Text Mining Tools for Narrative Analysis and Other Mixed-Method Research

Mary Kate Koch, M.A., A.B.D

SSEA Pre-conference Workshop

November 4th, 2021

Our general schedule

I. Topic modeling definitions and practical decisions (~30 minutes)

II. Code demonstration in R with *Big Mouth* data (~30 minutes)

III. Code walkthrough together with novel data (~90 minutes)

IV. Self-guided time (~30 minutes)

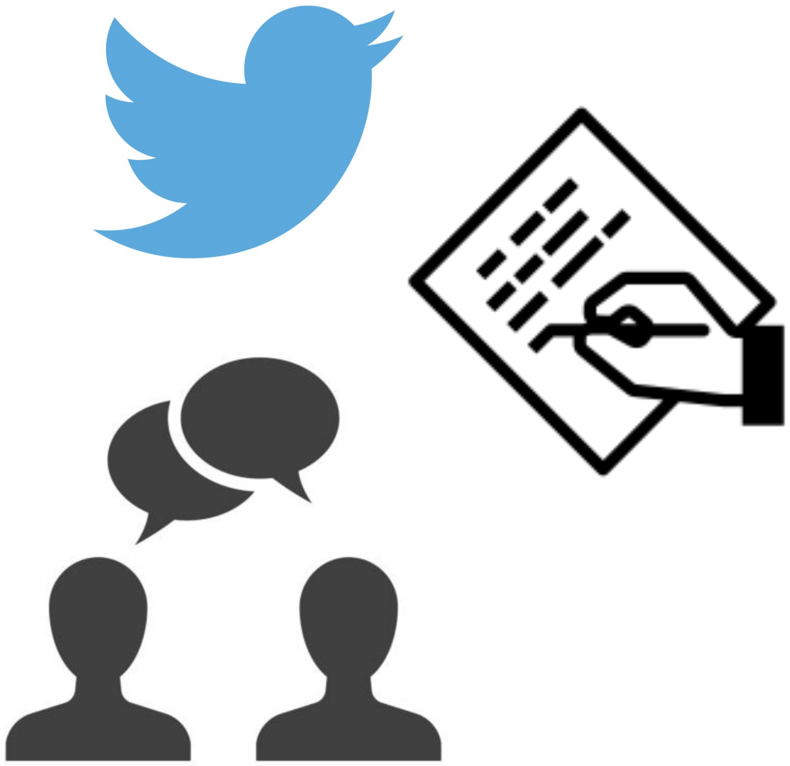
Some things to keep in mind

- Present majority of text mining articles are not written to be accessible or helpful to you

- Information science articles favor optimization whereas we may be more content and theory oriented

- Translating text mining from information science to the social sciences is ongoing work
  - Cross-collaboration, mixed methods, and establishing best practices for our field are important ongoing work

Opportunities of working with text



- Respondents can answer in their own voice
- Data can go beyond categories pre-established by researchers
- Collection and transcription easier than ever now

Challenges of working with text

- Often resource-intensive to analyze in a systematic way

- Calibrating inter-rater reliability across human raters can be tricky

- Structure of coding manuals may be restrictive regarding what text is coded and how broadly codes are labeled

- May not be intuitive how to integrate with quantitative methods

Atkins et al., 2012

What is NLP?

- NLP = Natural language processing
- Methods encompass:
  - Sentiment analysis
  - Part of speech (POS) tagging
  - Word embeddings
  - Topic modeling
  - Many more
- Today's focus is on topic modeling

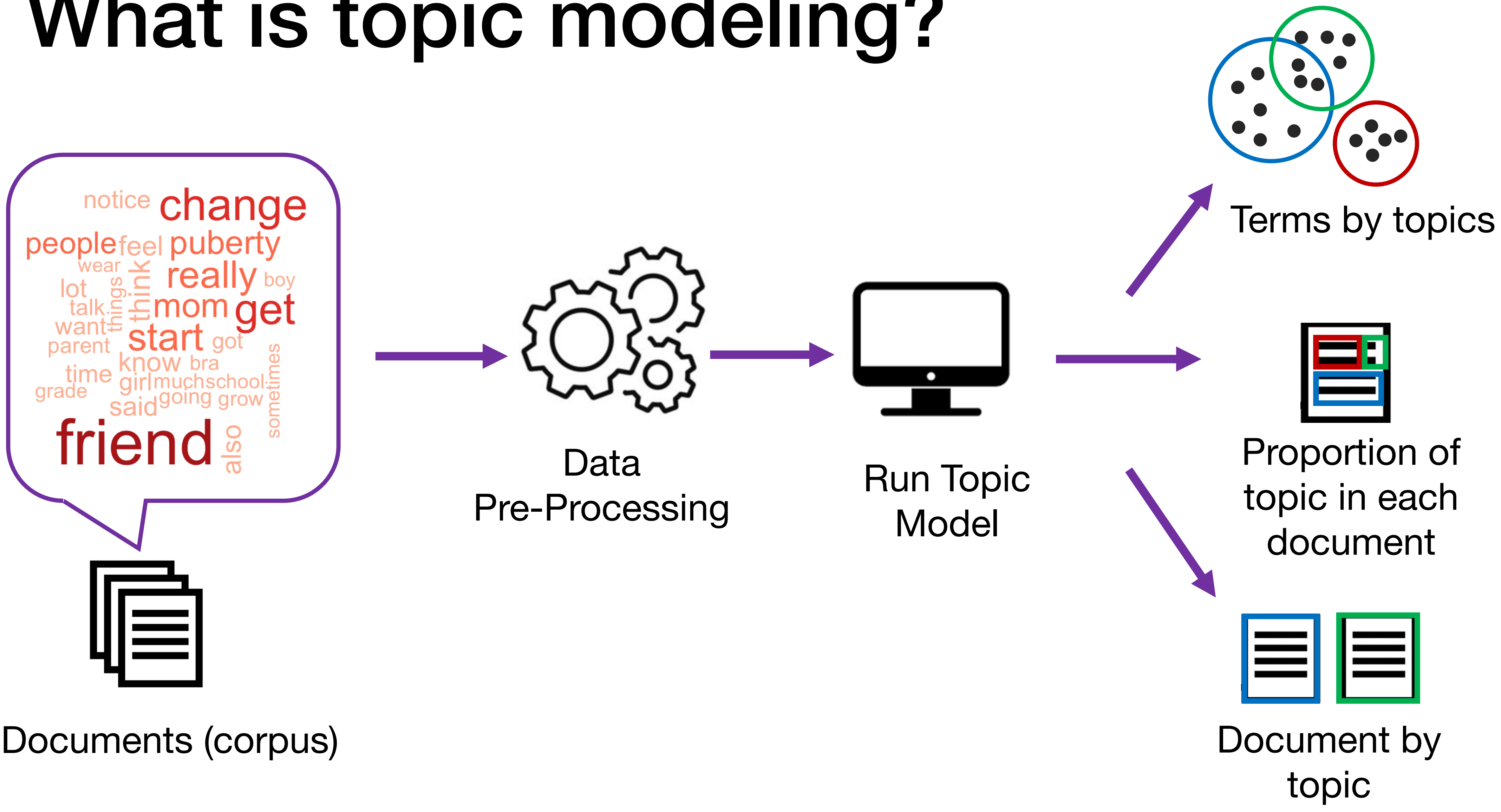
What is topic modeling?

- Data-driven method for analyzing content of text

- A dimensionality-reduction technique like principal component analysis (PCA) that transforms a large set of variables into a smaller set

- Transforms a collection of text (i.e., a corpus) into a smaller number of word clusters (i.e., topics) that typically provide an interpretable summary of the broader corpus

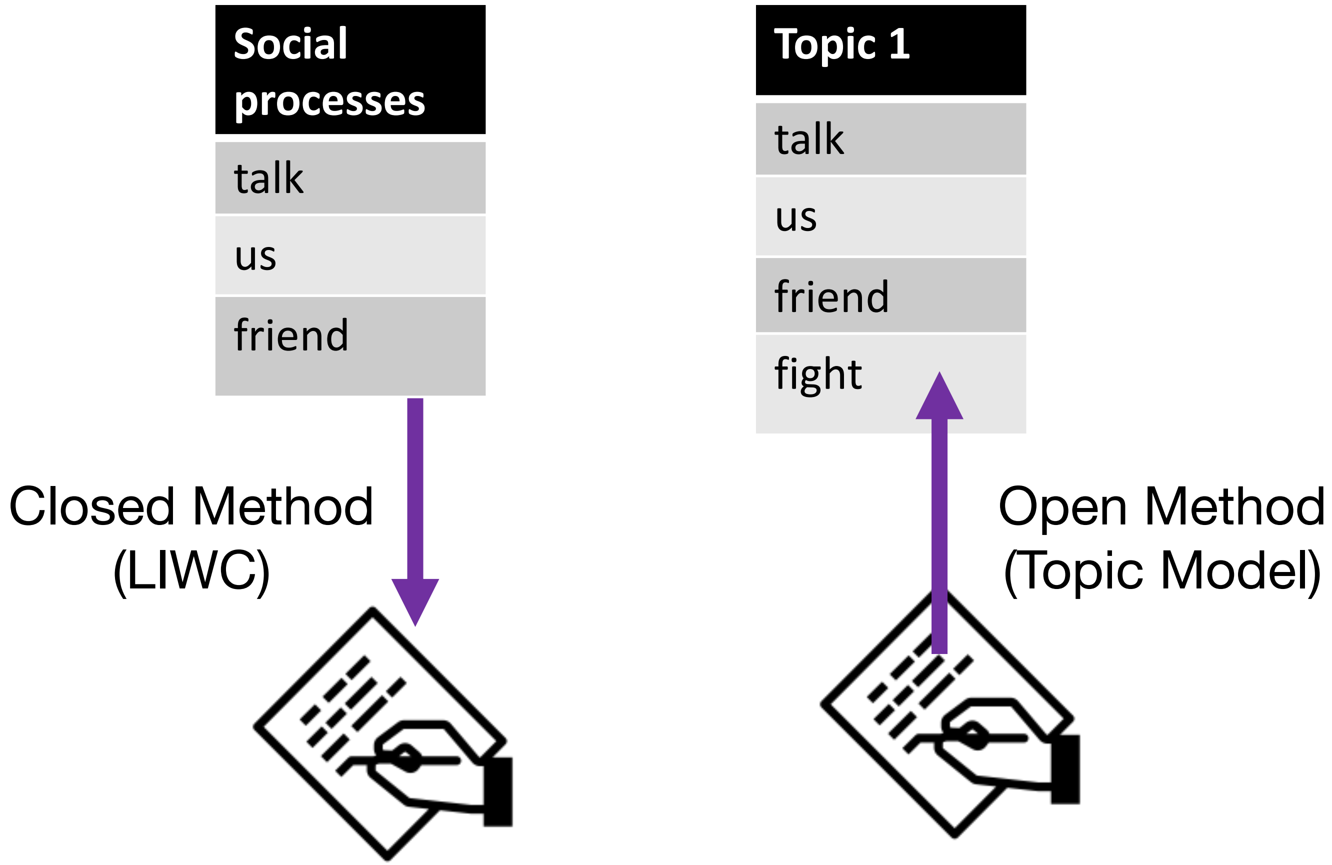
What is topic modeling?



How is this different from LIWC?

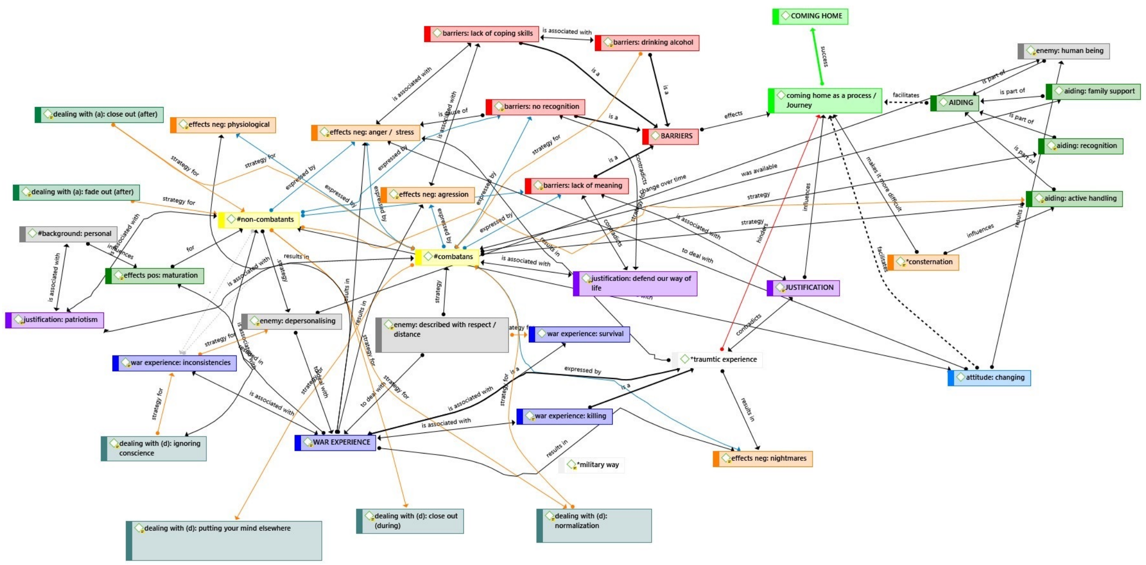
- Closed-vocabulary analysis = *a priori* assumptions about what words mean and which words to evaluate

- Open-vocabulary analysis = derives associations between words from the data itself; **no** ***a priori*** assumptions



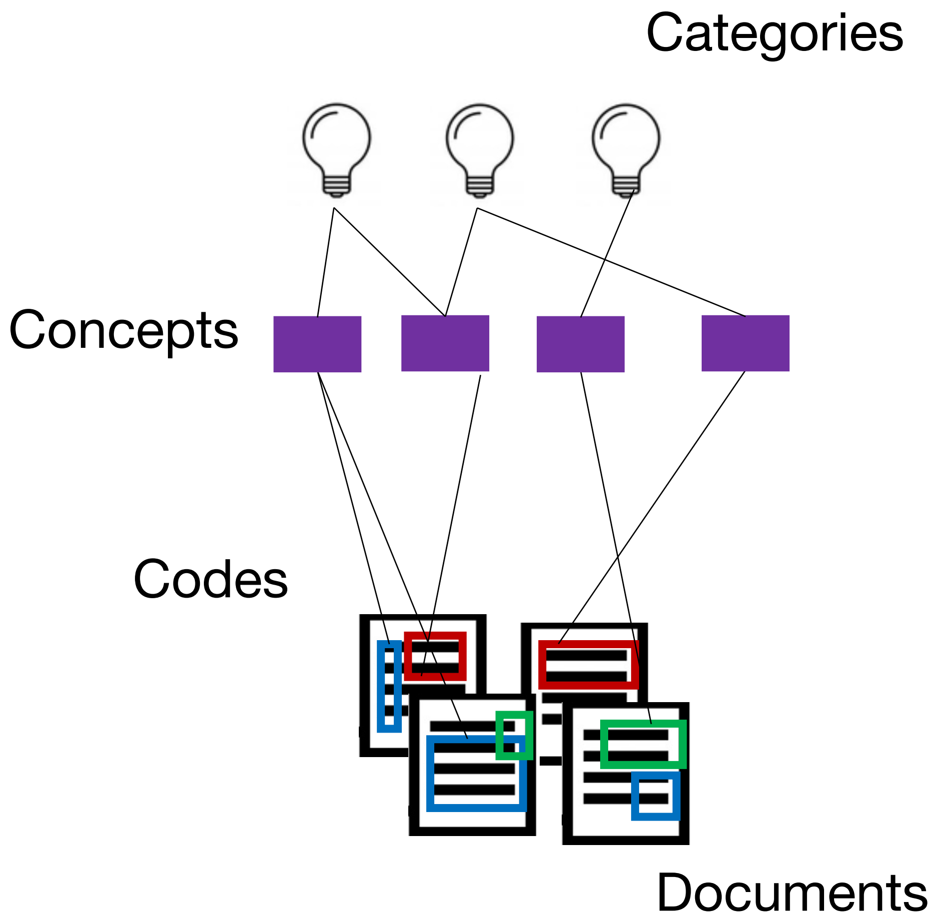
How is this different from ATLAS.ti?

- Not a specific software – can be run on R or Python
- Doesn't let you tag codes or visualize tags for you
- ATLAS.ti may be better to use when you want to easily code and cluster text



Convergence with grounded theory

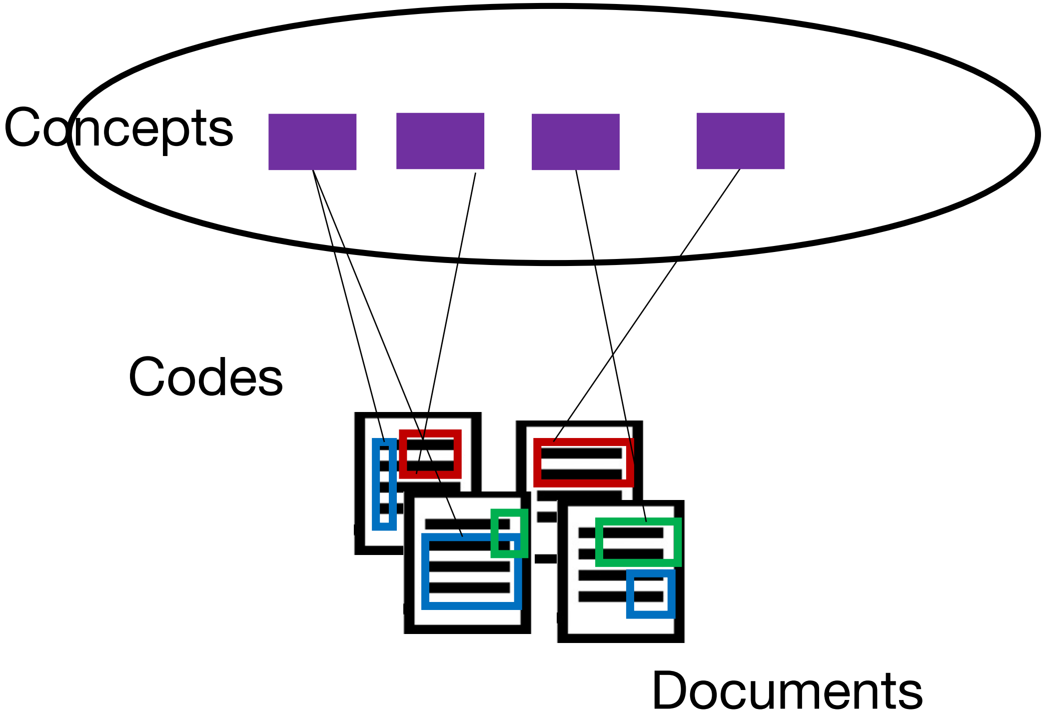
- Start with documents and read them over and over again
- Eventually, reach higher level insight about the text in collected documents



(Baumer et al., 2017)

Convergence with grounded theory

- Topic modeling sets an algorithm to read documents over and over again
- Iteratively build connections between aspects of these documents and the topics



(Baumer et al., 2017)

Convergence with grounded theory

	Triggers for returning	Communicative necessity	Morality	Renegotiated	Social reconnection	Friends' reactions
1. Positive response from friends						Friends had positive reactions
2. Necessary for communication (distance, tragedy, etc.)	Need to communicate as a type of trigger	Reasons for communicative necessity				
3. No reaction						Friends showed no reaction
4. Brief, focused, guilty return	Utilitarian (e.g., info seeking)		Qualified guilt			
5. Negative emotions (guilt, disappointment, addiction)			Guilt, let myself down	Addiction implies limited control		
7. Stories, obliged return, immediate reaction	Major life events as triggers				Welcomed back	Mixed reactions
8. Positive emotions, changed use				Increased self control	Positive about reconnecting	

(Baumer et al., 2017)

Decisions, Decisions: Pre-processing and Topic Number Selection

Pre-processing decisions

- Segmentation
- Stopwords
- Normalizing
- Stemming
- Tokenization
- K topic selection

Segmentation

- What text to include in the corpus and how to slice it up

- Topic models struggle to process text when there is too much variation in the document

- Focus group transcript vs. one paragraph response to open-ended survey questions --> different format = different needs

Stopwords

- Common words can crowd output and hinder clarity

- Two approaches to removing words:
    1. use the stopwords dictionary of the package you are using
    2. compile a list of stopwords specific to your corpus

- Removal of the most common words, determiners, conjunctions, and prepositions can improve model fit and quality
    - **But** further removal has no major effect on topic coherence or classification accuracy

(Schofield et al., 2017)

Normalizing

- Common to make all words lowercase and remove punctuation, numbers, and special characters

- Intended to reduce vocabulary size of corpus to improve the representation quality of topic terms

- However, use knowledge of data and research questions to inform these choices

Stemming

- Combines near-duplicate terms
  - e.g., *growing* and *grows* are combined into a single term: *grow*–

- Goal is to reduce vocabulary size and improve topic coherence but there are drawbacks:
  - return terms can be difficult to interpret (e.g., *stai*- is the stem of *stay*)
  - may conflate terms that have different semantic meanings (e.g., *apple* as a stem of both *apples* and *Apple)*

- Post-analysis stemming is usually better option because it retains the semantic context of words

(Schofield & Mimno, 2016)

Tokenization

- Individual words (i.e., unigrams) as the unit of analysis by default

- May be instances in which more than one word makes up a semantic unit of interest
  - *Middle* and *school* versus *middle school*

- Can tokenize individual terms with prior knowledge or through iterative process

Pre-processing sensitivity

- Each of these decisions can have significant impact on model output
  - Topic models are not especially stable
- Report whatever choices you make in the method section
  - Ideally, make these decisions before playing with your model
- Run independent initializations of model to get a sense of topic stability
- Denny and Spirling (2018) provide additional recommendations

Topic selection (K = ?)

- No “true” or “correct” number of topics in any given corpus
- It’s like asking how many slices of cake is the right amount
  - Fewer topics will translate to broader topic bins and more topics will translate to more fine-grained topics
- Can use metrics like adaptive-density to help pinpoint which topic numbers to try (Cao et al., 2009)
  - Will take some iterative exploration to see which model fits your analysis best

(Nguyen et al., 2019)

Interpreting output

- Topics are generally interpreted by their top-N terms, ranked based on the marginal probability $p(w_i \mid t_j)$ in that topic

2 Necessary for communication

Top 25 words: *with, friends, people, contact, because, way, only, other, family, some, keep, touch, needed, talk, could, who, need, really, also, phone, all, communicate, felt, miss, certain*

- Interpretable topics should be:
  1. Cohesive (i.e., high-probability words for the topic tend to co-occur within documents)
  2. Exclusive (i.e., top words for that topic are unlikely to appear within top words of other topics)

Interpreting output

- Posterior topic distribution – to what proportion does a document contain each topic
  - Can be used to quantify if certain groups write about some topics more
  - Or if writing about some topics more is associated with some outcome

- “Best fit” topic for each document

Reliability and validity

- Relying on model selection statistics will maximize model fit but not necessarily benefit substantive interpretation

- Validity measures may include:
  - pairing model output with human close reading of the text
  - comparing the model output to external measures that correspond

- Reliability measures may include independent initialization
  - Topics that don’t repeat (with small variation) may not be reliable

When to use topic models

- Topic models are great for insight-driven analysis
  - Looking for a “high-level” insight
  - Interested in **what** is being said

- If you are not interested in the topics themselves, there are better tools to use
  - For example, if you are interested in the affective tone (**how** things are being said), semantic analysis is a good tool choice