

Dartmouth College

## Dartmouth Digital Commons

---

Dartmouth College Undergraduate Theses

Theses, Dissertations, and Graduate Essays

---


Spring 6-3-2021

### Identifying optimal course structures using topic models

Tehut Tesfaye Biru

Tehut.T.Biru.21@Dartmouth.edu

Follow this and additional works at: [https://digitalcommons.dartmouth.edu/senior\\_theses](https://digitalcommons.dartmouth.edu/senior_theses)

 Part of the [Cognitive Psychology Commons](#), and the [Other Computer Sciences Commons](#)

---

#### Recommended Citation

Biru, Tehut Tesfaye, "Identifying optimal course structures using topic models" (2021). *Dartmouth College Undergraduate Theses*. 216.

[https://digitalcommons.dartmouth.edu/senior\\_theses/216](https://digitalcommons.dartmouth.edu/senior_theses/216)

This Thesis (Undergraduate) is brought to you for free and open access by the Theses, Dissertations, and Graduate Essays at Dartmouth Digital Commons. It has been accepted for inclusion in Dartmouth College Undergraduate Theses by an authorized administrator of Dartmouth Digital Commons. For more information, please contact [dartmouthdigitalcommons@groups.dartmouth.edu](mailto:dartmouthdigitalcommons@groups.dartmouth.edu).

# Identifying Optimal Course Structures Using Topic Models

Tehut T. Biru<sup>1,2</sup>

Honors Thesis Project

Professor Jeremy Manning<sup>2</sup>

May 10, 2021

<sup>1</sup>Department of Cognitive Science, Dartmouth College

<sup>2</sup>Department of Psychological and Brain Sciences, Dartmouth College



## **Acknowledgements**

This project would not have reached completion without the guidance of Professor Jeremy Manning, who provided expert and timely advice throughout the project. Members of the Contextual Dynamics Lab Caroline Lee and Paxton Fitzpatrick were also integral in overcoming obstacles encountered during this project. I would like to express special thanks for Caroline Lee who in addition to helping with the project, provided encouragement and support throughout. I would also like to thank Ugur Yavuz and Tibebu Biru, without whom some aspects of the project would not have come to fruition.

Further gratitude for Elshadai Biru, who helped provide some clarity of thought during complex stages of the project. I would like to thank my family, Rahel Bekele, Tesfaye Biru, Bezawit Biru and Hana Gebeyehu, without whose emotional support this project would not have been possible. Lastly, I would like to thank friends at Dartmouth College and extended family who provided me with support in various ways.

## **Abstract**

This research project investigates whether there exists an optimal way to structure topics in educational course content that results in higher levels of engagement among students. It is implemented by fitting topic models to transcripts of educational videos contained in the Khan Academy platform. The fitted models were used to extract topic trajectories across time for each video and subsequently clustered based on whether they have similar “shapes”. The differences in mean engagement metrics per cluster suggest that some course shapes are more palatable to students regardless of subject matter. Additionally, the topic trajectories suggest a constant progression of topics with little repetition is optimal for student engagement. The results from this project provide new methodologies to improve educational quality by focusing on the sequence of themes within instructional material.

*Keywords:* topic models, education, course effectiveness, natural language processing

# TABLE OF CONTENTS

<b>List of Figures.....</b>	<b>iii</b>
<b>List of Abbreviations .....</b>	<b>iv</b>
<b>1. Introduction.....</b>	<b>1</b>
1.1 Outline .....	2
<b>2. Literature review .....</b>	<b>3</b>
2.1 Word Embeddings.....	3
2.2 Topic Modelling Techniques .....	4
2.3 Education and Learning .....	5
<b>3. Methods.....</b>	<b>7</b>
3.1 Data Acquisition and Preprocessing .....	7
3.2 Analysis .....	9
<b>4. Evaluation and Results .....</b>	<b>11</b>
<b>5. Conclusion .....</b>	<b>21</b>
5.1 Discussion and Conclusion .....	21
5.2 Recommendations for Future Work .....	22
<b>References.....</b>	<b>22</b>

## List of Figures

Figure 1: Sample un-smoothed topic trajectory for a given video .....	13
Figure 2: Sample smoothed and resampled trajectory for a given video .....	13
Figure 3: Dot plot of topic trajectories colored by cluster .....	14
Figure 4: Dominant subject categories per cluster .....	15
Figure 5: Topic weights by cluster .....	16
Figure 6: Mean like, dislike and like to dislike ratio by cluster .....	17
Figure 7: Mean view count by cluster .....	17
Figure 8: Sample topic trajectory in cluster 7 .....	18
Figure 9: Average topic trajectory for cluster 7.....	18
Figure 10: Average topic trajectories for all 9 clusters .....	19

## List of Abbreviations

BAT	Bidirectional Adversarial Topic
BERT	Bidirectional Encoder Representation from Transformers
c-TF-IDF	class-based term frequency-inverse document frequency
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Algorithm
MCAT	Medical College Admissions Test
NCLEX-RN	National Council Licensure Examination for nursing
NLP	Natural Language Processing
NMF	Non-negative matrix factorization
USE	Universal Sentence Encoder



# Chapter 1

## 1. Introduction

Research in cognitive science has focused on topics ranging from how humans establish causality, how we draw conclusions, make decisions and more. Cognitive science also frequently draws from the field of computer science, with the classical view arguing that the mind is just a highly complicated computer involving the manipulation of formal symbols over a set of rules. (Cummins et al., 2000). A portion of cognitive science also investigates how people learn, how we encode and process information for various purposes. Computer science advancements, in addition to allowing for helpful analogies for the mind, have greatly increased our ability to assess and investigate the human learning processes and use it to abstractly map what we presume is occurring in the brain during learning. Together, these advancements can be valuable for making progress in educational practices as quality and accessible education is important for quality of life, access to opportunities and increasing our knowledge as a species. This thesis project aims to use computer science tools to learn more about how to structure educational content to ensure optimal learning.

The general approach followed in this project is to model course content using word embedding models. The idea of these models is to represent each concept as a feature vector, in which conceptually related words are nearby (i.e., in Euclidean distance). When we apply these models to each moment of a course video, we can obtain a trajectory, or “path”, through word embedding space that describes how the content of the course unfolds over time. This gives each course a unique geometric shape that can be compared or contrasted with other courses and leveraged to predict how someone will learn.

## Chapter 1:

For this Honors Thesis project, these word embedding models will be applied to a pool of approximately ten thousand course videos sourced from the Khan Academy platform (Khan Academy, n.d.), spanning a wide range of subject areas. My goal is to understand if particular classes of geometric shapes, normalizing for the specific subject area of each course, tend to be more or less effective at conveying knowledge to students, maintaining engagement, promoting interest, and so on. Discovering general patterns in effective courses' shapes that apply across subject areas could provide the first formal models of what makes an effective lecture (or teacher).

### **1.1 Outline**

This first chapter of this project work is an Introduction which includes background and purpose. The second chapter is the Literature Review, which presents relevant work on word embeddings, topic modeling techniques and education learning. Chapter three is the Methods, where the approach, data collection, data preprocessing, topic modeling method applied, and analysis are presented. Moreover, in chapter four, the evaluation and results are presented and discussed. The last chapter includes discussion, and recommendations for future work.

# Chapter 2

## 2. Literature review

### 2.1 Word Embeddings

Word embedding models help us create models of semantic content to extract meaning from words or texts and transform the field of computational linguistics. Latent Semantic analysis is a word embedding model that extrapolates global knowledge from local co-occurring data in a body of text (Landauer and Dumais 1997). Relatedly, Latent Dirichlet allocation is one of the most popular topic modelling methods that uses probabilistic models to identify the topics that a document belongs to based on the words in that document (Blei et al., 2003). Additional examples of word embedding models include Word2vec, which is a two-layer neural net that takes a text corpus as input and uses vectorization to output feature vectors that represent words in that corpus (Mikolov et al. 2013; Mikolov, Yih, and Zweig 2013) , and the Universal Sentence Encoder (USE), that creates numerical representations of textual data by encoding data into embeddings (high dimensional vectors) and is useful for transferring to other NLP methods such as text classification, semantic similarity and clustering (Cer et al., 2018). Relatedly, Top2vec utilizes distributed representations, joint document, and word semantic embedding, to identify topic vectors without having to specify the number of topics or address stop-word removal and lemmatization issues (Angelov, 2020).

The above mentioned word embedding models have various applications such as web search (Ding et al., 2004), online shopping (Niemann, Mochol, and Tolksdorf, 2009), unsupervised video

semantic partitioning (Quemy, 2018) and building cognitive models of learning and memory (Heusser, Fitzpatrick, and Manning, 2018)

## **2.2 Topic Modelling Techniques**

Topic modelling approaches such as the Correlated Topic Model, Pachinko Allocation Model, Time-based Topic Models, and Self-Aggregating Topic Models are also better able to track changes of topics over time, the correlation between topics and handle short amounts of text (Vayansk & Sathish, 2020). To address potential redundancy in topics when features of a dataset are biased, one topic modelling approach uses a nonnegative matrix factorization (NMF) model - usually used for dimensionality-reduction, latent topic modelling, and clustering tasks - and incorporates user-designed seed word supervision (Vendrow et al., 2021).

Pre-trained models help to get more accurate representations of words and sentences. They allow for more coherent outputs by providing more contextual knowledge. Bidirectional Encoder Representation from Transformers (BERT) is a pre-trained encoder stack that incorporates contextual word embeddings (different meanings based on words in close proximity), an attention mechanism to understand context, and feature extractor to achieve state-of-the-art results in various NLP tasks (Devlin et al., 2018). One related technique BERTopic uses BERT and class-based term frequency-inverse document frequency (c-TF-IDF), a statistical method to assess how relevant a word is to a document, to create dense clusters and increase the interpretability of topics (Maarten, 2020). Neural topic modelling uses black box inference methods with neural networks and one version, the Bidirectional Adversarial Topic (BAT) model, uses a two-way projection between document-topic distributions and the document-word distribution to improve on the coherence of topics extracted (Wang et al., 2020). Research on Contextualized Topic Models built with two components, a neural topic model and pre-trained models (SBERT embedded

representations), has shown that contextualization results in more meaningful and coherent topics (measured via human evaluation) (Bianchi, Terragni & Hovy, 2020).

Topic modelling algorithms have also been employed in education for various purposes. Previous applications have focused on educating patients about their treatment (Kandula, 2011), improving search engines for educational videos (Danciulescu, 2020) and conducting student assessments (Harden, 2019). Topic models have also been used to customize information presented to students based on models of learning and memory (Stern and Woolf, 2000). Researchers at the Contextual dynamics Lab at Dartmouth College used a similar approach to model the dynamic content of a television episode (Heusser, Fitzpatrick, and Manning, 2020). The main goals of that work were to compare how people’s memories of the television episode related to the original experience of watching the episode, and also to understand which brain regions were involved in learning and remembering that experience. This project work uses a similar modeling framework in order to understand whether there are universal “shapes” of online courses that tend to be more effective at teaching students and keeping them engaged and interested.

### **2.3 Education and Learning**

The relationship between presentation of learning material and how people learn has also been investigated from various perspectives. One such perspective that is largely refuted is the notion that students have different learning styles, and these learning styles can be classified based on sensory modalities into visual, auditory, and kinaesthetic. However, there is a lack of evidence to support these learning styles and the modification of teaching mediums to cater to different styles has not been shown to produce significant improvement in learning outcomes (Pashler et al. 2008; Geake, 2008). Although there is not much support for these learning styles, brain data indicates

## Chapter 2: Literature review

that modality of perceived information is significant for the way in which information is processed in our brain, with the fusiform gyrus active for pictorial stimuli and the supramarginal gyrus for verbal components (Kraemer, 2009).

Other ways to structure content to become more digestible to the human brain includes distributing learning (studying) time over a larger duration and interleaving (alternating between) different subjects (Rohrer, 2010). One project attempted to inform school curricula by building cognitive models of mathematics processing in students (Ritter, 2007). Some suggest that attention has a more crucial role in learning and consequently, due to the limitations of working memory, there should be decreased focus on the perceptual aspects of learning and more focus on frequent testing to improve information retention over a longer period of time (Miller, 2001). Additionally, the level of control given to students in the learning process, direct (guided) instruction vs. discovery (unguided) instruction is suggested to have different outcomes for learning. One study indicated that those in direct instruction conditions performed as well as or better than discovery instruction conditions and that these results indicate that some direct instruction is needed when young children attempt scientific investigation tasks (Klahr, 2004). The difference between physical and virtual learning materials has also been explored, with reports of different learning outcomes for these different learning materials. Proponents of physical (hands-on) activities for science learning suggest that it is essential for developing an experiential and concrete background for science learning (Flick, 1993). Some support the idea of embodied mathematics learning, which suggests that mathematics knowledge is perception and action based - supporting this claim using data from teacher's gestures while explaining mathematical concepts (Alibali, 2012). However, research done with engineering design projects for middle school students did not find any significant performance differences for those using physical materials as opposed to virtual and researchers

## Chapter 3: Methods

indicate that the practical benefits of virtual learning materials might make it a preferred method of instruction (Klahr, 2007).

In terms of teacher characteristics, studies have indicated that teachers who are better able to empathize with students and judge students' engagement levels to adjust content accordingly are rated as more effective (Jordan and Schwartz, 2018; Lavy 2015; Curran 2018). Overall, good course design and engagement have both been shown to increase satisfaction with classes, performance, and student focus (Black, 2014; Hosler, 2012; Richards, 2011).

# Chapter 3

## 3. Methods

### 3.1 Data Acquisition and Preprocessing

This project uses transcripts of videos contained in the Khan Academy YouTube channel. Data was acquired using the YouTube download library (Amine, 2011) which extracts text and time intervals for each set of words in the video. The total number of videos downloaded numbered 10,069 with a total size of 137.4 Mega Bytes. For each video, two files with the same file name were downloaded of file type Web Video Text Track (file extension '.vtt') and JavaScript Object Notation (JSON, file extension '.json'). The Web Video Text Track file contained the video transcript and time points while the JSON contained a dictionary with video details such as number of likes, dislikes, view counts and more which will be used to extract engagement data later during analysis. The videos vary in grade level (difficulty) and cover a wide range of subjects.

### Chapter 3: Methods

The video transcripts were pre-processed before analysis using the Natural Language Toolkit (Bird et al., 2009). Initial data cleaning and preparation involved lowercasing all text, removal of punctuation and tokenization - which separates lines of text into a list of words or terms - in order to standardize across all of the text and remove components that are not relevant for analysis. Subsequently, three separate files were outputted for each video transcript. The file primarily used for analysis pairs each word with the specific time point (in seconds) it occurred on. This was achieved using the time intervals saved with each series of words, taking the difference of the time intervals, and dividing it by the total numbers of tokens (words) in that interval. The second and third file output contain the (cleaned) text of the video and the part of speech tag of each word in the text.

The final text to be analyzed will also have stop words removed (words that appear frequently and do not contribute significant meaning to texts, e.g. 'the', 'so', 'and', 'with' etc...) and lemmatized (returning words to their base form while taking into account relevant prefixes, suffixes and part of speech tags for increased accuracy).

Five transcripts had frequent errors in pre-processing due to uncommon symbols and several misspelled words. These files were not included in analysis. Fifty transcripts were also removed because they did not cover educational content. Of those removed, twenty-one were talks, five were meetings with professionals, six were information for parents, eight were related to guided meditation and ten were help videos (helping khan academy, helping students figure out study strategies etc...). These videos were identified using keywords 'talk', 'meet', 'kickoff', 'meditation', 'parent', and 'help'.



### 3.2 Analysis

For analysis, a second set of text preparation had to be performed. Each time point document was converted into a data frame with its indices as the time points and column values as the corresponding words at that time point. These time points were also converted from milliseconds to seconds and cut off to two decimal points. To extract meaning from the data later, the file names were also stored alongside their contents and tracked throughout analysis. The main analyses for this project were implemented using the SciKit Learn library, which contains tools for statistical modelling and unsupervised machine learning (Scikit-learn: Machine Learning in Python).

In order for a topic model to be fitted, data must first be vectorized, i.e. turned into a document by word matrix distribution. To pass into the CountVectorizer function, each document was converted into one string with the elements separated by space. Parameters of minimum difference of 0.1 (words which appear in less than 10% of the documents are disregarded) and maximum difference of 0.25 (words which are particularly frequent or occur in more than 25% of the documents are disregarded) were also provided to the CountVectorizer. This function will provide a fitted vectorizer object that has learned the words in the corpus and a matrix of token counts (document by word matrix).

The matrix generated from the CountVectorizer function was used to fit a Latent Dirichlet Algorithm model. For this model, the learning method 'online' was used (faster for larger data sizes) and a prespecified number of 50 topics was provided. The purpose of this initial fit is for the model to learn all of the words in the corpus and their topic distribution. To obtain the topic trajectories and get shapes of videos, this fitted LDA model was transformed over sliding time windows of the video transcripts.

### Chapter 3: Methods

To obtain topic trajectories over the course of the video, each video string was divided into sliding time windows of 120 seconds durations and middle time points of each sliding window were recorded. This would save the words contained within the 120 seconds interval as one string and each successive window would only differ in the first and last words. These sliding windows were constructed so that co-occurring words can be considered together to improve topic identification and create topic vectors for each successive duration as the video progresses.

Using these sliding windows, the initial vectorizer object was transformed and the matrix returned was passed to the initial LDA model using the transform function. This allows for the identification of topic distributions of the sliding windows based on the initial model constructed using the entirety of the data. These topic trajectories were organized into a dataframe along with the middle time points the topic trajectories occurred on. The data frames were then resampled to the desired sample rate to ensure that all data have an equal length/distribution (fixed number of rows). Resampling the frames is useful for hyperalignment later and was implemented using the Scipy Python library's (Virtanen, 2020) pchip (Piecewise Cubic Hermite) interpolator which uses cubic splines to find the value of new points in the data.

For the subsequent analysis of hyperalignment and clustering using HDBSCAN, the data manipulation and visualization package hypertools was used (Heusser, 2017). The topic trajectories in the data frames were then aligned in a high dimensional space. The goal of hyperalignment is to create a shared information space in which the feature vectors that contain information about the topic trajectories are aligned. The aligned topic trajectories were then grouped using a data-driven way of automatically determining how many clusters there are and which videos go in which cluster, as well as which videos don't seem to fit into any particular

cluster. To test statistical significance, paired t-tests were performed comparing the view counts, like counts, dislike counts, and like/ dislike ratio of videos in each cluster.

# Chapter 4

## 4. Evaluation and Results

The top ten words for the 50 topics specified (shown in Table 1) indicate how well the model generated is working to capture the themes within the video transcripts. This is done through a qualitative assessment to check whether words in the same topic are conceptually related. From topic 36, we see words such as particle, kinetic, cycle, molecule, which are all related in physics concepts and we see that words in topic 42 are conceptually related to governmental institutions.

**Table 1: The top 10 words for each of the 50 topics identified in the model are shown above.**

**Qualitative assessments of the conceptual relationship between these words indicates whether our model worked well in capturing themes in the corpus.**

<b>Topic 0:</b>	statement, true, integer, david, difference, inequality, compare, woman, dog, great
<b>Topic 1:</b>	light, sun, relative, ice, surface, hot, small, level, room, water
<b>Topic 2:</b>	pound, eat, apple, store, code, sugar, food, paper, fruit, piece
<b>Topic 3:</b>	bank, buy, dollar, pay, price, sell, market, asset, income, money
<b>Topic 4:</b>	social, child, life, feel, family, society, study, friend, play, person
<b>Topic 5:</b>	distance, rate, minute, foot, car, mile, travel, average, round, hour
<b>Topic 6:</b>	20, 25, 30, 15, 50, 40, 60, subtract, 80, 100
<b>Topic 7:</b>	khan, academy, school, teacher, kid, learn, year, grade, great, student
<b>Topic 8:</b>	fourth, exponent, tenth, raise, hundredth, sixth, decimal, eighth, pattern, power
<b>Topic 9:</b>	group, bond, hydrogen, ring, molecule, double, methyl, chain, ahead, carbon
<b>Topic 10:</b>	base, natural, cost, total, scenario, sentence, clause, orange, green, log
<b>Topic 11:</b>	derivative, respect, prime, partial, interval, input, constant, rule, maximum, function
<b>Topic 12:</b>	root, squared, positive, polynomial, coefficient, quadratic, formula, simplify, radical, square
<b>Topic 13:</b>	american, state, empire, united, history, period, power, year, north, war
<b>Topic 14:</b>	heart, pressure, vessel, muscle, lung, oxygen, body, artery, left, blood
<b>Topic 15:</b>	company, business, job, spend, bar, plan, project, technology, build, share
<b>Topic 16:</b>	velocity, speed, ball, acceleration, initial, position, direction, final, momentum, meter
<b>Topic 17:</b>	neuron, brain, membrane, inside, motor, nervous, action, error, movement

## Chapter 4: Evaluation and Results

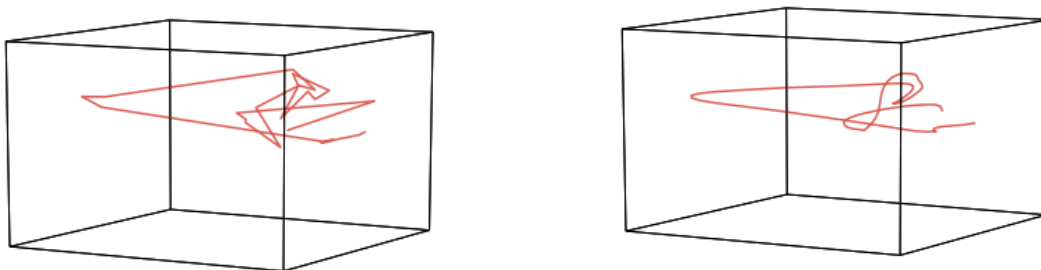
<b>Topic 18:</b>	height, volume, rectangle, centimeter, length, surface, base, width, unit, potential. area
<b>Topic 19:</b>	open, ion, calcium, inch, channel, chloride, sphere, close, salt, sodium
<b>Topic 20:</b>	13, choice, 14, 16, 18, problem, 11, 23, 24, 12
<b>Topic 21:</b>	limit, infinity, state, president, party, large, vote, close, closer, approach
<b>Topic 22:</b>	hole, black, building, city, wood, walk, stone, sort, leg, voiceover
<b>Topic 23:</b>	solution, solve, subtract, absolute, lefthand, problem, cancel, satisfy, righthand, equation
<b>Topic 24:</b>	day, month, pay, age, option, period, rate, tax, 10000, year
<b>Topic 25:</b>	half, represent, blue, large, left, parallel, pause, split, copy, place
<b>Topic 26:</b>	head, wire, flip, tail, switch, coin, iron, cup, connect, probability
<b>Topic 27:</b>	set, random, define, box, sum, possibility, pick, element, sequence, variable
<b>Topic 28:</b>	wave, protein, signal, body, frequency, red, node, layer, release, cell
<b>Topic 29:</b>	learn, problem, class, computer, math, game, model, watch, didn't, sal
<b>Topic 30:</b>	price, curve, distribution, population, cost, average, standard, data, quantity, sample
<b>Topic 31:</b>	read, book, story, language, text, sound, passage, today, character, word
<b>Topic 32:</b>	oxygen, hydrogen, proton, charge, acid, bond, reaction, molecule, pair, electron
<b>Topic 33:</b>	coordinate, axis, inverse, comma, shift, real, vertical, plot, xaxis, graph
<b>Topic 34:</b>	zucker, roman, greek, beth, steven, ancient, rome, sculpture, frame, harris
<b>Topic 35:</b>	matrix, row, column, dot, transformation, guy, space, plane, product, vector
<b>Topic 36:</b>	particle, kinetic, period, cycle, molecule, glucose, high, spring, potential
<b>Topic 37:</b>	speaker, capital, growth, week, labor, grow, machine, large, size, ratio
<b>Topic 38:</b>	current, concentration, constant, delta, voltage, pressure, heat, gas, reaction, temperature
<b>Topic 39:</b>	human, star, population, gene, animal, type, life, virus, bacteria, dna
<b>Topic 40:</b>	sine, theta, integral, evaluate, antiderivative, series, unit, function, definite, cosine
<b>Topic 41:</b>	field, direction, newton, block, electric, net, pull, gravity, magnitude, force
<b>Topic 42:</b>	government, law, constitution, court, amendment, congress, federal, house, supreme, state
<b>Topic 43:</b>	object, circle, image, distance, earth, mirror, ray, parabola, reflect, center
<b>Topic 44:</b>	paint, steven, art, male, beth, female, artist, woman, music, painting
<b>Topic 45:</b>	triangle, degree, length, circle, 90, congruent, 180, opposite, measure, angle
<b>Topic 46:</b>	symptom, patient, type, infection, test, person, drug, disorder, brain, disease
<b>Topic 47:</b>	mole, gram, unit, kilogram, liter, weight, atomic, atom, proportional, mass
<b>Topic 48:</b>	denominator, expression, numerator, common, simplify, divisible, 2x, fraction, rewrite, factor
<b>Topic 49:</b>	slope, increase, charge, decrease, density, sign, blue, close, cross, positive

Topic trajectories for each video were obtained by building a feature vector for sliding windows of one-hundred twenty second durations. Feature vectors for conceptually-related words are nearby one another. This allows us to map the semantic content of the video over time and obtain

## Chapter 4: Evaluation and Results

a trajectory, or “path” through word embedding space that describes how the content of the course unfolds over time. This gives each course a unique geometric shape that can be compared or contrasted with other courses (across subjects), leveraged to predict how someone will learn, and so on. Since the videos are different lengths, each video was resampled to 100 time points and a smoothing procedure was applied to even out some of the bumps in trajectories. The differences between the initial trajectories and the resampled and smoothed ones are shown using a sample video’s trajectory in Figure 1 and Figure 2. As compared to the trajectory of Figure 1, the smoothed trajectory in Figure 2 has significantly less sharp corners with smoother curves. This allows us to standardize each trajectory for easier comparison later.

After obtaining the trajectories for each video, the videos are clustered based on similarity in topic trajectories. Using HDBSCAN, nine clusters were identified in the transcripts with a 10th cluster, -1, containing all the videos that were not able to be clustered. The videos in cluster -1 were removed from graphing and analysis in order to better focus on those videos that had clear distinct categories in topic trajectories.

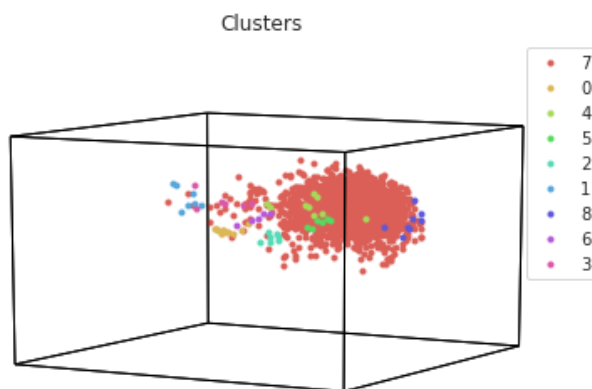


**Figure 1, left, and Figure 2, right, show un-smoothed and smoothed topic trajectories for a sample video. Figure 2 is the resulting trajectory after videos have been resampled to 100 time points and contains less sharp turns and smoother curves.**

## Chapter 4: Evaluation and Results

By projecting the word embedding vectors into a 3D space, we represent each video's topic trajectory as a point in a hyperdimensional space. The values plotted are categorized by cluster (Figure 3); and the clusters are signified by different colors.

We can see that videos with similar topic trajectories are nearby to each other and thus colored the same, and the spatial differences imply that there is a clear delineation between the clusters in terms of progression of conceptual content. The differences in the number of dots by cluster also indicates that the clusters are of different sizes.

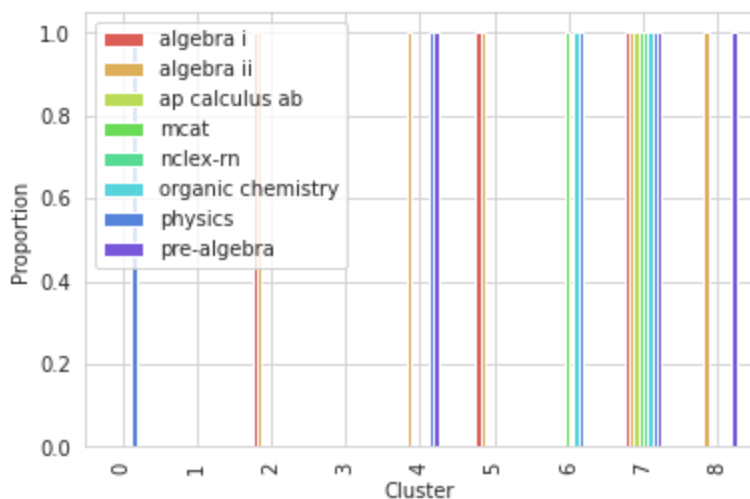


**Figure 3: Dot plot of topic trajectories colored by cluster. The plot shows that videos clustered together are nearby to each other and while videos in different clusters are far apart.**

The video categories by cluster (Figure 4) indicate that there is a difference in the predominant subject category in each cluster. Cluster 1 and 3 have no values here because the size of the clusters, or the number of videos in the clusters, was too small to have the minimum count of videos in a category required to be captured in this graph. We see that clusters 0, 2, 4, 5, and 8 contain predominantly physics and algebra while cluster 6 and 7 contain predominantly organic chemistry, MCAT (medical college admissions test), and nclex-rn (nursing license examination) with a few other categories mixed in. One way to interpret this figure is to recognize the distinct subject areas

## Chapter 4: Evaluation and Results

in each cluster and that these subject areas may also be interrelated or interdependent. For instance from the data, we can identify the subject area MCAT (Medical College Admission Test) in cluster 7 and recognize that the other subject areas enclosed in that cluster (organic chemistry, AP calculus AB, physics, algebra and more) are often subject areas required for and related to the MCAT. This is relevant because it can be used to identify optimal topic progressions that may result in better performance on the MCAT. Physics concepts also require a basis in mathematics, which cluster 0 and 4 demonstrate by being composed of physics and mathematics videos predominantly.

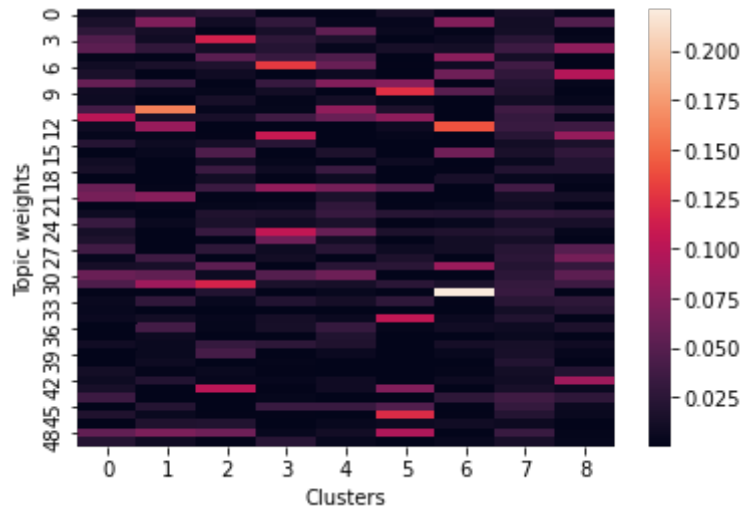


**Figure 4: shows the predominant subject categories by cluster. While cluster 1 and 3 are too small to be captured in this plot, other clusters show a clear delineation in the dominant categories.**

We can also further expand this using the initial 50 topics generated from our model, and create a heatmap to show the relative weight of each topic in each cluster (Figure 5), which intersecting with information about subject categories in each cluster can help us evaluate our model and unpack each cluster further. Taking cluster 4 as a sample to investigate our model, we see that topics 11 and 12 are the most weighed in this cluster. Both topics cover mathematics and physics concepts and this is in line with the subject categories in cluster 4, which predominantly

## Chapter 4: Evaluation and Results

cover pre-algebra, algebra 2 and physics. Hence, a qualitative assessment drawn from these two types of details from each cluster can emphasize how well the model is working in capturing concepts.



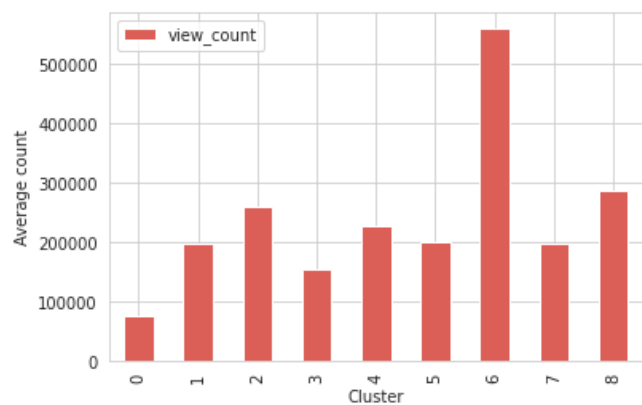
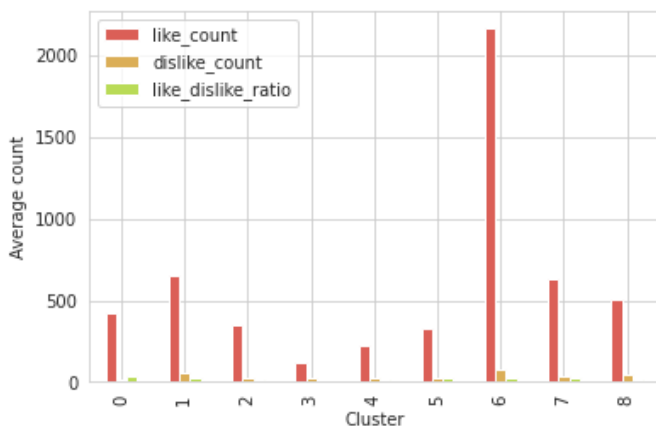
**Figure 5: The topic weights by cluster are shown in the heat map above, where lighter colors indicate a heavier presence of that topic in that cluster. This allows us to cross-check the subject categories in each cluster using the highest weighted topics in that cluster.**

Subsequently, the engagement metrics per cluster can demonstrate if particular clusters are more popular than others. Plotting the average like count, dislike count and like to dislike ratio for each cluster (Figure 6) shows that different clusters have different mean values, suggesting that some clusters are more popular than others. Cluster 6 is the most liked, while cluster 3 is the least liked, indicating that the topic trajectories within cluster 6 are more popular among students. Performing unpaired, two-tailed t-tests shows that a number of these differences in means are statistically significant. For mean like count, cluster 3's value was significantly different from cluster 5's value,  $t(11) = -2.65, p = 0.02 < 0.05$ . For dislikes, the mean value of cluster 6 was significantly different from cluster 8,  $t(9297) = -3.08, p = 0.002 < 0.05$ . Several of the differences in mean like to dislike



## Chapter 4: Evaluation and Results

ratios were also statistically significant. Cluster 0 was significantly different from cluster 2 and 7 with values of  $t(7) = 2.47$ ,  $p = 0.04 < 0.05$  and  $t(1) = 2.39$ ,  $p = 0.03 < 0.05$  respectively. Cluster 5 had significantly different like to dislike ratios with clusters 2, 3 and 7 with values  $t(6) = -6.32$ ,  $p < 0.001$ ,  $t(1) = -2.53$ ,  $p = 0.03 < 0.05$  and,  $t(9) = 4.02$ ,  $p = 0.003 < 0.05$  respectively. Lastly, cluster 6 had a significantly different mean value from cluster 8,  $t(9296) = 2.65$ ,  $p = 0.008 < 0.05$ . These results indicate that the differences in mean like to dislike ratios occur above chance for a number of clusters. The like to dislike ratio also controls for the varying total like and dislike count for each cluster and expresses the proportion of those who liked the video to those who disliked it. Therefore, some ‘shapes’ of courses or trajectories of concepts in a course are more palatable to students when compared to others.



**Figure 6, left, and Figure 7, right show the average like, dislike, like to dislike ratio by cluster and the mean view count by cluster. The differences between these mean values indicate that some course shapes produce higher engagement.**

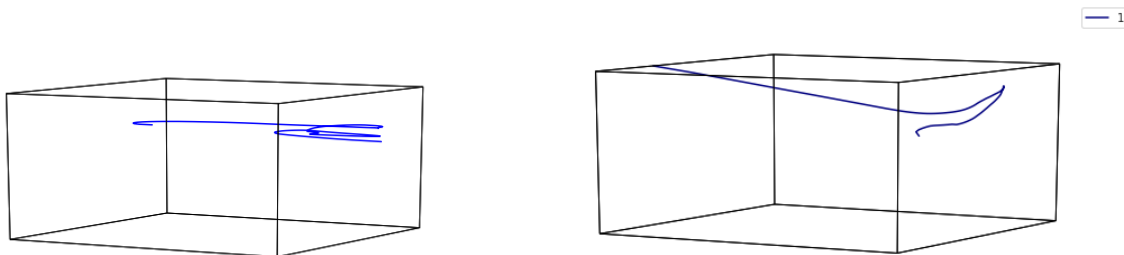
Looking at the mean view counts by cluster (Figure 7), we can see that cluster 6 is similarly popular by having the highest view count by far, while cluster 0 has the least number of view counts. The

## Chapter 4: Evaluation and Results

mean view count of cluster 5 is statistically different from those of cluster 6 and 8, with values  $t(2918) = 2.22, p = 0.02 < 0.05$  and  $t(6889) = 2.16, p = 0.03 < 0.05$  respectively.

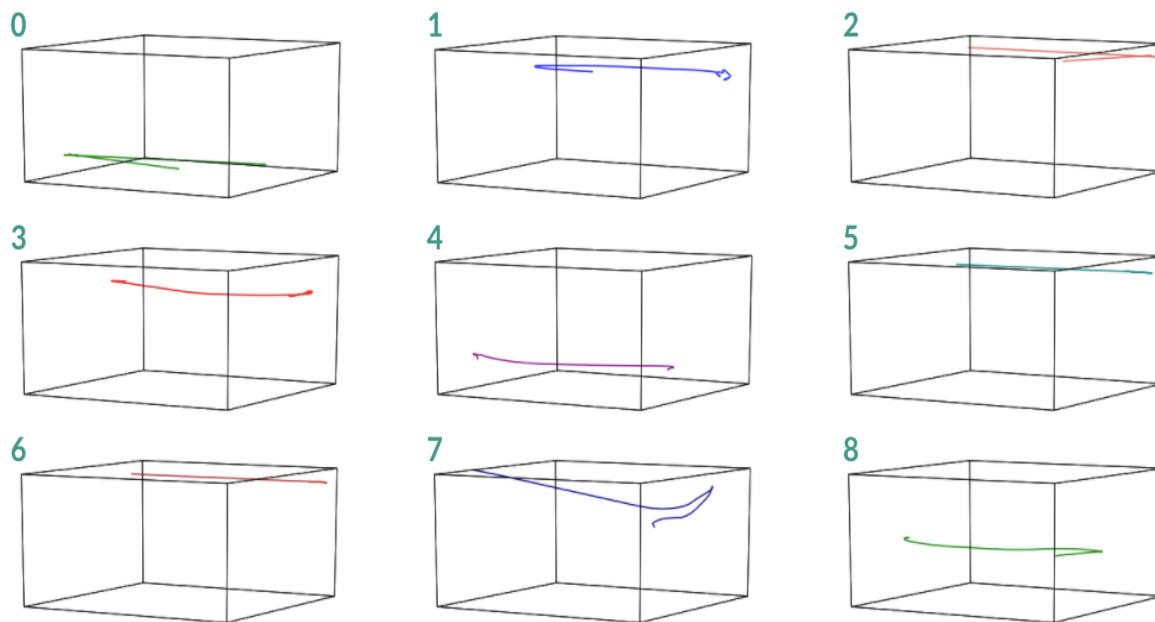
These differences in means indicate that the distinction between the course shapes of clusters matters for how popular they are among students. We can unpack this more by taking a closer look at a few clusters.

We can plot the average trajectory for each cluster and make comparisons between the structures of the topic trajectories to make inferences about why they have different engagement metrics. Figure 8 and Figure 9 show a sample video trajectory from cluster 7 and the average topic trajectory for that cluster. We can see that the average topic trajectory (Figure 9) has much less winds and loops when averaged across all videos within that cluster and captures a simple structural trajectory.



**Figure 8, left, and Figure 9, right, show a sample topic trajectory from cluster 7 and the average topic trajectory from that cluster. The average trajectory represents a simplified version as seen by the reduced variations in the trajectory.**

Below are the average topic trajectories for the 9 clusters identified from the videos (Figure 10). Some clusters, such as cluster 6 and cluster 5 seem to have a somewhat smooth progression of topics while other clusters like cluster 0 and 1 have more bends, indicating a sharp change in topics or revisitation of similar topics.



**Figure 10: Average topic trajectories for each of the 9 clusters identified. We can see a clear variation in the trajectories with, for instance, clusters 0, 1, 7 having more bends and clusters 5 and 6 having somewhat straight lines.**

Using these average trajectories, we can look into the most popular cluster, cluster 6, and cluster 8 which has a significantly less like to dislike ratio as compared to cluster 6. The average topic trajectory for videos in cluster 6 seems to be a somewhat straight line, which can indicate a smooth progression of topics throughout the videos without much repetition or sudden switches in the themes or topics. We can make this inference because if there were a revisiting of concepts, we would expect to see the trajectory have some form of looped shape in certain sections. The predominant video categories within this cluster are organic chemistry, physics, and test preparation for the MCAT (Medical college admissions test), and the highest weighted topics within this cluster, topics 13 and 32, can give some insights on why these videos are liked. Topic 13, which contains words such as state, empire, united, history, year, north, can be explained by

## Chapter 4: Evaluation and Results

the fact that the MCAT has a verbal reasoning section which can include literature or history text, and topic 32, which contains the words electron, oxygen, hydrogen, proton, charge, is in line with the video categories of physics and chemistry as well. So instructional videos on these subject categories tend to have a smoother progression in topics, which is more popular among students and produces more engagement. Alternatively, the less popular cluster 8 predominantly contains videos on algebraic mathematics. Looking at the average topic trajectory for this cluster, the main difference we can notice is a sharp change in topics at one end of the videos. This can indicate that some topics are being revisited in close proximity to each other. This can be interpreted to mean that this revisitation can result in educational videos not being as effective or engaging and suggesting that instructional videos in these subject categories mostly have fewer smooth ends and are not as popular.

# Chapter 5

## 5. Conclusion

### 5.1 Discussion and Conclusion

Although the videos on the Khan Academy platform cover a large range of concepts over the 10,000 videos, the shapes of the courses tend to fall in a few groups. The mean engagement metrics showed that some course shapes are more popular, liked and interacted with more than others. Results also showed the videos with a constant progression of topics, with less loops or sharp changes in themes, tend to be more effective at engaging students. Alternatively, course shapes that tend to revisit concepts in close temporal proximity to each other, tend to be less effective at engaging students, as shown by the engagement metrics by cluster. This would suggest that certain “shapes” of courses, irrespective of subject area, are more palatable or satisfying to students than others. This is interesting because it indicates a possibility of tweaking courses to be more effective at engaging students, by reordering topics (maybe in terms of themes in syllabi or lecture slides) to best match a particularly optimal trajectory shape, without changing the overall conceptual content of the courses.

This project is limited in that course effectiveness may not be fully captured by the engagement metrics used in this study. There could be other confounding variables that can lead to some courses being more popular than others, for instance because videos on some subject categories are more in demand than others. The engagement metrics used also do not capture student’s engagement throughout watching the video. In an experimental setting, the use of eye-tracking and brain data to measure attention and engagement during the video would better indicate overall engagement levels for each course video. Moreover, an assessment of students’ comprehension

through frequent evaluations would provide further information on the effectiveness of some shapes of course videos when compared to others.

### **5.2 Recommendations for Future Work**

Future research could also utilize pre-trained language models to improve vectorization. This will allow for fine-tuning of models already trained with large data sets for specific NLP tasks, which makes the process faster and produces better models. For example, this project used Latent Dirichlet Allocation to build the topic models; however, it can be beneficial to explore this project's goals in alternative ways. Methodologies like the Correlated Topic Model, Time-based Topic Models, and others which can better track changes of topics over time, the correlation between topics and handling short amounts of text could provide more insight in this topic.

Moreover, further calibration of hyper-parameters passed to the model constructors could improve their performance. Subsequent research can also focus on non-video based educational materials and whether the effectiveness of in-person, collaborative and other learning environments can be evaluated using natural language processing methodologies.

Other methods to expand this project are to go beyond engagement metrics and investigate if some course shapes produce more comprehension in students than others. Additionally, this methodology can be used to investigate if the formulation of individualized lesson plans to adhere to distinct ways of consuming information between students can optimize learning.

Access to the highest quality education can substantially improve quality of life and access to opportunities. Education is also central to expanding our knowledge as a species. Advances in education technology, learning theory, pedagogy, and related fields also improve the quality of

## Chapter 5: Conclusion

education. Applications of these technologies can help us address educational inequity and ensure students have the best educational experiences.

## References

- Amine, R. et al. (2011). youtube-dl, Github, <https://github.com/ytdl-org/youtube-dl>
- Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. arXiv preprint arXiv:2008.09470.
- API Reference | YouTube Data API | Google Developers. (n.d.). (Retrieved October 07, 2020), from <https://developers.google.com/youtube/v3/docs>
- Bianchi, F., Terragni, S., & Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. arXiv preprint arXiv:2004.03974.
- Bird, S., Loper, E., & Klein, E. (2009), Natural Language Processing with Python. O'Reilly Media Inc. <https://www.nltk.org/>
- Black, G. S., Daughtrey, C. L., & Lewis, J. S. (2014). The Importance of Course Design on Classroom Performance of Marketing Students. *Marketing Education Review*, 24 (3), 213-226. doi:10.2753/mer1052-8008240303
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225-255.
- Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.
- Contextual Dynamics Lab. (n.d.). (Retrieved October 07, 2020), from <http://www.context-lab.com/>



- Curran, Sam. 2018. What Makes an Effective Teacher? GRIN Verlag.
- Cummins R. & Cummins, D.D. (2000) Minds, Brains, and Computers: The foundations of cognitive science. Blackwell.
- Danciulescu, T.I., Mihaescu, M.C., Heras, S., Palanca, J., & Julián, V. (2020). More Data and Better Keywords Imply Better Educational Transcript Classification?  
[https://educationalLDAtaming.org/files/conferences/EDM2020/papers/paper\\_98.pdf](https://educationalLDAtaming.org/files/conferences/EDM2020/papers/paper_98.pdf)
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Ding, Li, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, V. C. Doshi, and Joel Sachs. "Swoogle: A semantic web search and metadata engine." In Proc. *13th ACM Conf. on Information and Knowledge Management*, vol. 304, pp. 10-1145. 2004.
- Flick, L. (1993). The Meanings of Hands-On Science. *Journal of Science Teacher Education*, 4(1), 1-8, from <http://www.jstor.org/stable/43155957>
- Geake, J. (2008). Neuromythologies in education. *Educational Research*, 50(2), 123-133. doi: 10.1080/00131880802082518
- Harden, M. (2019). First-Year Students and the Framework: Using Topic Modeling to Analyze Student Understanding of the Framework for Information Literacy for Higher Education. *Evidence Based Library and Information Practice*, 14 (2), 51-69.  
doi:10.18438/eblip29514
- Haxby, J., Guntupalli, J., Nastase, S., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, 9.  
<https://doi.org/10.7554/eLife.56601>

- Heusser, A. C., Fitzpatrick, P. C., & Manning, J. R. (2018). How Is Experience Transformed into Memory? *bioRxiv* 409987.
- Heusser, A. C., Ziman, K., Owen, L. L., & Manning, J. R. (2017). HyperTools: a Python toolbox for gaining geometric insights into high-dimensional data. *The Journal of Machine Learning Research*, 18(1), 5589-5594.
- Heusser, A. C., Fitzpatrick, P. C., & Manning, J. R. (2020). Geometric models reveal behavioral and neural signatures of transforming naturalistic experiences into episodic memories. *Life Science Weekly*, 1838. doi:10.1101/409987
- Hosler, K., & Arend, B. (2012). The importance of course design, feedback, and facilitation: student perceptions of the relationship between teaching presence and cognitive presence. *Educational Media International*, 49(3), 217–229.  
<https://doi.org/10.1080/09523987.2012.738014>
- Jacobi, C., Atteveldt, W. V., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4 (1), 89-106.  
doi:10.1080/21670811.2015.1093271
- Jordan, J. V., & Schwartz, H. L. (2018). Radical empathy in teaching. *New Directions for Teaching and Learning*, 2018(153), 25-35.
- K-means clustering algorithm. (n.d.). (Retrieved October 07, 2020), from  
<https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm>
- Kandula, S., Curtis, D., Hill, B., & Zeng-Treitler, Q. (2011). Use of topic modeling for recommending relevant education material to diabetic patients. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2011*, 674–682.
- Khan Academy. (n.d.). Home [YouTube Channel]. YouTube. (Retrieved January, 2020), from

<https://www.youtube.com/user/khanacademy>

- Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661–667.
- Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching, 44*(1), 183–203.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*(2), 211.
- Lavy, V. (2015). What Makes an Effective Teacher? Quasi-Experimental Evidence. *CESifo Economic Studies, 62*(1), 88–125.
- Maarten, G. (2020). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics, v0.7.0, Zenodo, doi:10.5281/zenodo.4381785
- Mikolov, T., K. Chen, G. Corrado, and J. Dean. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint, 1301.3781*.
- Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 746-751*.
- Miller, M.D. (2011). What college teachers should know about memory: a perspective from cognitive psychology. *College Teaching, 59*(3), 117-122.  
doi: 10.1080/87567555.2011.580636
- Niemann, M., Mochol, M., & Tolksdorf, R. (2009). Smart Shop Assistant—Using Semantic

- Technologies to Improve Online Shopping. In Future Internet Symposium. Springer, Berlin, Heidelberg, 106-115.
- Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological science in the public interest*, 9(3), 105-119.
- Quemy, A., Jamrog, K., & Janiszewski, M. (2018). Unsupervised Video Semantic Partitioning Using IBM Watson and Topic Modelling. *EDBT/ICDT Workshops*.
- Richards, G. (2011). Measuring engagement: Learning analytics in online learning. Retrieved from [http://www.academia.edu/779650/Measuring\\_Engagement\\_Learning\\_Analytics\\_in\\_Online\\_Learning](http://www.academia.edu/779650/Measuring_Engagement_Learning_Analytics_in_Online_Learning)
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, 14(2), 249–255.
- Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
- Stern, M. K., & Woolf, B. P. (2000). Adaptive content in an online lecture system. In *International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*. Springer, Berlin, Heidelberg, 227-238.
- Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582.
- Virtanen, P. et al. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272. <https://www.scipy.org/>
- Vendrow, J., Haddock, J., Rebrova, E., & Needell, D. (2020). On a Guided Nonnegative Matrix Factorization. *arXiv preprint arXiv:2010.11365*.
- Wang, R., Hu, X., Zhou, D., He, Y., Xiong, Y., Ye, C., & Xu, H. (2020). Neural topic modeling with bidirectional adversarial training. *arXiv preprint arXiv:2004.12331*.