JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume x(x), 1—10. http://dx.doi.org/xxx-xxx-xxx

Analyzing Educational Priorities Over Time Using Topic Modeling

Madison Hobbs[1], Evan Liang[2], Aanya Alwani[3], Jean Selasi Adedze[4], Dominique Macias[5], Dr. Talithia Williams[6]

Abstract

How can the overarching curriculum of educational institutions be studied over time? We leverage unsupervised topic modeling to analyze a corpus of syllabi spanning 13 years from Oxford College at Emory University. By generating topics and comparing these across the years, we can estimate how much each topic is emphasized in the curriculum of Oxford College at Emory University over time.

Keywords
topic modeling, education, syllabi.

Submitted: mm/dd/yy — **Accepted:** mm/dd/yy — **Published:** mm/dd/yy

[1] *Department of Mathematics, Harvey Mudd College, Claremont, USA. email icon: mhobbs@g.hmc.edu*
[2] *Department of Mathematics, Harvey Mudd College, Claremont, USA. email icon: eliang@g.hmc.edu*
[3] *Department of Mathematics, Harvey Mudd College, Claremont, USA. email icon: aalwani@g.hmc.edu*
[4] *Department of Mathematics, Harvey Mudd College, Claremont, USA. email icon: sadedze@g.hmc.edu*
[5] *Department of Mathematics, Harvey Mudd College, Claremont, USA. email icon: dmacias@g.hmc.edu*
[6] *Department of Mathematics, Harvey Mudd College, Claremont, USA. email icon: twilliams@g.hmc.edu*

1. Introduction

In recent times, topic modeling has emerged as an innovative tool to extract meaningful information from large sources of text. Part of the surge in attention to topic modeling lies in its ability to uncover insights from documents that would otherwise never be hard to discern in aggregate by a human. As educational institutions increasingly adopt data driven approaches in their administrative practices, we see immense potential in leveraging topic modeling techniques to extract insights from large collection of documents like written work, handbooks, syllabi, course evaluations, textbooks among others that are abundant in these institutions.

In this paper, we analyze the changes in the school curriculum over time. We leverage topic modeling to analyze a corpus of syllabi spanning 13 years from Oxford College at Emory University. By generating topics and comparing these across the years, we can estimate how much each topic is emphasized in the curriculum of Oxford College at Emory University over time. This type of analysis has the potential to inform parents, policy makers and school administrator of how priorities of educational institutions have changed in time.

2. Related Work

Sekiya et al. present a similar investigation to ours with a slightly different angle (Sekiya, Matsuda, & Yamaguchi, 2017). Rather than looking at all course subjects across time at one school, they narrow the focus to computer science curriculum across multiple schools within the same time frame. Furthermore, rather than having to train a topic model themselves, they leverage the pre-existing CS2013 Body of Knowledge (BOK), produced by the ACM and IEEE Computer Society and detailing the 18 primary topics in Computer Science curriculum as of 2013. Sekiya et al. used those 18 topics to train a simplified, supervised Latent Dirichlet Allocation model (ssLDA) which then outputs, for a given unseen computer science syllabus, how much each of those 18 core topics were represented. However, since we lack predefined topics, our work involves training unsupervised topic models to discover the core topics across all disciplines and observe how these have changed over an 13 year period at Oxford College of Emory University.

Another publication more similar to our temporal analysis though not using syllabi is “Studying the History of Ideas Using Topic Models” (Hall, Jurafsky, & Manning, 2008). Like us, they have time series data; in their case, journal publications from the ACL Anthology (a compilation of natural language processing papers) spanning multiple years. Also like us, they train a

single topic model on their entire corpus, spanning all years. Unlike us, they have 12,500 documents compared to our 3,778. They also only mention trying one algorithm on a fixed number of topics (100 topics with LDA), whereas we select a best model across multiple algorithms and numbers of topics. After training a model, Hall et al. hand-select a subset of the topics with which to proceed, then observe the prevalence of those topics for documents from each year. This technique inspired the way we approach the temporal analysis in our paper.

3. Methods

3.1 Data

Our dataset consists of 3,778 syllabi from Oxford College of Emory University from 2001 to 2014. This was the largest and most temporally expansive collection of syllabi we could find on a university website for scraping. After scraping syllabi from Oxford's website, we discarded syllabi which were scanned copies because text could not be easily extracted in such format. This explains why we consider syllabi from 2001 to present even though Oxford College has syllabi from 1990.

3.2 Processing

First, we use regular expressions to extract words composed of letters from the English alphabet. For consistent representation, we convert all words to lower-case. In order to improve the effectiveness of topic modeling algorithms, we remove stop words such as "the", "a", "he", etc using NLTK's pre-defined list of stop words (Loper & Bird, 2002). Using NLTK again, we further filter out words not in the English language. Finally, we use TF-IDF from the Scikit-learn package in Python (Pedregosa, 2011) to filter out any other words which do not provide differentiable course-related information for topic modeling algorithms to leverage.

3.3 Topic Modeling

Topic modeling is an unsupervised learning algorithm that takes in a large corpus of text data and returns a specified number of representative topics found in the corpus. A topic is defined as probability distribution over fixed vocabulary. Words in a topic are sorted in descending order using the probability assigned to each word in the topic. The top k words in a topic (for some k defined by the user) reflect the overarching and related concepts of the topic. Topic modeling therefore provides a method of extracting insights about the high level meaning of a text. An example of an output of topic modelling is shown in figure 1.

Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10	Topic # 11	Topic # 12	Topic # 13	Topic # 14	Topic # 15	Topic # 16
independent	fitness	math	theater	basketball	web	network	wine	swimming	painting	yoga
project	exercise	linear	musical	game	design	configure	winery	polo	color	relaxation
study	training	algebra	music	intercollegiate	site	cisco	grape	swim	drawing	breathing
end	strength	solve	production	team	create	routing	tasting	water	design	strength
noted	endurance	exponential	performance	competition	data	security	vineyard	backstroke	studio	flexibility
semester	aerobic	quadratic	ensemble	shooting	use	configuration	sensory	training	art	yo
develop	walking	rational	acting	participation	office	operating	production	stroke	critique	balance
form	kin	logarithmic	vocal	flag	lab	server	viticulture	butterfly	value	kin
instructor	muscular	intermediate	jazz	passing	page	lan	fermentation	kin	lighting	mat
lab	heart	learning	stage	football	user	wireless	world	competitive	composition	increase

Figure 1. Example distribution of topics output by a trained topic model

We considered two well-known methods methods for topic modeling: Latent Dirichlet Allocation and Non-negative Matrix Factorization.

3.3.1 Non-negative Matrix Factorization

The first method employed uses a novel interpretation of low-rank approximation of matrices. The Non-negative Matrix Factorization (NMF for short) method takes in a bag-of-words $n \times m$ matrix A whose entry A_{ij} is the number of occurrences of word j in document i. NMF seeks to find the closest approximation of A as a product of a $n \times k$ matrix W and a $k \times m$ matrix H with the condition that all entries of W and H are non-negative. In other words, NMF outputs W, H with the prescribed dimensions such that $\|A - WH\|_F$ is minimized.

To interpret the output matrices W and H, we call W the *basis matrix* and H the *coefficient matrix*. The prescribed number k here represents the number of topics we wish to extract from the text. To find out the top words are associated with the i-th

topic, we examine the i-th row of the H matrix and take the words whose position in that row has a high value. In other words, the entry H_{ij} measures how relevant word j is for topic i. Next, if we want to find out what topics are the most prevalent in the i-th corpus document, we examine the i-th row of the W matrix and take the topics whose position in that row has a high value. Although there are methods to perform this task with unseen documents that were not used in training, we omit their discussions here as they were not used for our experiments.

3.3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation, commonly known as LDA, is a generative probabilistic model that prescribes each document in a corpus with a finite mixture of topics from an underlying topic distribution. LDA views each document in a corpus as a generated item from a collection in order to infer the topic distribution. The generative process that LDA uses to infer the underlying topic distribution represents the topic distribution θ_m for each document m as a random variable from a Dirichlet distribution with sparse priors, where each topic is a distribution over all of the words. For each of N words in document m, a topic $z_n \sim \text{Multinomial}(\theta)$ is chosen and a word w_n is chosen from a multinomial distribution conditioned on z_n. LDA requires the modeller to input number of topics. In our experiments, we built our LDA models from LdaModel in the Python `gensim` package.

3.4 Training & Selecting Topics

We first train both LDA and NMF models on all Oxford College syllabi from 2001 to 2014 with various numbers of topics (50, 100, 150, and 200). Within LDA and NMF each, we compare the topics generated and select the best model by plotting and manual inspection of the topics. Then, comparing our best models from both NMF and LDA we determine which model outperforms the other. From this best model, we hand-select 43 topics with which to proceed in our analysis.

3.5 Assessing Change Through Time

We aim to assess the changes in courses offered over time using the topic model. To do so, we classify a syllabus by the topic that yields the highest relevance over the hand-selected topics. This is possible due to the topic distribution feature both techniques possess. For each year, we assess what proportion of syllabi was most related to each topic. In the next section, we will present graphs of these analysis as well as the hand-selected topics.

4. Results and Discussion

4.1 Selected Model

After comparing topic coherence and manually observing the topics output by LDA and NMF for 50, 100, 150, and 200 topics, we decided that NMF with 100 topics gave the best results. Even the best LDA model gave much worse topics than NMF which is surprising since LDA is often the model more preferred in literature (as was the case with both (Hall et al., 2008) and (Sekiya et al., 2017)). However, on small and sparse corpuses, NMF does as well and oftentimes better than LDA as shown by (Chen, Zhang, Liu, Ye, & Lin, 2019). For this reason, it is perfectly reasonable that NMF was more successful in our task of under 4,000 one-to-two-page syllabi.

4.2 Selected Topics

The hand-selected 43 topics are displayed in the table below.

Topic Description	Word 1	Word 2	Word 3	Word 4	Word 5
Anatomy	dissection	lab	physiology	anatomy	laboratory
Anthropology	anthropology	park	culture	cultural	archaeology
Art	studio	drawing	color	charcoal	art
Astronomy	laboratory	astronomy	observation	universe	heavens
Biology	biology	lab	genetics	laboratory	scientific
Botany	field	plant	woody	trip	identification
Child Development	development	child	discussion	childhood	group
Classical Studies	metamorphoses	homer	myth	mythology	tragedy
Dance	dance	ballroom	folk	cha	cultural
Economics	economic	march	policy	demand	market
Environmental Science	environmental	ozone	lab	stream	science
Ethics	ethics	ethical	morals	utilitarianism	philosophy
Finance	accounting	financial	business	time	assets
French	sur	pour	dissertation	reprise	lire
Geology	geology	earth	lab	laboratory	geologic
German	mitt	german	thema	die	sie
Gerontology	aging	aged	dying	death	surrounding
Golf	golf	game	score	chipping	swing
Health	fitness	activity	physical	training	running
Linear Algebra	linear	algebra	differential	matrices	problem
Literature	fiction	poetry	portfolio	short	march
Logic	logic	reasoning	categorical	syllogism	ordinary
Mandarin Chinese	dialogue	workbook	character	mandarin	cheng
Martial Arts	tai	skill	chi	practice	form
Mathematics	gateway	calculus	trigonometric	derivative	logarithmic
Media Studies	screening	film	cinema	reserve	sound
Meteorology	lab	weather	climate	meteorology	atmospheric
Music Education	music	musical	western	concert	classical
Musical Performance	dress	rehearsal	black	concert	music
Philosophy	philosophy	philosophical	philosopher	reverse	ken
Physical Education	cycling	fitness	indoor	workout	physical
Poetry	poetry	workshop	mid	poem	story
Political Philosophy	republic	utopia	book	political	politics
Political Science	politics	political	science	syllabus	international
Probability and Statistics	statistics	statistical	probability	data	hypothesis
Proofs	mathematics	theory	mathematical	landau	analysis
Racket Sports	singles	play	smash	badminton	net
Spanish	leer	lunes	antes	para	las
Swimming	pool	swim	swimming	water	underwater
Theater	theater	play	theatrical	performance	production
US Government and Politics	political	federalist	march	federalism	bureaucracy
Weight Training	lift	weight	training	fitness	muscle
Western History	art	ancient	architecture	paleolithic	aesthetic

4.3 Historical Trends in Emory University Courses

We now present some plots of syllabus frequency over time. Specifically, we examine examples of topics that increase or decrease in prevalence. Due to the anomalous nature of the number of syllabi in recent years, we will restrict the time of observation from 2001 to 2014. To visualize the trend of a selected topic, we plot the proportion of syllabi that are best represented by that topic.

4.3.1 Topics Increasing in Prevalence

We see an increase in relative frequency of syllabi that are best represented by the *Biology* and Mandarin Chinese topics. Both observations could be explained by their emerging popularity as subjects. For example, more people might be realizing that

Mandarin is a useful language to learn as it is spoken by many people.

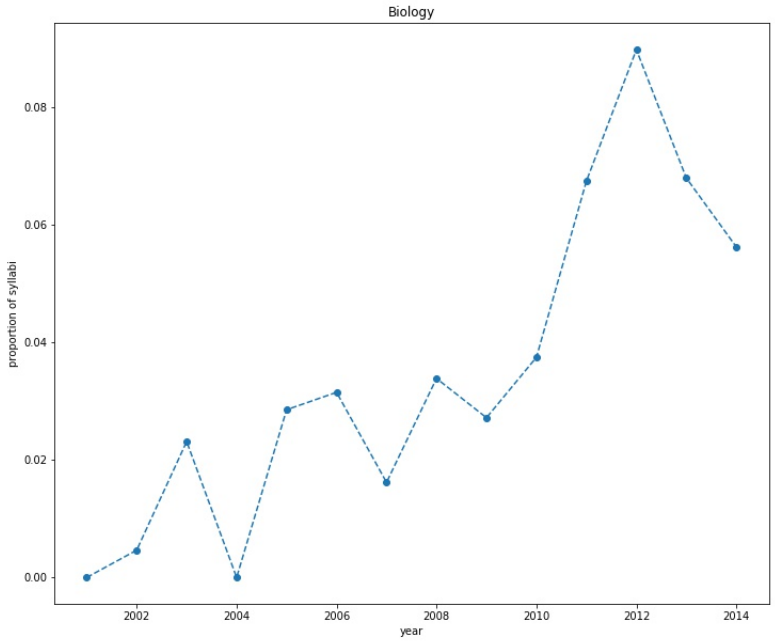

Figure 2. This plot displays the relative frequency of syllabi that are best represented by the *Biology* topic between 2001 and 2014.

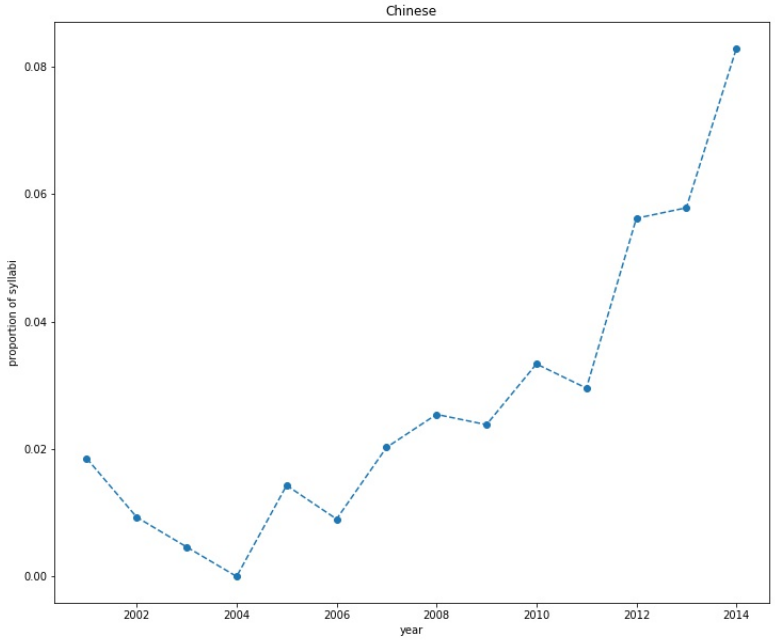

Figure 3. This plot displays the relative frequency of syllabi that are best represented by the *Mandarin Chinese* topic between 2001 and 2014.

4.3.2 Topics Diminishing in Prevalence

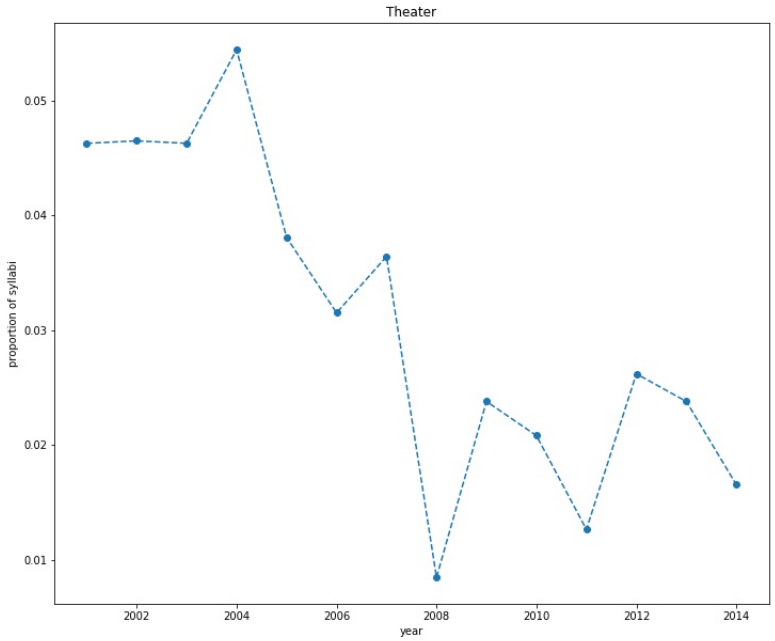

Figure 4. This plot displays the relative frequency of syllabi that are best represented by the *Theater* topic between 2001 and 2014.

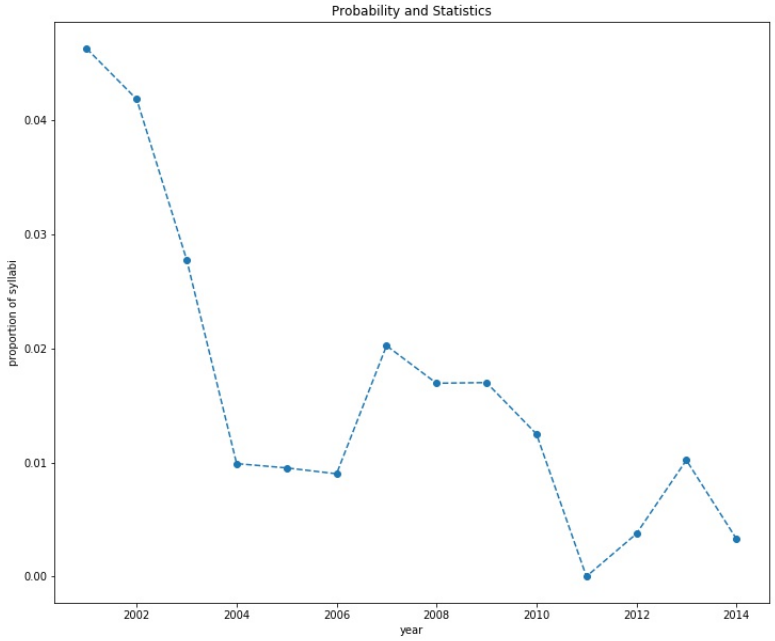

Figure 5. This plot displays the relative frequency of syllabi that are best represented by the *Probability and Statistics* topic between 2001 and 2014.

5. Conclusion

5.1 Remarks

Our research demonstrates the power of topic modeling to uncover the priorities in a university over time. At Oxford College of Emory University, we discover that certain subjects like Biology and Mandarin Chinese have increased in prevalence and certain subjects like Theater and Probability and Statistics have decreased in prevalence between 2001 and 2014.

5.2 Future Work

As our research presented in this paper figures into one part of a larger project with various other stakeholders and other priorities, we were under significant time constraints. As such, our choice of university was influenced by finding a public website we could easily web-scrape which also contained at least 10 years of syllabi. Although it would have been ideal to have had a larger corpus, our work establishes a reproducible pipeline which, upon acquiring more data from a source like the Open Syllabus Project, we would be able to compare multiple universities across time.

It would be interesting to perform the same type of analysis on a university offering thousands, rather than hundreds, of courses each year. It would also be fascinating to compare the syllabi of private and public universities, or even high schools to community colleges. Finally, Hall et al. were able to compare the papers from three Natural Language Processing conferences over time. Similarly, we could compare three or more universities over the same time frame. We were challenged to find comparable corpuses of syllabi freely available online, but with access to the Open Syllabus Project, this endeavor would be very feasible.

Another idea we discussed, though veered away from due to time limitations and sample size concerns, is analyzing how one course subject (such as literature, statistics, engineering, etc.) changes over time within a single school or across institutions. Say, for example, we focused on statistics. We could then train a topic model on a corpus of syllabi from only statistics courses and generate a number of topics from within statistics. Or, inspired by the work of Sekiya et al., we could find some pre-defined curriculum topics established an educational institution and use classify syllabi based on these topics. Expanding beyond the work of Sekiya et al., we could analyze how statistics courses have changed over time within a single institution or across a sample of universities across the country.

Acknowledgments

We greatly appreciate Derick Lee, the PilotCity local team, Dr. Talithia Williams, Dr. Weiqing Gu, and the Institute of Education Sciences for supporting our project.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study was financially supported by The Institute of Education Sciences.

References

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019, Jan). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, 1–13.

David M. Blei, A. Y. N., & Jordan, M. I. (2003). Latent dirichlet allocation. In *Journal of machine learning research.* Retrieved from `http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf`

Greene, D. (2017). *Parameter Selection for NMF.* Available online at `gitHub.com`.

Hall, D. L., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Empirical methods in natural language processing (emnlp).* Retrieved from `pubs/hall-emnlp08.pdf`

Le, Q., & Mikolov., T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning*, 1188–1196.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. In *In proceedings of the acl workshop on effective tools and methodologies for teaching natural language processing and computational linguistics. philadelphia: Association for computational linguistics.*

McCormick, C. (2018). *Word2Vec Tutorial - The Skip-Gram Model.* Available online at mccormickml.com.

Pedregosa, F. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12.*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora.

Sekiya, T., Matsuda, Y., & Yamaguchi, K. (2017, Oct). A web-based curriculum engineering tool for investigating syllabi in topic space of standard computer science curricula. In *2017 ieee frontiers in education conference (fie)* (p. 1-9). doi: 10.1109/FIE.2017.8190598

Shperber, G. (2017). *A Gentle Introduction to Doc2Vec - ScaleAbout - Medium.* Available online at medium.com.

Shuai, W. (2016). Topic Modeling and t-SNE Visualization. *Shuai's AI Data Blog.*.

Las Positas College. (2019). *Las Positas College CurricUNET.* Available online at `http://www.laspositascollege.edu/onlinelearning/faculty/distance_education/handbook/syllabus.php`.

Oxford College, Emery University. (2019). *Electronic syllabi.* Available online at `https://app.oxford.emory.edu/WebApps/Directories/EResources/index.cfm?FuseAction=List&Criteria=All`.

Xu, J. (2018). *Topic modeling with LSA, PSLA, LDA Lda2Vec - NanoNets - medium.* Available online at medium.com.