Topic Modeling For Qualitative Content Analysis

ResBaz 2021

Dr Aneesha Bakharia
ITaLI
The University of Queensland
Email: aneesha.bakharia@gmail.com
Twitter: @aneesha

Topic Modeling

- Non-negative Matrix Factorization (**NMF**) and Latent Dirichlet Allocation (**LDA**) are two popular and useful algorithms that are able to find latent topics within document collections.

- While NMF and LDA stem from different mathematical underpinnings, both algorithms are able to map documents to topics and words to topics.

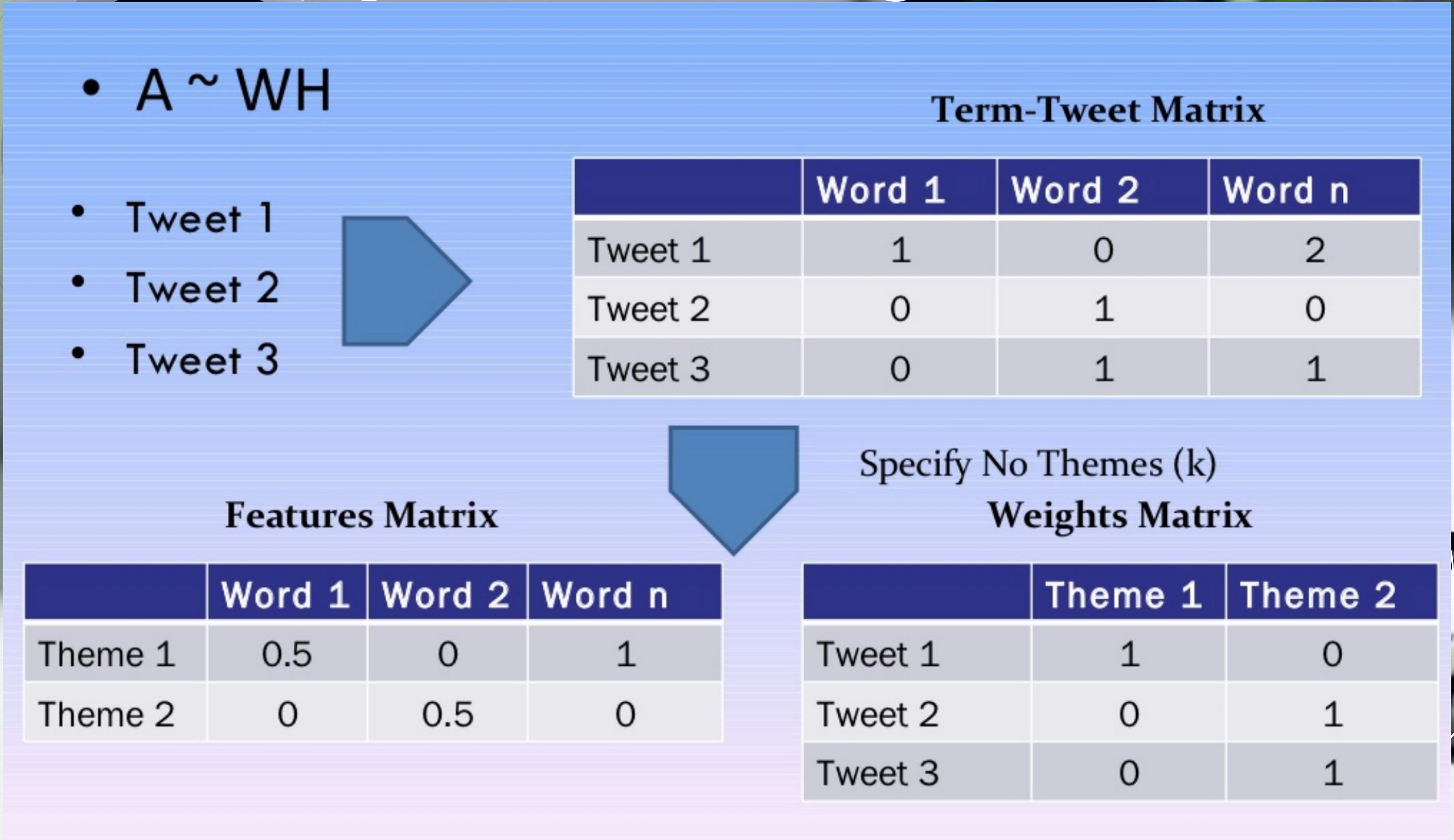
Surveys, Workshop Feedback, Interviews, Transcripts, Novels, Reports, Tweets, Blog comments

Types of Qualitative Content Analysis

Coding Approach	Study Begins With	Derivation of Codes	Algorithms
Summative	Keywords	Keywords identified before and during analysis	Unsupervised and semi-supervised algorithms: Non Negative Matrix Factorization (**NMF)**, Latent Dirichlet Allocation (**LDA),** neural inspired **Contextual Topic Model (CTM)** and traditional clustering algorithms.
Conventional (Inductive)	Observation	Categories developed during analysis	
Directed (Deductive)	Theory	Categories derived from pre-existing theory prior to analysis	Supervised classification algorithms: Support Vector Machines (SVM)

(Hsieh and Shannon, 2006)

Topic Modeling – NMF



- A ~ WH

- Tweet 1
- Tweet 2
- Tweet 3

Term-Tweet Matrix

	Word 1	Word 2	Word n
Tweet 1	1	0	2
Tweet 2	0	1	0
Tweet 3	0	1	1

Specify No Themes (k)

Features Matrix

	Word 1	Word 2	Word n
Theme 1	0.5	0	1
Theme 2	0	0.5	0

Weights Matrix

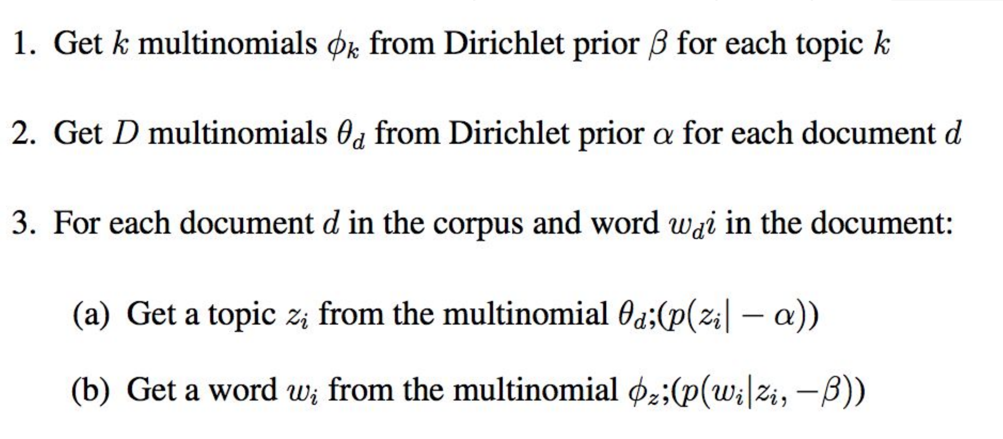
	Theme 1	Theme 2
Tweet 1	1	0
Tweet 2	0	1
Tweet 3	0	1

Topic Modeling – LDA

- Every document is a mixture of topics
  Eg Doc 1 is 90% Topic 1 and 10% Topic 2

- Every topic is a mixture of words
  Example
  Topic 1 on Politics,
  Topic 2 on Business
  Both Topic 1 and Topic 2
  share the word “Budget”

- LDA is a probabilistic model to estimate both
- words in a topic and documents in a topic



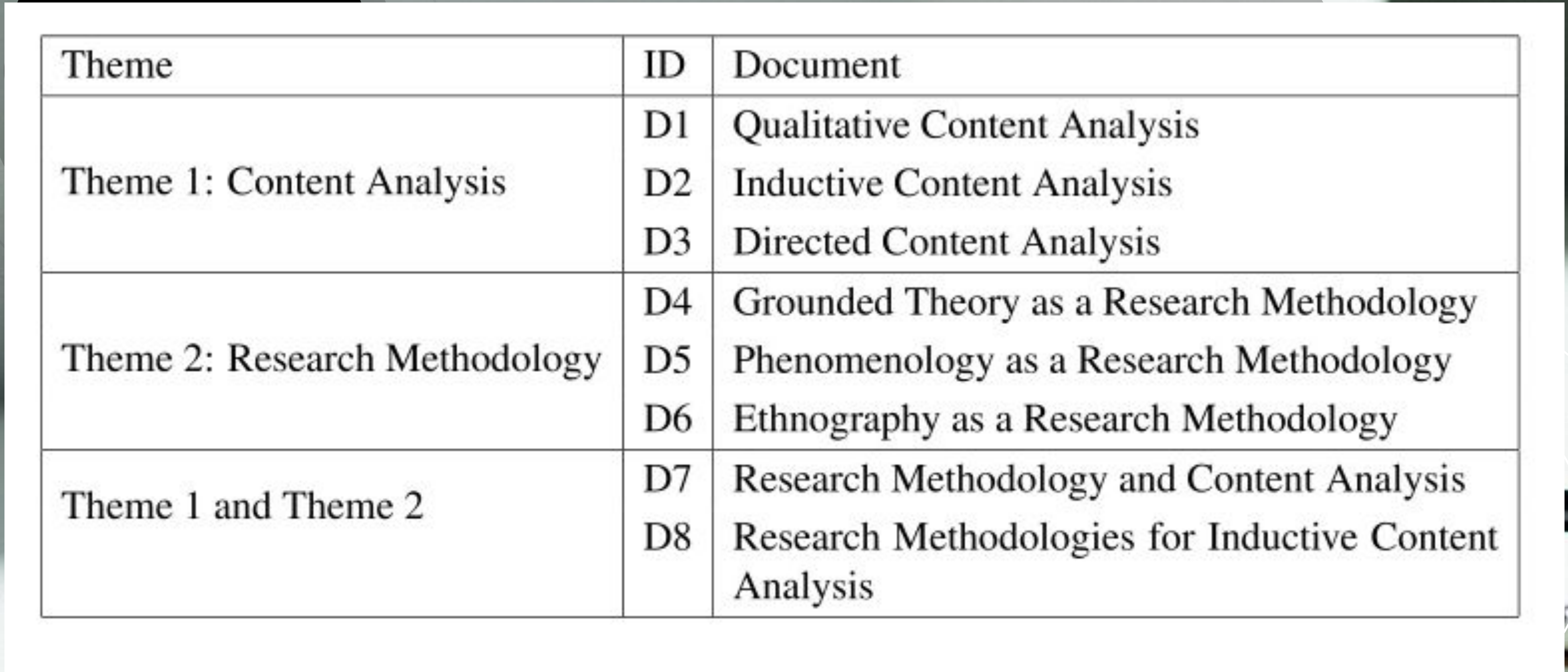
1. Get k multinomials ϕ_k from Dirichlet prior β for each topic k

2. Get D multinomials θ_d from Dirichlet prior α for each document d

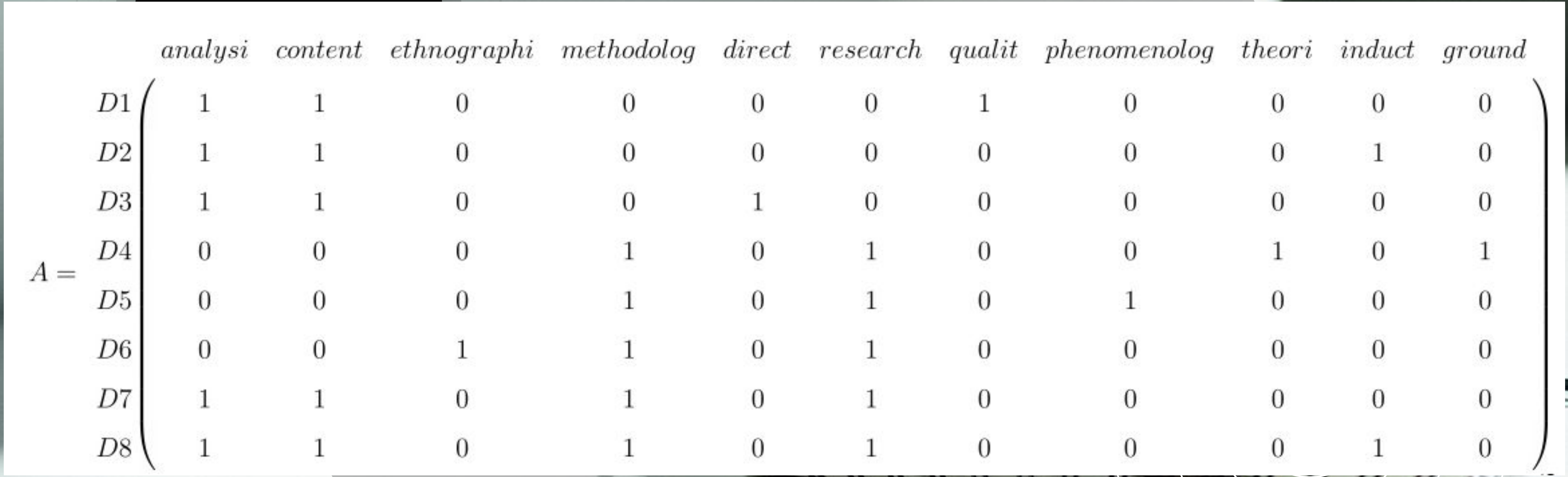
3. For each document d in the corpus and word w_di in the document:

   (a) Get a topic z_i from the multinomial θ_d;$(p(z_i| - \alpha))$

   (b) Get a word w_i from the multinomial ϕ_z;$(p(w_i|z_i, -\beta))$



Theme	ID	Document
Theme 1: Content Analysis	D1	Qualitative Content Analysis
	D2	Inductive Content Analysis
	D3	Directed Content Analysis
Theme 2: Research Methodology	D4	Grounded Theory as a Research Methodology
	D5	Phenomenology as a Research Methodology
	D6	Ethnography as a Research Methodology
Theme 1 and Theme 2	D7	Research Methodology and Content Analysis
	D8	Research Methodologies for Inductive Content Analysis



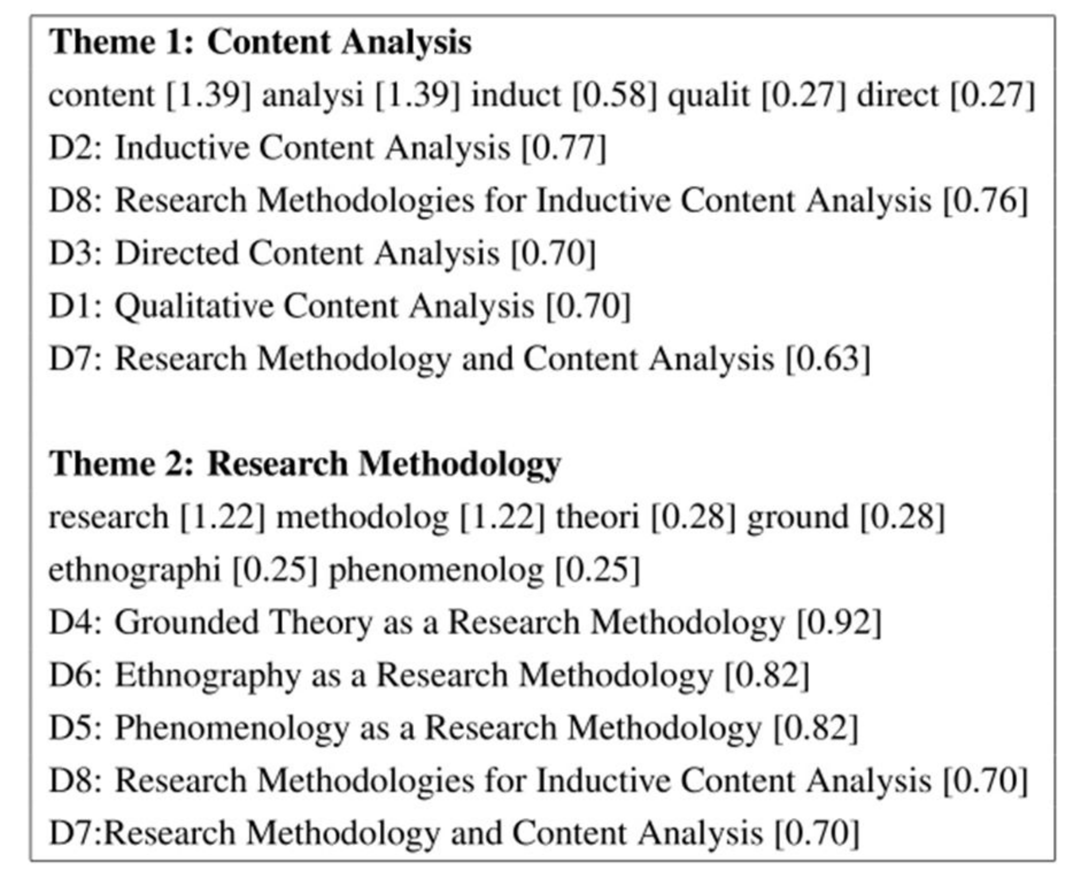
$$A = \begin{array}{c} \\ D1 \\ D2 \\ D3 \\ D4 \\ D5 \\ D6 \\ D7 \\ D8 \end{array} \begin{array}{ccccccccccc} analysi & content & ethnographi & methodolog & direct & research & qualit & phenomenolog & theori & induct & ground \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{array}$$

$$A \approx WH$$

$$W = \begin{array}{c} \\ D1 \\ D2 \\ D3 \\ D4 \\ D5 \\ D6 \\ D7 \\ D8 \end{array} \begin{array}{c} \begin{array}{cc} Theme1 & Theme2 \end{array} \\ \begin{pmatrix} 0.69 & 0.00 \\ 0.76 & 0.00 \\ 0.69 & 0.00 \\ 0.00 & 0.91 \\ 0.00 & 0.81 \\ 0.00 & 0.81 \\ 0.62 & 0.70 \\ 0.76 & 0.70 \end{pmatrix} \end{array}$$

Number of Themes = k; (Must be specified)

$$H = \begin{array}{c} \\ Theme1 \\ Theme2 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} analysi & content & ethnographi & methodolog & direct & research & qualit & phenomenolog & theori & induct & ground \end{array} \\ \begin{pmatrix} 1.39 & 1.39 & 0.00 & 0.07 & 0.27 & 0.07 & 0.27 & 0.00 & 0.00 & 0.58 & 0.00 \\ 0.01 & 0.01 & 0.25 & 1.22 & 0.00 & 1.22 & 0.00 & 0.25 & 0.28 & 0.04 & 0.28 \end{pmatrix} \end{array}$$



Theme 1: Content Analysis

content [1.39] analysi [1.39] induct [0.58] qualit [0.27] direct [0.27]

D2: Inductive Content Analysis [0.77]

D8: Research Methodologies for Inductive Content Analysis [0.76]

D3: Directed Content Analysis [0.70]

D1: Qualitative Content Analysis [0.70]

D7: Research Methodology and Content Analysis [0.63]

Theme 2: Research Methodology

research [1.22] methodolog [1.22] theori [0.28] ground [0.28] ethnographi [0.25] phenomenolog [0.25]

D4: Grounded Theory as a Research Methodology [0.92]

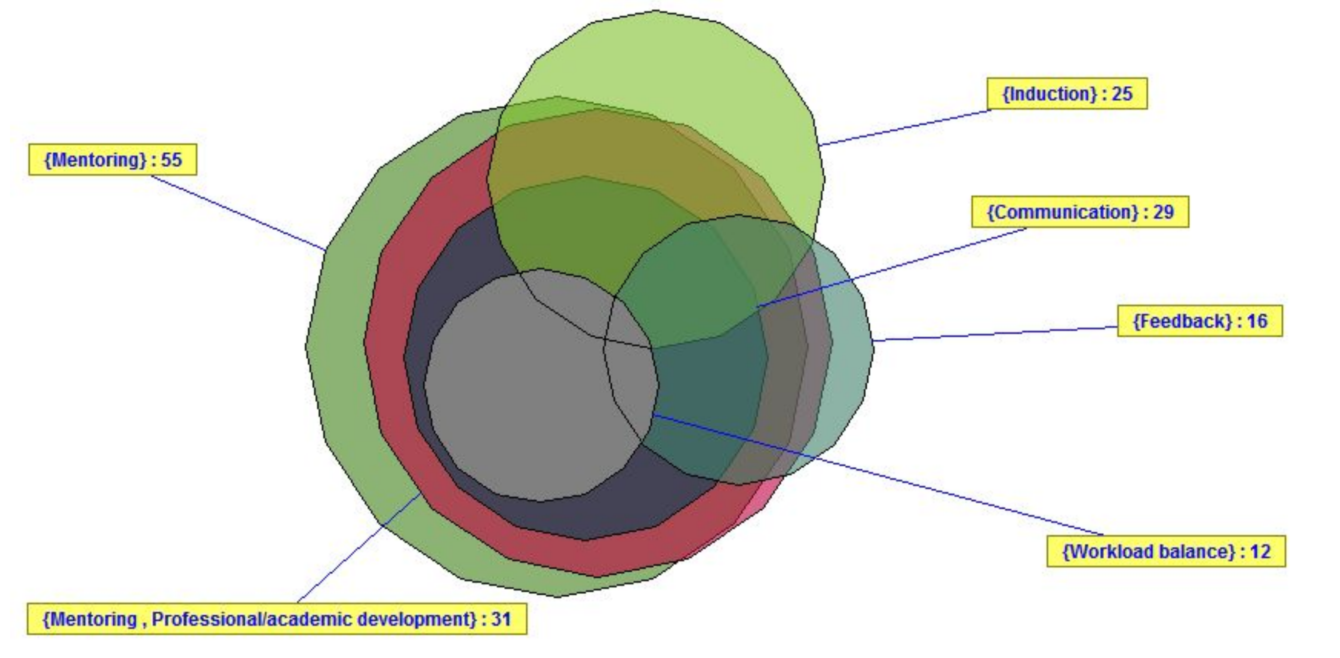
D6: Ethnography as a Research Methodology [0.82]

D5: Phenomenology as a Research Methodology [0.82]

D8: Research Methodologies for Inductive Content Analysis [0.70]

D7:Research Methodology and Content Analysis [0.70]

Why Topic Modeling for Inductive Content Analysis?



Euler Diagram of Topic Overlap by a Manual Coder

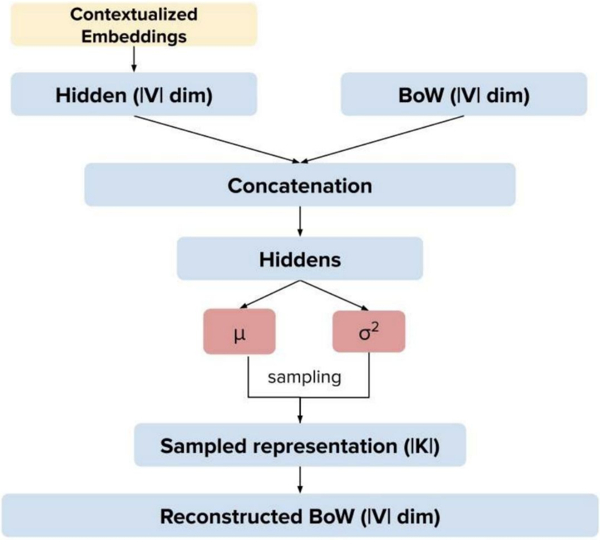
Experiment to compare human coded topics with NMF derived topics.

Humans group documents together with overlap

Bakharia, Aneesha (2014) *Interactive content analysis : evaluating interactive variants of non-negative Matrix Factorisation and Latent Dirichlet Allocation as qualitative content analysis aids.* PhD thesis, Queensland University of Technology.

Contextual Topic Model

- Concatenates a vector obtained from a transformer model such as BERT to the BoW matrix
- Able to handle unseen words
- Improved semantic relatedness between full documents
- Support for multilingual Topic Modeling
- More parameters to tune (no layers and neurons)
- Longer time to run


Contextualized Embeddings
Hidden (|V| dim)
BoW (|V| dim)
Concatenation
Hiddens
μ
σ²
sampling
Sampled representation (|K|)
Reconstructed BoW (|V| dim)

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. ACL. https://aclanthology.org/2021.acl-short.96/

What are the challenges in using topic models as computational aids in qualitative content analysis?

  - Text data pre-processing
  - Parameter selection (eg number of topics)
  - Quality (Topic Coherence & Diversity)
  - Visualisation that aids topic interpretation, evidence gathering and trust

- How can we use Python packages to address these?
- How can we use Jupyter Notebooks to make Topic Modeling accessible to non-coders?

Choosing a Python Library

	Preprocessing	NMF	LDA	CTM	Eval Metrics	Visualisation
Spacy	Y					
Scikit-Learn	Y	Y	Y			
Gensim	Y	Y	Y$^+$		Y	
OCTIS	Y	Y	Y	Y	Y	
LDAVis						Y

OCTIS: Optimizing and Comparing Topic Models is Simple!

Also includes Parameter Tuning via Bayesian Optimization and lots of evaluation measures

https://github.com/MIND-Lab/OCTIS

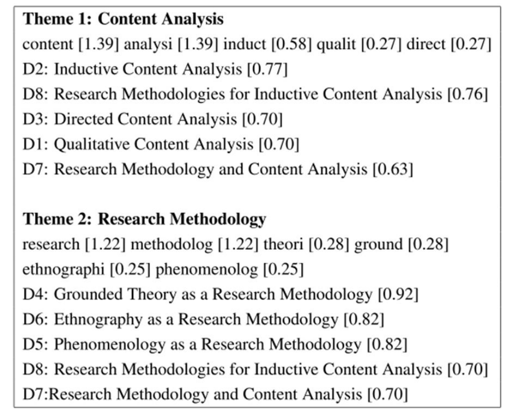
Google Colab Notebook for Topic Modeling（Demo）

- Upload CSV dataset of Twitter Airline Sentiment Tweets
- Preprocess using Spacy + tweet-preprocessor
    - Stop word, hashtag and @reply removal
    - Custom stop words
    - Lemmatization
- Parameter Tune using Coherence and Diversity Metrics from OCTIS (i.e. use to select number of topics)
- Uses Scikit Learn NMF
- View top words and documents to aid with interpretation and evidence gathering

Google Colab Notebook for Topic Modeling（Demo）

- Summary of Workflow
  - Use Coherence and Diversity as a guide to find a good starting number of topics
  - Tweak custom stop words as required after reviewing generated topics
  - Repeat process

Visualising Topics



Theme 1: Content Analysis

content [1.39] analysi [1.39] induct [0.58] qualit [0.27] direct [0.27]

D2: Inductive Content Analysis [0.77]

D8: Research Methodologies for Inductive Content Analysis [0.76]

D3: Directed Content Analysis [0.70]

D1: Qualitative Content Analysis [0.70]

D7: Research Methodology and Content Analysis [0.63]

Theme 2: Research Methodology

research [1.22] methodolog [1.22] theori [0.28] ground [0.28] ethnographi [0.25] phenomenolog [0.25]

D4: Grounded Theory as a Research Methodology [0.92]

D6: Ethnography as a Research Methodology [0.82]

D5: Phenomenology as a Research Methodology [0.82]

D8: Research Methodologies for Inductive Content Analysis [0.70]

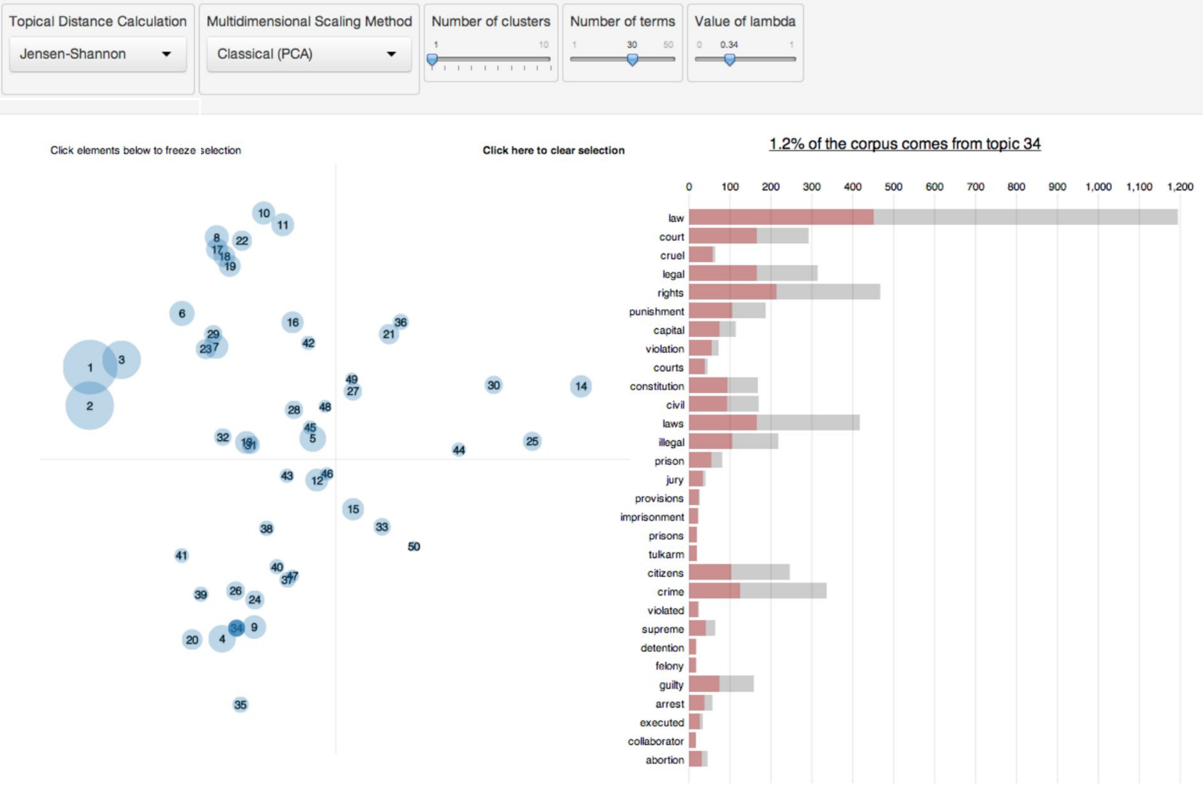
D7:Research Methodology and Content Analysis [0.70]

- Only shown the top words in a topic and asked to evaluate the topic
- But ....
  - Users (researchers and content analysts) need to answer research questions and gather evidence
  - Users need the clustered document grouping information to make decisions
- The context of where the top words occurred (i.e., location within the document) is very important.

Visualising Topics (LDAVis)

Topic Size

Topic Overlap



Topical Distance Calculation: Jensen-Shannon

Multidimensional Scaling Method: Classical (PCA)

Number of clusters: 1

Number of terms: 30

Value of lambda: 0.34

Click elements below to freeze selection

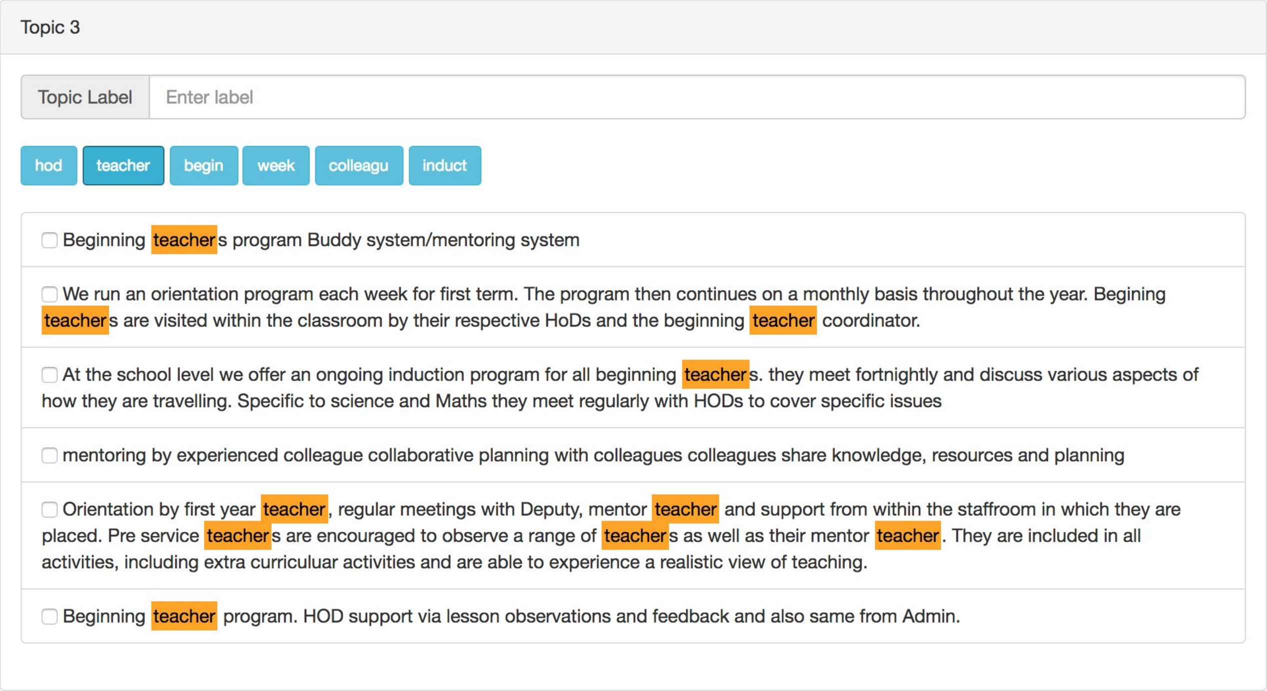
Click here to clear selection

1.2% of the corpus comes from topic 34

Term Relevance In Selected Topic And Corpus

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Interpreting Topics



□ Beginning teachers program Buddy system/mentoring system

□ We run an orientation program each week for first term. The program then continues on a monthly basis throughout the year. Begining teachers are visited within the classroom by their respective HoDs and the beginning teacher coordinator.

□ At the school level we offer an ongoing induction program for all beginning teachers. they meet fortnightly and discuss various aspects of how they are travelling. Specific to science and Maths they meet regularly with HODs to cover specific issues

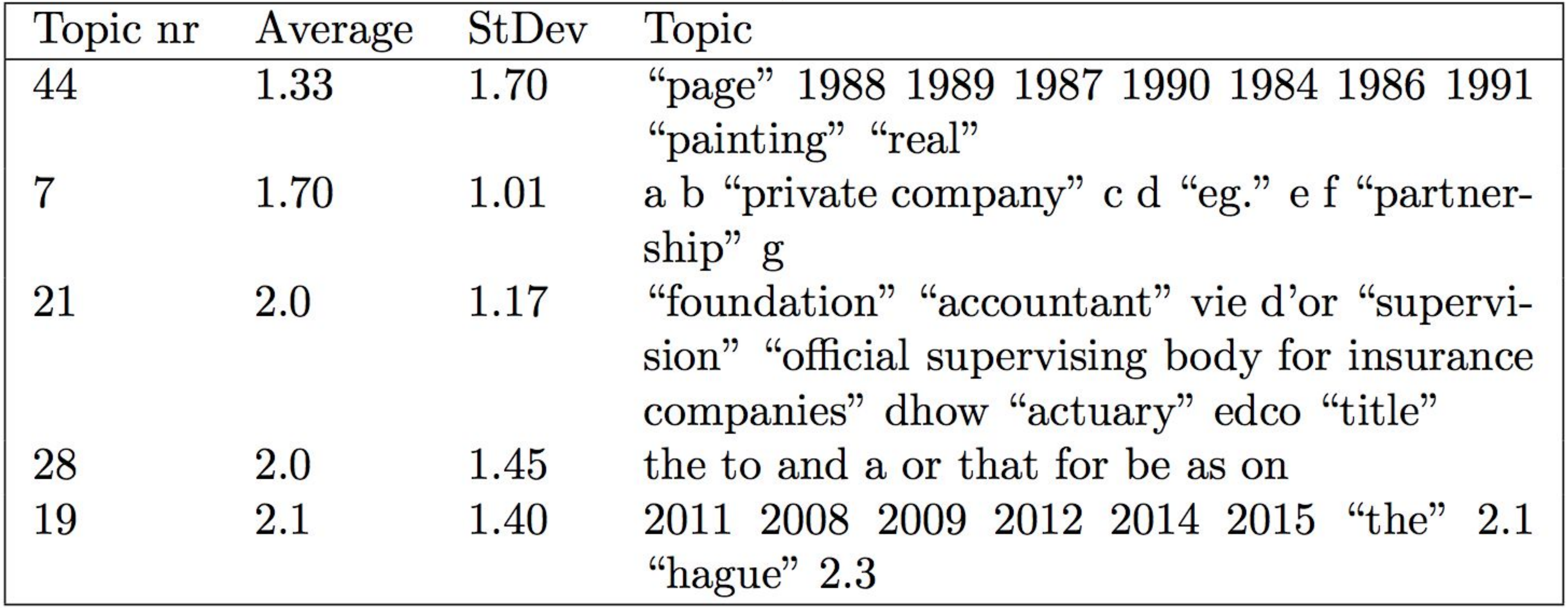
□ mentoring by experienced colleague collaborative planning with colleagues colleagues share knowledge, resources and planning

□ Orientation by first year teacher, regular meetings with Deputy, mentor teacher and support from within the staffroom in which they are placed. Pre service teachers are encouraged to observe a range of teachers as well as their mentor teacher. They are included in all activities, including extra curriculuar activities and are able to experience a realistic view of teaching.

□ Beginning teacher program. HOD support via lesson observations and feedback and also same from Admin.

- Content Analysts need to see the occurrence of the top words within the top documents

Interpreting Topics – Beware Low Quality Topics



Topic nr	Average	StDev	Topic
44	1.33	1.70	“page” 1988 1989 1987 1990 1984 1986 1991 “painting” “real”
7	1.70	1.01	a b “private company” c d “eg.” e f “partnership” g
21	2.0	1.17	“foundation” “accountant” vie d’or “supervision” “official supervising body for insurance companies” dhow “actuary” edco “title”
28	2.0	1.45	the to and a or that for be as on
19	2.1	1.40	2011 2008 2009 2012 2014 2015 “the” 2.1 “hague” 2.3

Low quality topics from:
Finding the Topics of Case Law: Latent Dirichlet Allocation on Supreme Court Decisions
http://theses.ubn.ru.nl/bitstream/handle/123456789/5218/Remmits%2C%20Y._BSc_Thesis_2017.pdf?sequence=1

What about R users?

Tutorial from Language Technology and Data Analysis Lab (LADAL), School of Languages and Culture, UQ

https://slcladal.github.io/topicmodels.html

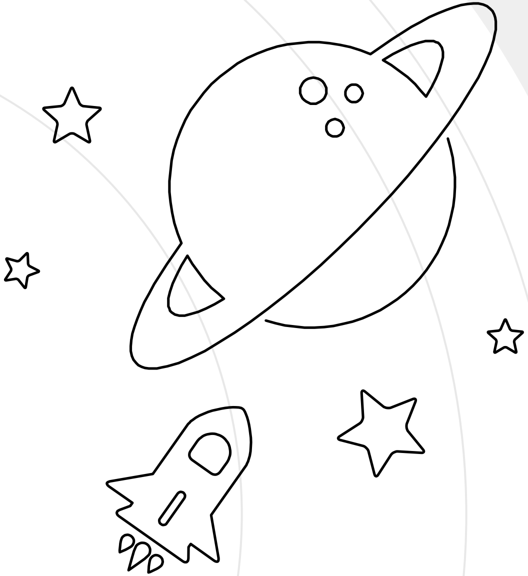
Readings

- Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Schmid-Petri, H. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93-118.

- Bakharia, A., Bruza, P., Watters, J., Narayan, B., & Sitbon, L. (2016). Interactive topic modeling for aiding qualitative content analysis. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*(pp. 213-222). ACM.

- Bakharia, A. (2019, October). On the equivalence of inductive content analysis and topic modeling. In *International Conference on Quantitative Ethnography* (pp. 291-298). Springer, Cham.

Questions



Slides & Code:

https://github.com/aneesha/Resbaz2021_QualTopicModeling

Dr Aneesha Bakharia
Institute of Teaching and Learning Innovation
The University of Queensland
Email: aneesha.bakharia@gmail.com
Twitter: @aneesha

Presentation design

This presentation uses the following typographies:

- Titles: Poppins Bold
- Body copy: Poppins Light

You can download the fonts at:

https://www.fontsquirrel.com/fonts/poppins

You don’t need to keep this slide in your presentation. It’s only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®