. Ultimately, despite the difference in their mathematical underpinnings, the results of each model were quite similar, which may be unsurprising as in each case, their vectorizers used the same parameters and the number of topics was the same. The

[2] For brevity's sake, only the top 20 results of departments are shown; this standard is adopted throughout the rest of the report for evaluating topics as well. To see the full list of departments, topics, and more, visit the project website at http://mphilli.github.io/ETD-topic-modeling

following chart in (3) displays an abridged selection of 15 topics, annotated using personal judgment, and the respective LDA and NMF "version" of that topic.

(3) Topic results comparison

NMF	Topic	LDA
application performance parallel memory design data software code	Computer Science	network service design performance management based decision organizational data communication
students course academic classroom grade university achievement	Academia	student education college participant data experience learning academic teaching students
policy foreign state decision international security united domestic	Politics	policy state political government war case politics economic institution country
child family parent mother children father parental involvement	Family	family woman child social life experience parent relationship interview mother
tax income rate essay effect local capital bond	Finance	market essay firm effect impact cost industry increase business income
intervention treatment feedback behavior training control assessment condition	Behavioral Statistics	intervention treatment significant control score test effect result related measure
school education district reform educational principal	Education	school teacher education district mathematics child practice teachers educational teaching
cell gene substrate growth formation membrane effect culture	Biology	cell protein growth drug sequence membrane gene effect binding interaction
metal compound ligand chemistry complex material synthesis synthetic	Chemistry	material metal synthesis property compound complex polymer corrosion novel composite
writing literacy rhetorical rhetoric composition feedback discourse practice	Literacy	discourse literacy language social communication way writing narrative
reaction product synthesis acid organic yield temperature natural	Physics	phase reaction temperature energy product transition state surface thermal transfer
political state history century new economic historical cultural	Local History	new york century state pattern native historical landscape period resource
relationship factor participant data behavior individual level process	Psychology	self behavior female male adult sexual disability gender young men
soil water concentration forest organic lake stream deposition	Botany	water soil concentration stream plant specie organic deposition capacity forest
religious religion philosophy tradition life argue modern moral	Philosophy	reform moral conflict account reason argue argument agent view claim

As we can see, for any given topic of one model, an approximate topic can be found in the other model, displaying their similarity. They were of course not perfectly similar, as (3) shows, but the exact extent of their difference is not measured in the current study. As mentioned, the topics assigned in (3) were generated by personal judgment, and the easiness or difficulty of

annotating topics varied, although most as in (3) seemed straightforward. The biggest concern regarding the results is that some topics were very similar, such as the NMF discovered topics in (4):

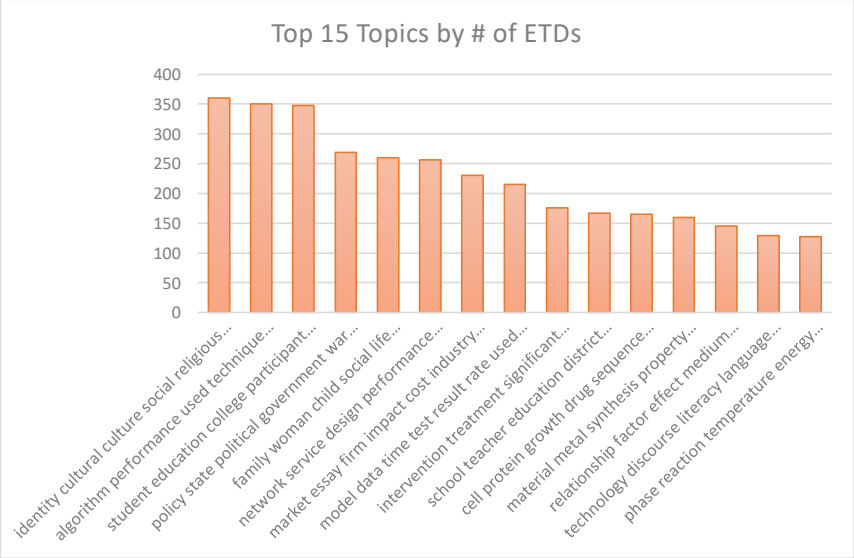
(4) Similar topics (NMF)

i. college university student year institution higher education faculty
ii. student students course academic classroom grade university achievement
iii. learning teaching knowledge instructional education mathematics course

Perhaps these topics could be "collapsed" into a single topic by reducing the number of topics to <50 moving forward.

In addition to producing the 50 topics, each model also produces a document-topic matrix populated with scores which signify the strength of the relationship between each topic and document. Using this information, we can find the "top" topic for each document, and create a set of "top topics", i.e., topics which had the highest score for the highest number of documents. The top 20 most prominent topics are shown in (5):

(5) Top topics (by document-topic scores)



Ideally, the results here would somewhat reflect the frequency distribution of ETDs by department shown in (2), as this metric was used to inform the number of topics in the first place. To some extent, this is the case. The 2nd and 6th topics (from the left) capture many of the ETDs from the EECS department, which had the highest number of ETDs overall. Meanwhile the first topic, likely labelable as "culture", is a broad topic that is likely associated with multiple departments, hence its ranking as #1. This chart helps to develop the goal of the study, which is to

explore what topics graduate students at Syracuse University tend to write their theses and dissertations on. We can consider this to be more informative than the raw frequency of ETDs by department, because these topic rankings are able to consider data from all of the departments, and bring to the surface those topics which are the most prominent across multiple departments.

Commented [BY3]: Yes – iterative data analysis with new hypothesis generates from prior results

Thus, one reason the department frequency and topic frequency do not seem to directly correlate is that some departments are more diverse than others with respect to their topics, and have their respective ETDs split among many different topics. The chart in (5) thus reflects not just the departments, but the diversity of the departments. This is explored further in 3.2.

3.2 Topic diversity by department

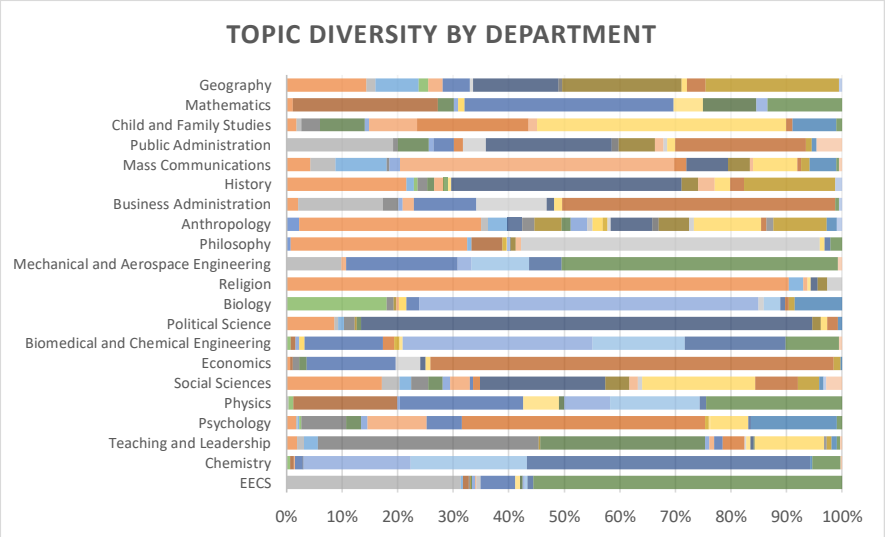
In addition to evaluating the topics themselves, we can also evaluate their distribution over the documents (ETD abstracts) by department, again leveraging the ETD metadata to assist in the evaluation. This is accomplished with the following steps:

I. There is a score for the strength of the relationship between each topic and each document. Among these, find the highest scoring topic for each document.

II. Find the department associated with each document and associate the document's highest scoring topic with that department.

The result is a matrix which is populated with scores between each department and each topic, based on the documents and their respective topic and department associations. The following chart in (6) displays a 100% stacked bar for a set of 10 departments. Moving forward, only the LDA algorithm was used to produce these results (which was arbitrarily picked due to their similar results of each model).

(6) Diversity in the Topic-Department matrix



This visual normalizes the department-topic scores to occupy the same 0 to 100% space. It seems to be the case for most departments that there is one or two dominant topics, followed by 5-10 others. Clearly, the least diverse department is Religion, for which 90% of its 94 ETDs most strongly corresponded to the LDA-discovered topic "identity culture social religious way practice context" (in orange). The largest department by ETDs, EECS, was prominently defined by two topics: "network service design performance management decision organizational" (grey) and "algorithm performance used technique result field detection application" (green).

From these results, we can also explore relationships between the different departments. For example, the most prominent topic for Biology is "cell protein growth drug sequence membrane gene" (~65%) and the most prominent topic for Chemistry is "material metal synthesis property compound complex polymer corrosion" (~55%) are both present in the Biomedical and Chemical Engineering department, at ~33% and ~18% respectively.

4. Conclusion

In this study, topic modeling using LDA and NMF algorithms to explore ETD abstracts from Syracuse University was conducted. Metadata of department associations was leveraged to guide the number of topics, K, to be 50, a number slightly less than the unique number of departments. The two models displayed similar results, as was seen in (3). (3) also demonstrated the apparent success of the topic modeling, as the words used to create the topics by each model seemed coherently related. Each topic was annotated using personal judgment, although the opportunity to evaluate the results using intrusion (Chang et al., 2009) or PMI (Newman et al., 2010) would help improve the evaluation. One problem observed was that some of the topics seemed to be quite similar, which could perhaps be alleviated by adjusting K or other model/vectorizer parameters.

It was observed, through both department frequency and topic frequency, that some of the top topics or areas of study for Syracuse University seemed to be "Culture", "Computer Science", "Teaching/Education", and "Politics", to name a few (per (5)). As potential future work, we may wish to provide a department-by-department topic modeling analysis, but the interpretation of the results would require esoteric knowledge about the field(s) in question.

Gauging the diversity of topics proved to be a particularly difficult yet interesting undertaking. It helps us to explore not only what topics people write about but also how and to what degree the different departments at the university are related to one another based on the topics. However, the accuracy of this diversity is entirely contingent upon optimizing the results of our model, including and perhaps most importantly the number of topics.

Commented [BY4]: Excellent! 15/15

References

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009) Reading tea leaves: how humans interpret topic models. *Advances in neural information processing systems 22*, pages 288-296.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016, April). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5).

Hong, L. & Davison, B. (2010). Empircal study of topic modeling in twitter. In *Proceedings of the SIGKDD Workshop on SMA.*

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010, June). Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 215-224).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 448-456).