Topic Modeling #STEM Dialogue from Twitter

Byron Biney

Research Overview & RQs:

Dataset of 198,030 tweets collected between March 14th 2018 and June 3rd 2018 (Tweepy API, Karbasian, H.)

Each tweet mentions the hashtag “#STEM” or keyword for STEM

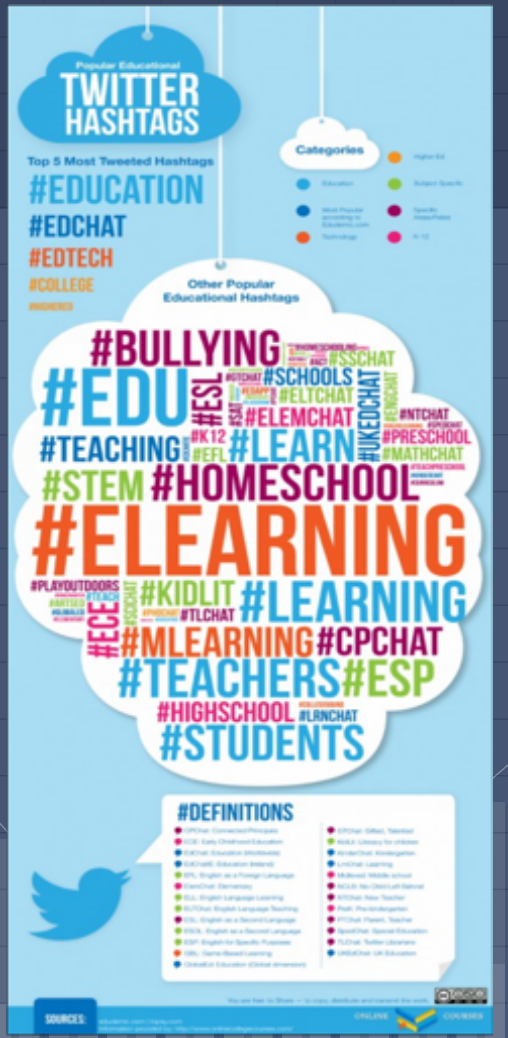
Used LDA Topic modeling algorithm to cluster tweet text into distinct topic distributions

What do people talk about in the #STEM hashtag?

What conversational patterns take place within the #STEM hashtag?
What are the most popular topics to discuss?
What terms and phrases are likely to reoccur among topics?
What do topic modeling results reveal about the scope and relevance of STEM in culture and society?

Research Motivations

- Few studies on textual content of #STEM hashtag
- Few studies summarizing and examining conversations about education
- Few studies looking for data-mining social media solely from an educators' perspective, not from a supply-chain oriented view



Related Projects: STEM Education on Social Media

Yuen & Pickering (2016): STEM Conversations

- Social Network Analysis of #STEM tweets spread throughout year
- Findings: Content analysis using hashtags absence of students despite target marketing of educational resources for students

Chen et al. (2014): #EngineeringProblems

- Content Analysis of 19000 tweets made at Purdue University
- Findings: Many posts centered around depicting high levels of student stress and lack of free time.
- Naive-Bayes Multi-label Classifier for posts about engineering problems

Related Work: Theoretical Framing

1. Definition of social media
2. Social Code, the expectations of human sociality in social media
3. Convergent Media Culture
4. Networked Individualism

Related Work: Social Media and Culture

Defining Social Media:

Social media includes a variety of internet-based tools that users engage with by maintaining an individual profile and interacting with others based on network of connections (Xenos, 2014)

Followers Lists and newsfeed content are self curated *within* constraints of adaptive algorithms and site design

Users don't just articulate networks and make friends lists out of thin air, they are presented to the user by a combination of their interaction with the website and their personal choice (

Related Work: Social Code

Code understood as the instructional techniques that enable digital materials to form…

Code can be understood in more than just technical dimensions.

Williamson (2015): Writing code is a technical process but generated through social processes and human actions. Production of code reflects technical aims in Computer Science and entrepreneurial + social domains in software development.
Williamson (2015): Software development is concerned with the production of tangible artefacts and computational science treats the computer as a scientific instrument.

Facebook is built upon particular understandings of people as socially networked beings.

'people you know' algorithm seeks to optimize users' network sociality by establishing a kind of algorithmic normality for social relations. The concepts underpinning such algorithms articulate their own 'authoritative' accounts of humans as social networking creatures.
Williamson (2015) Terms such as "participatory culture," "networked individualism," etc… express something natural about human sociality as if evolution has actually demanded social network media.

- Mass media sources such as newspapers and televisions and personal communications intersect in the landscape provided by social media (Williamson, 2015).
- Intersection leads to a participatory culture surrounding content creation.
- Convergent media cultures are simultaneously conducive to increased ownership and commodification among commercial media producers.

Convergent Media Culture

Related Work: Networked Individualism

Williamson (2015): Networked individualism is a culture that starts with the values and projects of the individual who interacts with others following their own choices, values, and interests, rather than by tradition and hierarchy.

Can lead to social change by the sharing of multicultural and cosmopolitan values
Can entrench individuals further into consumer-media culture

- Latour (2005): The social is not a thing or domain of reality; it does not explain, it is precisely what needs explaining. This is remarkably easy to forget, as social media platforms constantly suggest the opposite, take the social for granted, naturalize it, make the social equal happiness, inclusion, the good life.

Methods

Dataset: 198,030 tweets containing the hashtag "#STEM" or keyword
Preprocessing:

- Converted all characters to lowercase
- Omitted Spanish and English stopwords + punctuation
- Lemmatized each word to noun form

"No_filtering" enabled models: No additional preprocessing
"All_filtering" enabled models: Additional preprocessing, removed words occurring in >60% of tweets and <5%
Created 10 of each LDA model, varied topics from 10-40, increasing number 5 for each model

LDA Plate Notation

Larger rectangle M: total number of documents

N: Number of words within a document

Alpha: hyperparameter on the per-document topic distribution

Beta: hyperparameter on the per-topic word distribution

Theta: Topic distribution for document m

z: a topic

w: a word

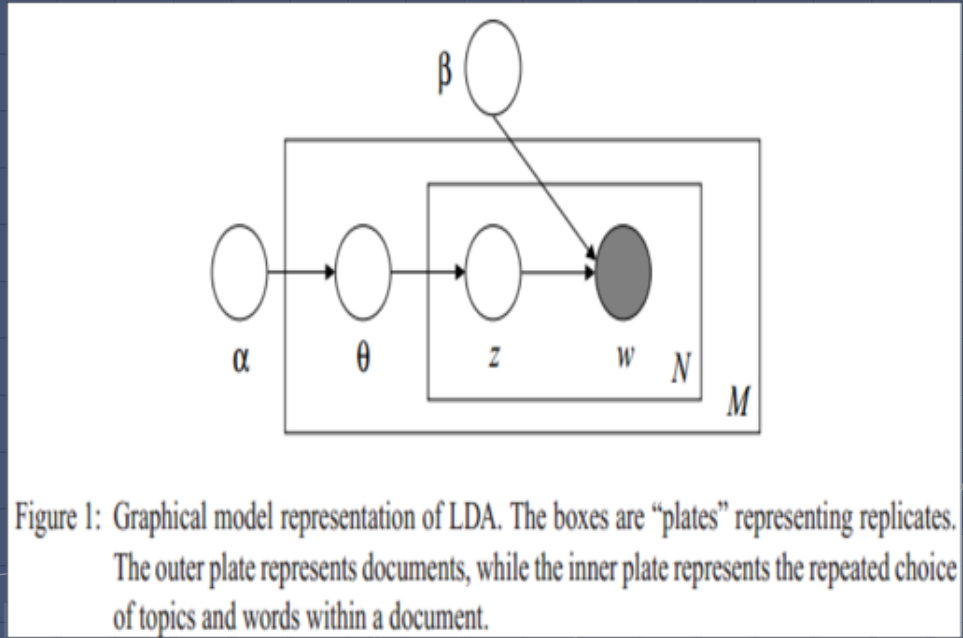
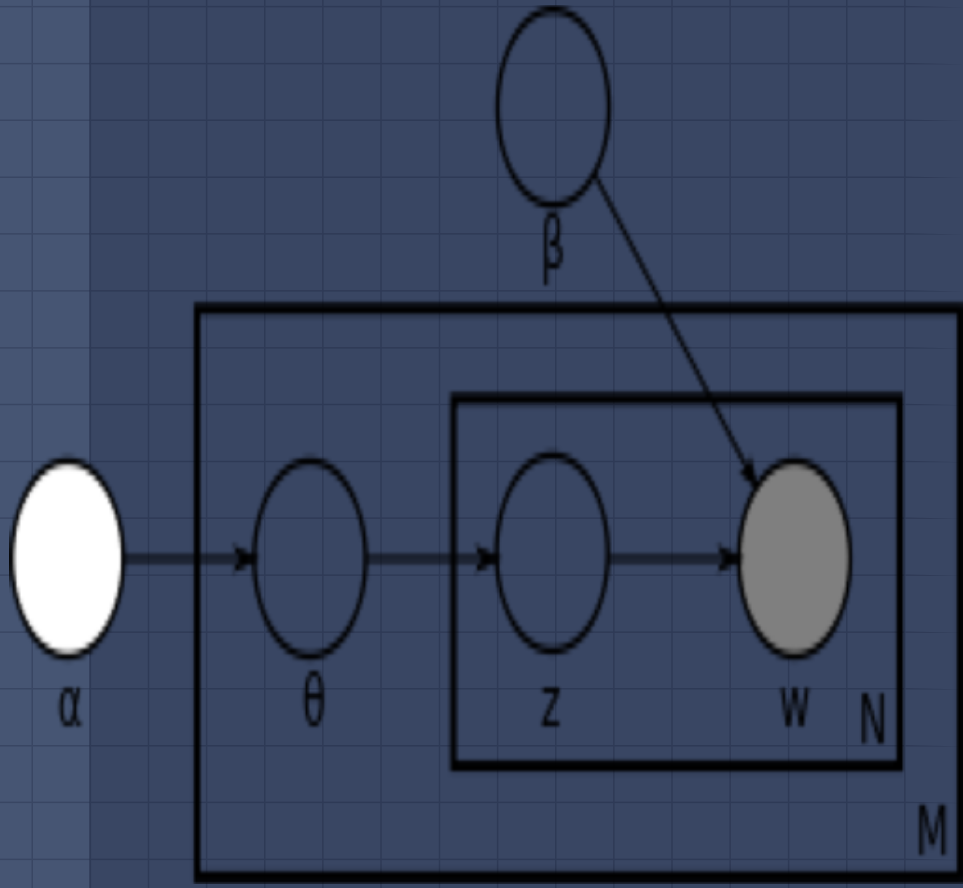


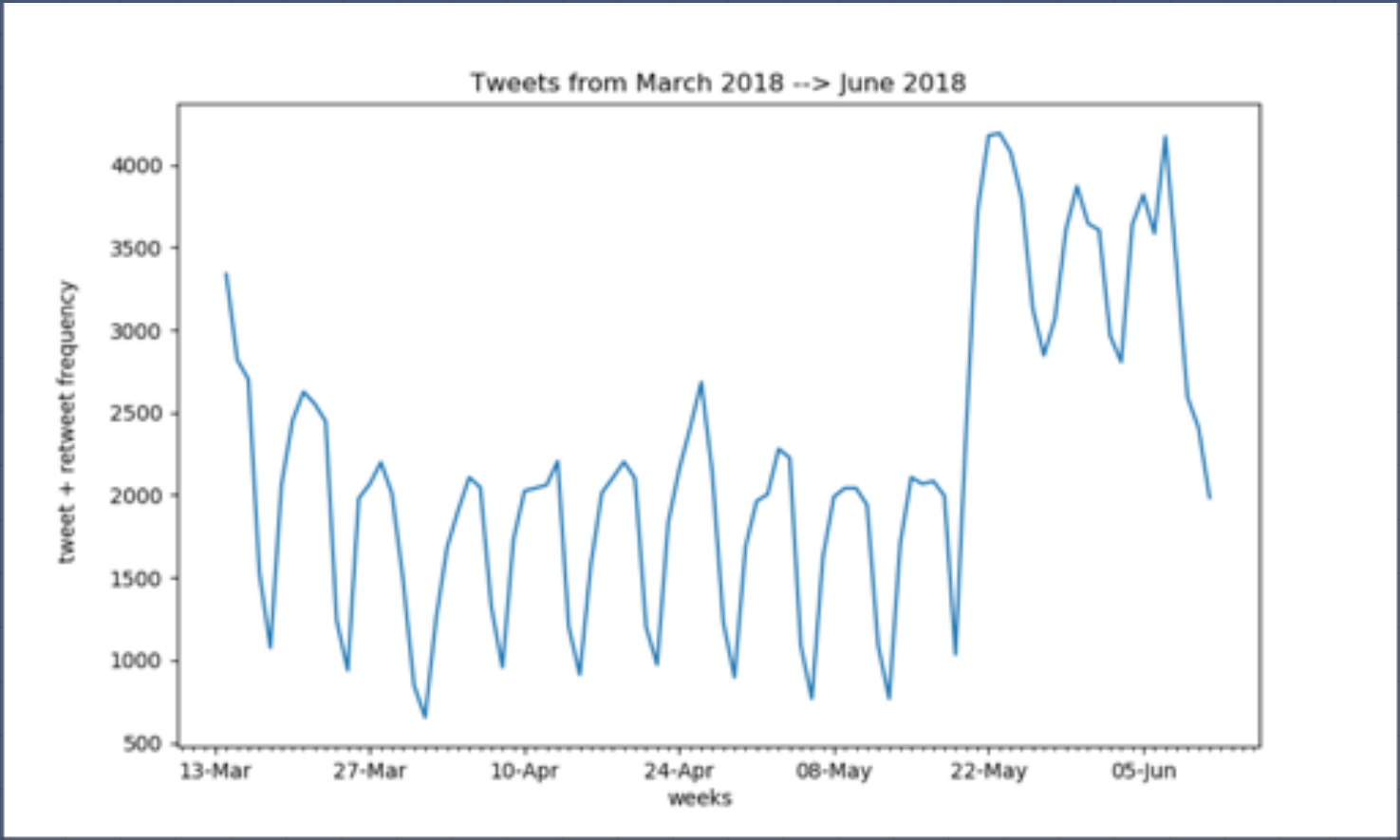
Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Generative Process for LDA

The LDA algorithm assumes that each document or tweet is produced by the following generative process:

1. Choose a number of words used a document
2. Choose a topic mixture over a fixed set of topics (i.e. topic distribution for documents).
3. Generate words by
    a. First, picking a topic from the document's multinomial distribution of topics...
    b. Then, picking a word based on the topic's multinomial distribution of words...

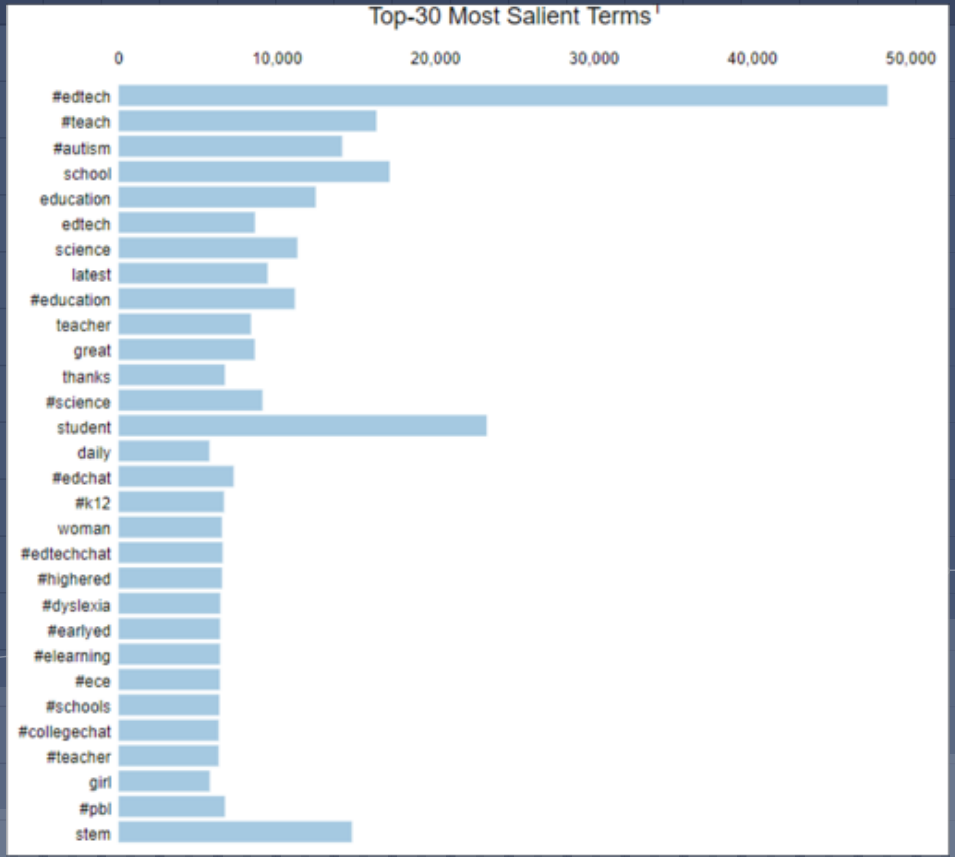




Temporal Visualization of #STEM posting

Descriptive Analysis Results

Total Tweets: *198,030*	Tweets with photo: *104,063 (52.49%)*	Average Word Count Per Post: *~13*
Verified User Count: *8,696 (4.39%)*	Tweets with video: *7,324 (3.698%)*	Unique Hashtags: *92,252 (46.58%)*
Tweets with URLs: *68,867 (34.78%)*	Tweets with GIF: *1717 (.867%)*	Tweets in English: *188,061 (94.97%)*

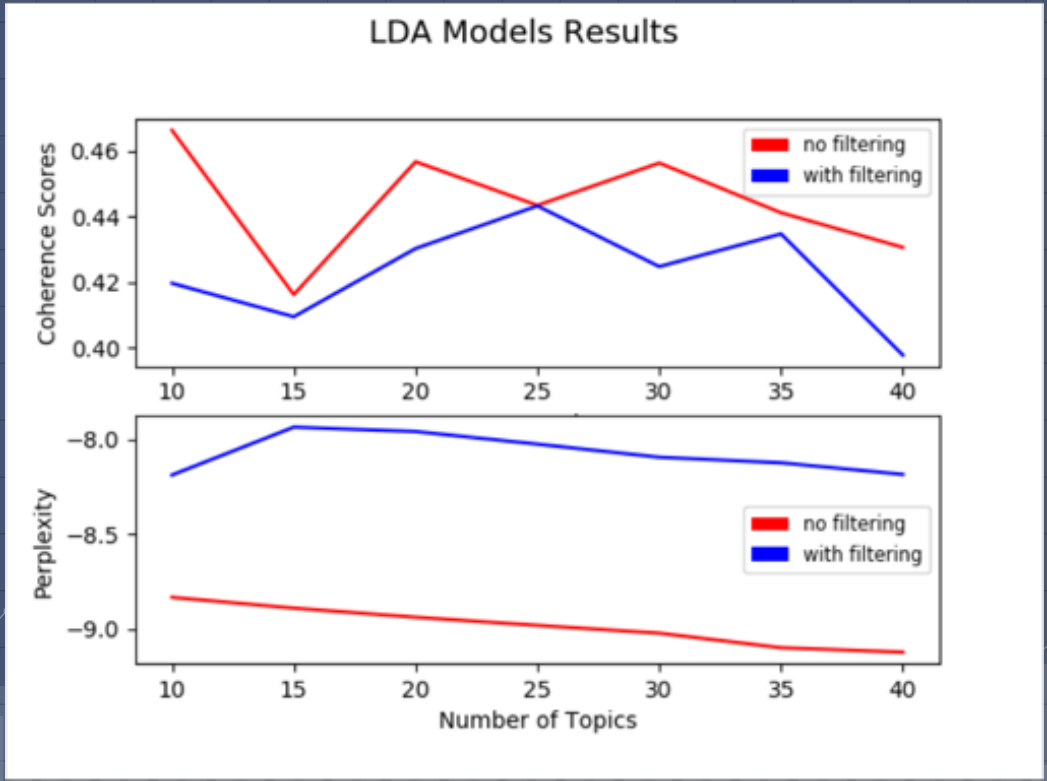


LDA Topic Model Metrics

Perplexity: A comparative metric for how well the theorized probability distribution by an LDA model compares to the actual distribution of words among texts. **The lower the perplexity, the better.**

Coherence Score: A comparative metric for representing how coherent a set of words is with one aggregated number value. **The higher the coherence score, the better.** The coherence metric uses the top n number of words (top_n_words) that LDA uses to represent each topic and computes the average pairwise similarity for each term in each topic. It considers all the ways in which words from the same topic can be paired, references their probability of occurring among the texts, and evaluates how much each topic semantically supports each word pairing with respect to its probability.

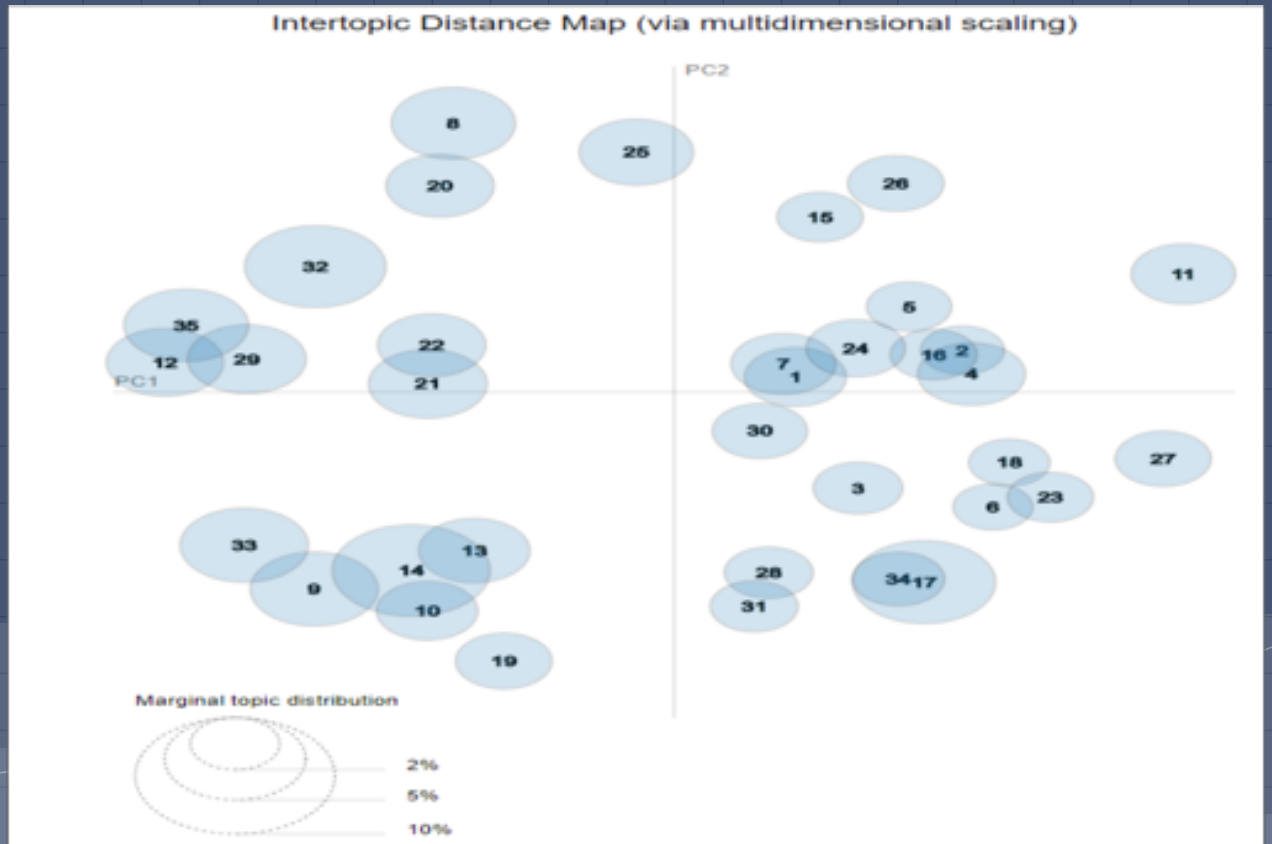
LDA Metrics Results



1st Finding: Metric Scores Indicate Noise of Twitter Data

From reviewing the data on the two metrics we recorded for our lda models, it isn't entirely clear which model is truly optimal. As the number of topics change, values were not observed as consistently increasing or decreasing. It is also surprising that the models with no filtering of frequent or infrequent words and phrases received better metrics (lower perplexity values and higher coherency scores) than the models that used such filtering.

pyLDAvis: Intertopic Distance



Interesting Findings & Discussion

STEM Conversation is usually about students

- When STEM is discussed, so are keywords like “education,” “student,” and “young”
- Stringing together a “culturally relevant pedagogy” from this data is not intuitive

Edtech plays a major role in representing STEM

- Edtech-based companies and blogs, individuals from edtech, even the keyword “edtech” is frequently mentioned throughout the hashtag
- Do entrepreneurial goals of Edtech and moral mission of pedagogy line up?

Interesting Findings & Discussion

Observation of extracurriculars in interdisciplinary areas relating to technology (robotics, biotechnical sciences) commonly promoted for “fun” and “play”

- The term “summer” and “camp” would frequently be given its own topic or a very high association with its topic word distribution
- Few studies on extracurriculars and long-term career trajectory. Data from Sahin (2017) found that one of the greatest predictors for STEM majors is perceived efficacy in math, not external programs and expectations.
- Binns & Conrad (2016) found that optional extracurriculars can increase confidence in STEM fields for juniors in HS

Interesting Findings & Discussion

Suggestion: Aggregating educational resources in the #STEM hashtag

- This topic modeling showed similar results to Yuen & Pickering (2016) in finding that a large amount of educational resources are given on this hashtag
- Usually updates about people's activities, not straightforward tutorials
- Propose a dashboard that can help educators mark certain content as providing "intellectual" and creative resources

Limitations and Future Work

- There were complications with analyzing the external urls of tweets. One difficulty is because the urls themselves may contain image, video, or content that cannot simply be text-mined due to its format. Another complication is due to the fact that many posts contain shortened urls which require more time to issue requests for unshortening en masse. Sites such as Instagram cannot be properly disseminated without analysis of other forms of media
- Content analysis and user profile analysis is necessary in order to fully understand the textual content of posts in this hashtag. As Chen et al. (2012) and other authors state, manual interpretation is crucial and pragmatic for social media analytics. Even optimal topic learning models can still display some of the ambiguities that were encountered from the word distributions.
- An optimal model was not easily identifiable, and filtering did not improve metric scores unanimously. This points to the noise of twitter data and also suggests that further or more elaborate preprocessing should take place with respect to spelling errors, short acronyms and slang commonly used in tweets, and other features.

References

1. http://journals.sagepub.com/doi/pdf/10.1177/2056305115578138
2. In progress for this slideshow! However…

1. Binns, I. Conrad, J. Student Perceptions of a Summer Venture and Mathematics Camp Experience. School, Science, and Mathematics. 116(8):420-429.
2. Chen, X. Vorvoreanu, M. Madhavan, K. 2014. Mining Social Media Data for Understanding Students' Learning Experiences. IEEE Transactions on Learning Technologies, 7(3):246-259.
3. Hogan, B. 2010. The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. Bulletin of Science, Technology, And Society. 30(6):377-386. Jacksonville, (FL).
4. Linvill, D. Brandon, B.C. Grant, W.J. 2018, Jan 29. "Back-Stage" Dissent: Student Twitter Use Addressing Instructor Ideology. Communication Education. 67(2):125-143
5. Pickering, T.A. Yuen, T.T. Wang, T. 2016. STEM Conversations on Social Media: Implications for STEM Education. IEEE. Bangkok, Thailand IEEE International Conference on Teaching, Assessment, and Learning Engineering
6. Sahin, A. 2017 Jun 22. The Relationships Among High School STEM Learning Experiences, Expectations, and Mathematics and Science Efficacy and the Likelihood of Majoring in College. 39(11):1549-1572.
7. Sievert, C & Shirley, K. 2014. LDAvis: A Method for Visualizing Topic Models. Paper presented at Proceedings of The Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore, (MA)
8. Williamson, B. 2013. The Future of Curriculum: School Knowledge in the Digital Age. Cambridge (MA). The MIT Press. Chapter 6: Globalizing Cultures of Lifelong Learning