Foundations in Text Analysis

Alina Arseniev-Koehler

Outline

1. Text Analysis Overview
2. Basic Text Cleaning and Wrangling
3. Dictionary Methods
   - Lexicon, Sentiment Analysis
4. A Teaser on Other Methods
   - Text Network, Topic Models, Word Embeddings, Supervised Learning
5. Intro to group exercise

1. Text Analysis Overview

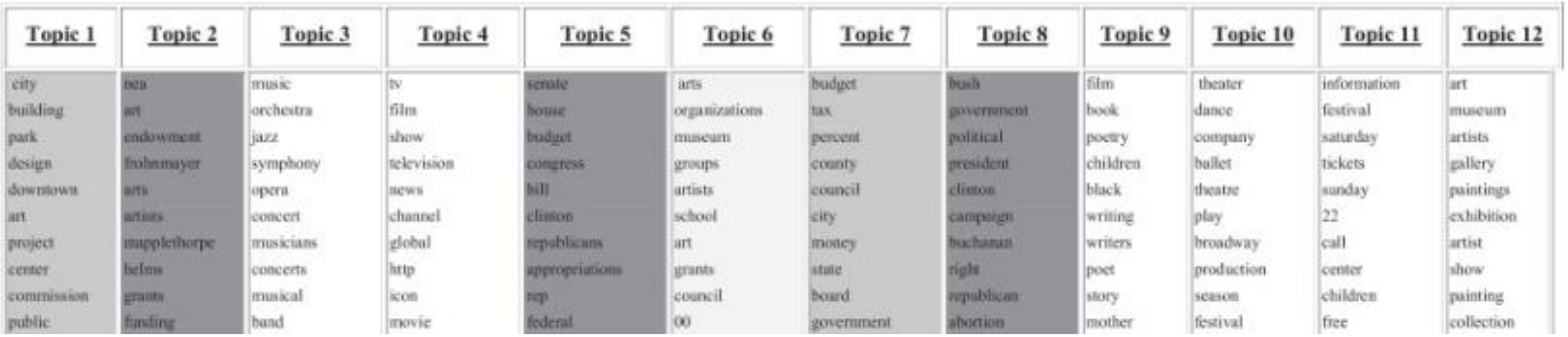
Why Automated Text Analysis?

- Why text?
    - Rich in meaning
    - Reflects + influences social life (cultural)
- *Goal:* Learning from text data in ways that are scalable *or* provide a different lens than qualitative analysis

- Limitation: quantifying meaning?

Sample Research 1

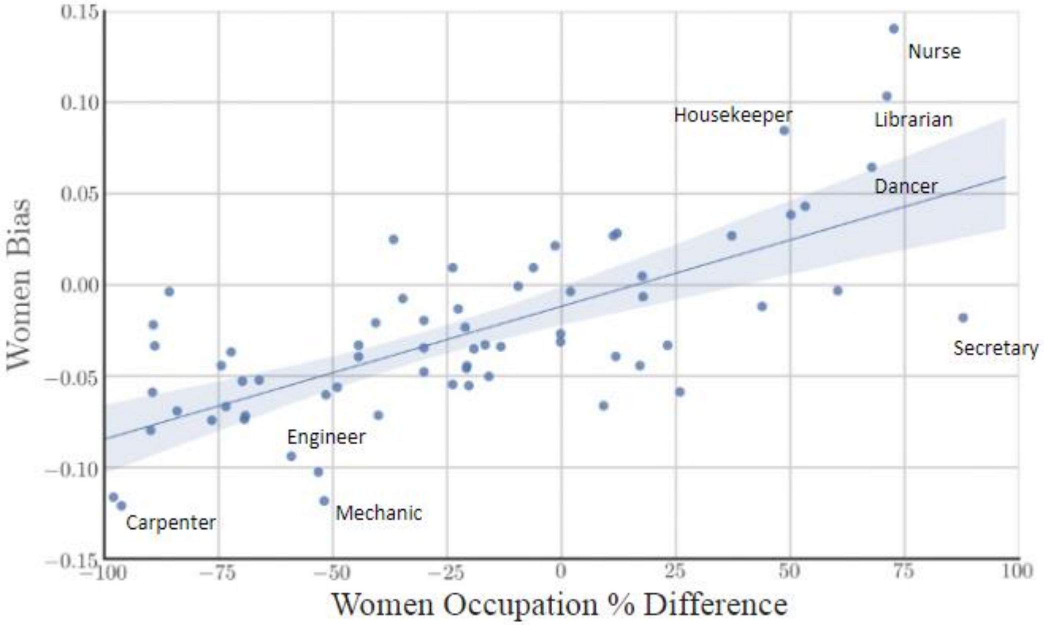
- Identifies how newspapers frame government assistance to artists and arts organizations

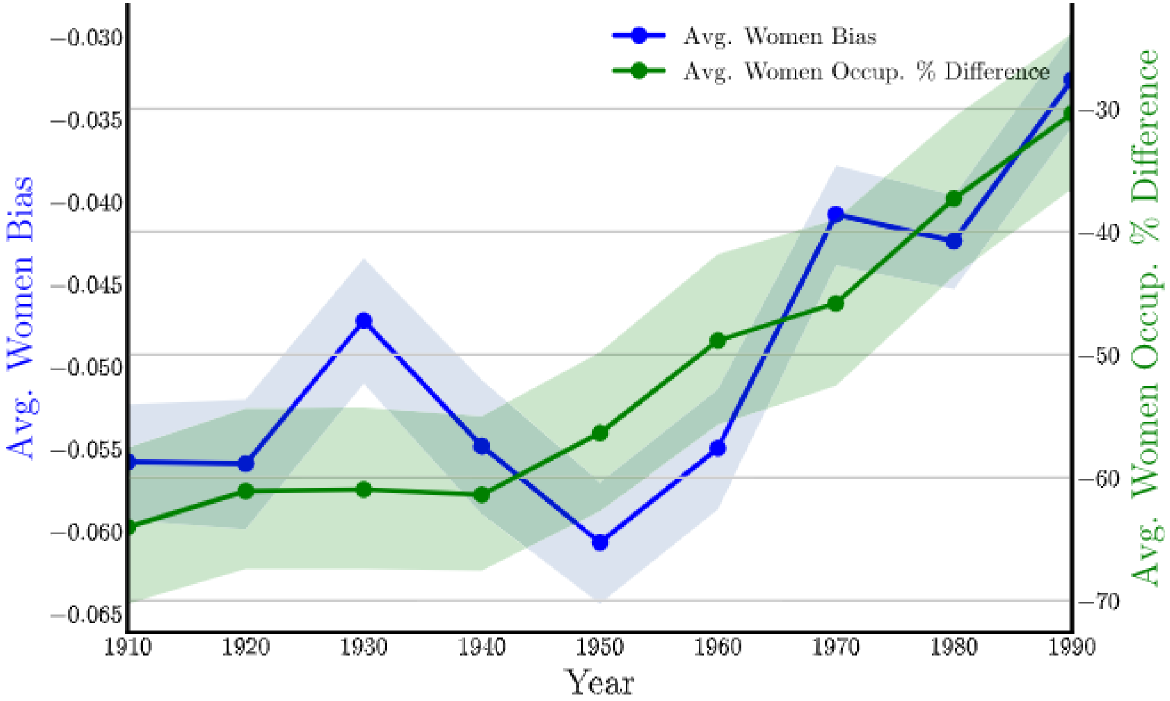


Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
city	nea	music	tv	senate	arts	budget	bush	film	theater	information	art
building	art	orchestra	film	house	organizations	tax	government	book	dance	festival	museum
park	endowment	jazz	show	budget	museum	percent	political	poetry	company	saturday	artists
design	frohnmayer	symphony	television	congress	groups	county	president	children	ballet	tickets	gallery
downtown	arts	opera	news	bill	artists	council	clinton	black	theatre	sunday	paintings
art	artists	concert	channel	clinton	school	city	campaign	writing	play	22	exhibition
project	mapplethorpe	musicians	global	republicans	art	money	buchanan	writers	broadway	call	artist
center	helms	concerts	http	appropriations	grants	state	right	poet	production	center	show
commission	grants	musical	icon	rep	council	board	republican	story	season	children	painting
public	funding	band	movie	federal	00	government	abortion	mother	festival	free	collection

DiMaggio, Paul, Manish Nag, and David Blei. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding." *Poetics* 41.6 (2013): 570-606.

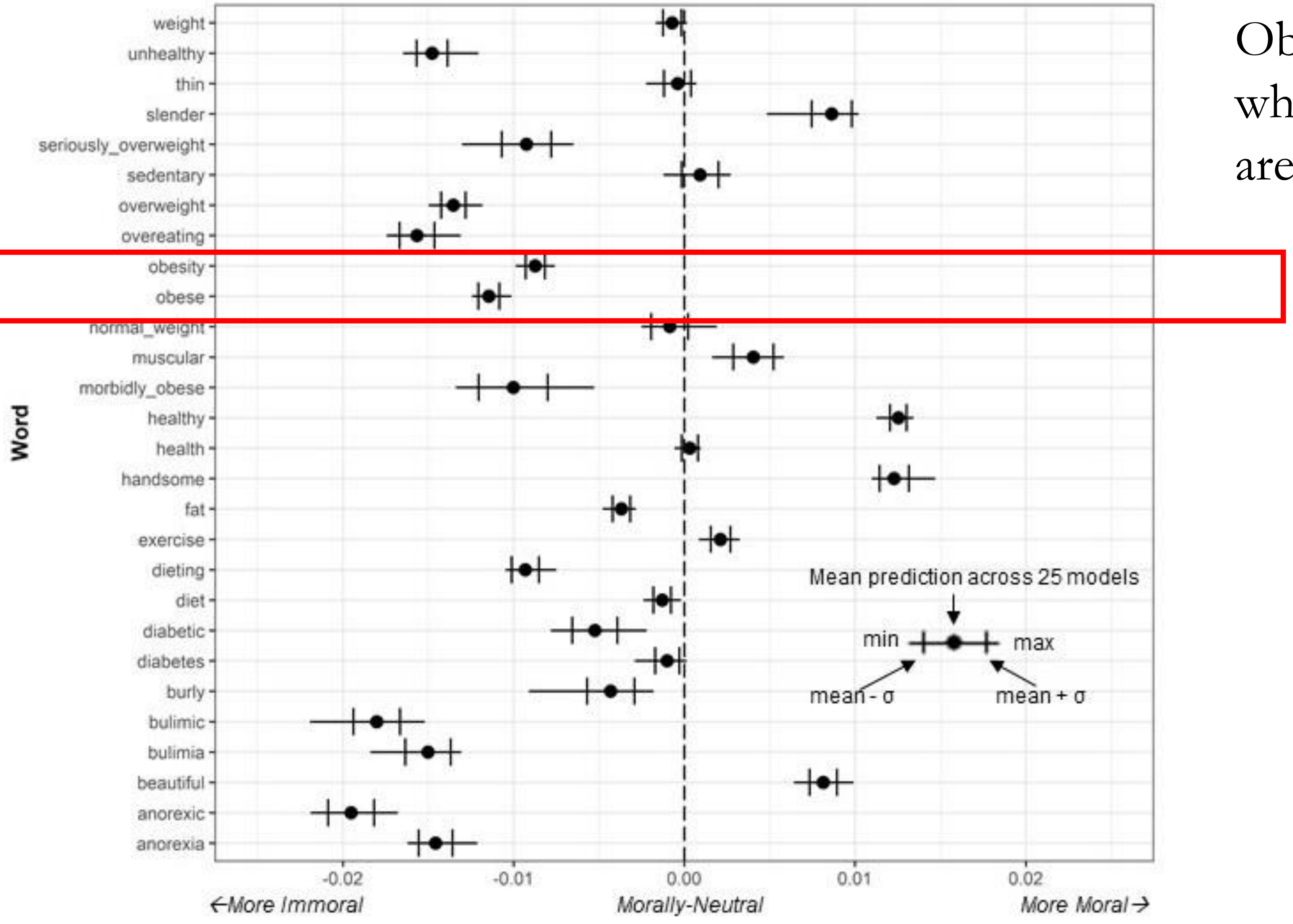
Sample Research 2





Garg, Nikhil, et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115.16 (2018): E3635-E3644.

Sample Research 3



Morality in *New York Times*

Obesity is immoral… while health and beauty are moral

Arseniev-Koehler and Foster

What do we do with text data?

- Study culture
  - Themes, topics, content, style
  - Across time, contexts
  - Content + form, meaning processes

- Extract variables from text (e.g., certainty, sentiment, gender of the writer)
  - Then, use these variables to *explain* another outcome

Data

- Oral histories, interviews
- Social media
- Scientific articles
- Fiction
- Movie/TV scripts
- News, magazines
- Narratives
- Medical notes
- Blogs, Wikipedia
- Reviews (professors, products, movies)
- Other ideas?

- Written vs spoken
- Asynchronous vs synchronous

Questions

2. Basic Text Cleaning & Wrangling

Before analysis begins…

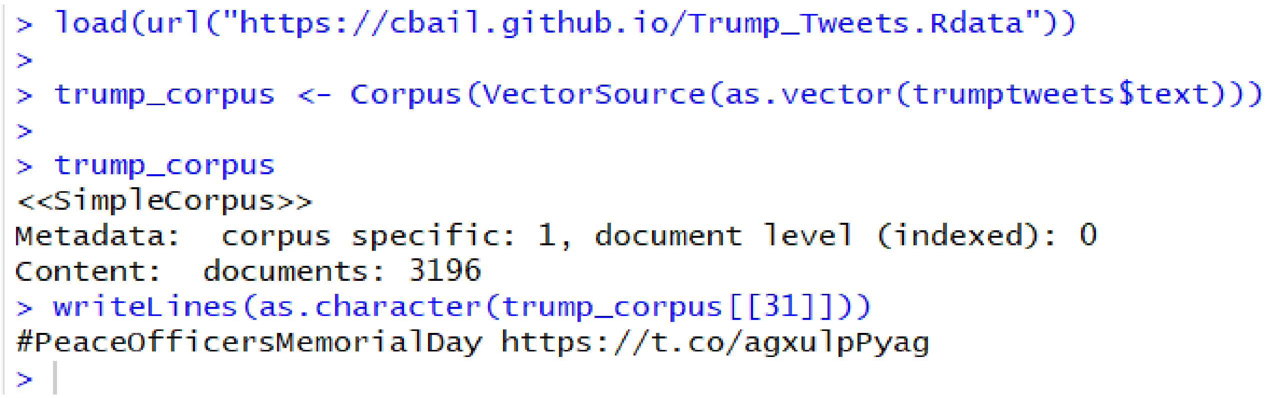
- Loading in your data

- Clean the text to fit *your* needs

Raw: “The quick brown fox jumped over the lazy dog. The cat saat on the hat.”

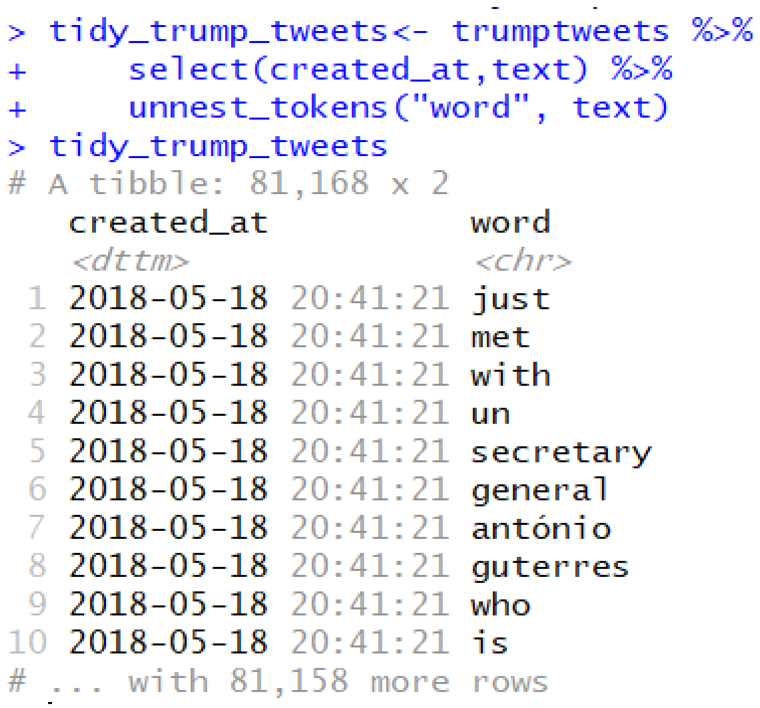
Cleaned: [[“the”, “quick”, “brown”, “fox”, “jumped”, “over”, “the”, “lazy”, “dog”], [“the”, cat”, “sat”, “on”, “the”, “hat”]]

Corpus Object



```
> load(url("https://cbail.github.io/Trump_Tweets.Rdata"))
>
> trump_corpus <- Corpus(VectorSource(as.vector(trumptweets$text)))
>
> trump_corpus
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 3196
> writeLines(as.character(trump_corpus[[31]]))
#PeaceOfficersMemorialDay https://t.co/agxulpPyag
> |
```

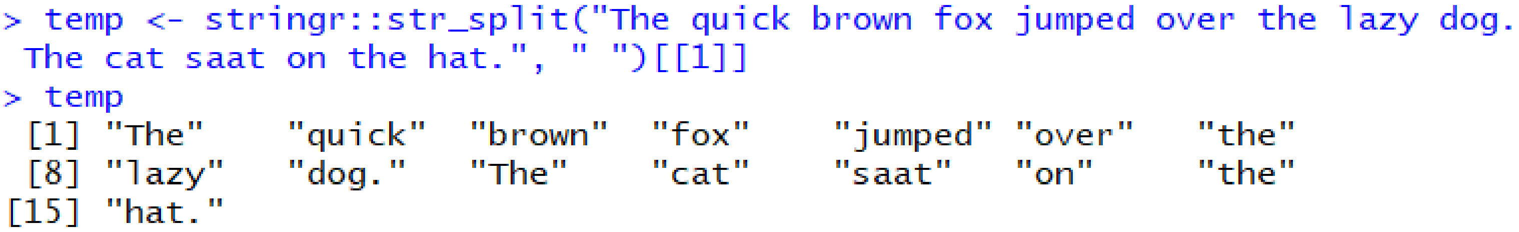
Tidy-Text



```
> tidy_trump_tweets<- trumptweets %>%
+     select(created_at,text) %>%
+     unnest_tokens("word", text)
> tidy_trump_tweets
# A tibble: 81,168 x 2
   created_at          word
   <dttm>              <chr>
 1 2018-05-18 20:41:21 just
 2 2018-05-18 20:41:21 met
 3 2018-05-18 20:41:21 with
 4 2018-05-18 20:41:21 un
 5 2018-05-18 20:41:21 secretary
 6 2018-05-18 20:41:21 general
 7 2018-05-18 20:41:21 antónio
 8 2018-05-18 20:41:21 guterres
 9 2018-05-18 20:41:21 who
10 2018-05-18 20:41:21 is
# ... with 81,158 more rows
```

**This slide and its images are adapted from Chris Bail's SICSS-Duke Workshop R tutorial*

Cleaning Text Data



```
> temp <- stringr::str_split("The quick brown fox jumped over the lazy dog.
 The cat saat on the hat.", " ")[[1]]
> temp
 [1] "The"    "quick"  "brown"  "fox"    "jumped" "over"   "the"
 [8] "lazy"   "dog."   "The"    "cat"    "saat"   "on"     "the"
[15] "hat."
```

- Tokenize, lowercase, correct spelling, remove punctuation & numbers
- 1-grams vs bigrams: “New” “York” v.s. “New_York”

Cleaning Text Data

- Stem

  Cleaned: [[“the”, “quick”, “brown”, “fox”, “jump”, “over”, “the”, “lazy”, “dog”], [“the”, cat”, “sat”, “on”, “the”, “hat”]]

- Stem, and Remove Stop-Words

  Cleaned: [[“quick”, “brown”, “fox”, “jump”, “over”, “lazy”, “dog”], [“cat”, “sat”, “on”, “hat”]]

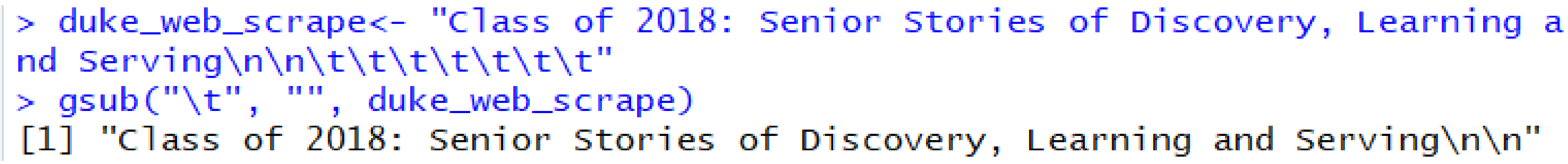
- Whitespaces

  Tokenized & Stemmed: [[“quick ”, “brown”, “ fox”, “jump”, “over”, “lazy   ”, “dog”], [“cat”, “sat”, “on”, “hat”]]

  Cleaned: [[“quick ”, “brown”, “fox”, “jump”, “over”, “lazy”, “dog”], [“cat”, “sat”, “on”, “hat”]]

GREP / Regular Expressions

- GREP is a tool that helps you search for occurrences of a specific pattern. Some examples…
  - Replace all @handlenames in tweets with a standard token
  - Find all instances where a specific illness is mentioned, and then disambiguate (e.g., "alcoholic", "overweight," "stroke")
  - Standardize the ways in which an illness (or other entity) is referenced



```
> duke_web_scrape<- "Class of 2018: Senior Stories of Discovery, Learning a
nd Serving\n\n\t\t\t\t\t\t\t"
> gsub("\t", "", duke_web_scrape)
[1] "Class of 2018: Senior Stories of Discovery, Learning and Serving\n\n"
```

**This slide image is adapted from Chris Bail's SICSS-Duke Workshop R tutorial*

Corpus level Co-Occurrence Matrix

"the increasing prevalence of obesity is like a hundred car freight train going downhill with no brakes"

"a national epidemic of childhood obesity"

"obesity is on the increase"

Vocabulary size: 24

Tokens: 28

	A	Brakes	Childhood	Downhill	Epidemic	…
A	0	1	1	1	1	…
Brakes	1	0	0	1	0	…
Childhood	1	0	0	0	1	…
Downhill	1	1	0	0	0	…
Epidemic	1	0	1	0	0	…
…	…	…	…	…	…	…

- each word represented as a (sparse) vector
- measure similarity

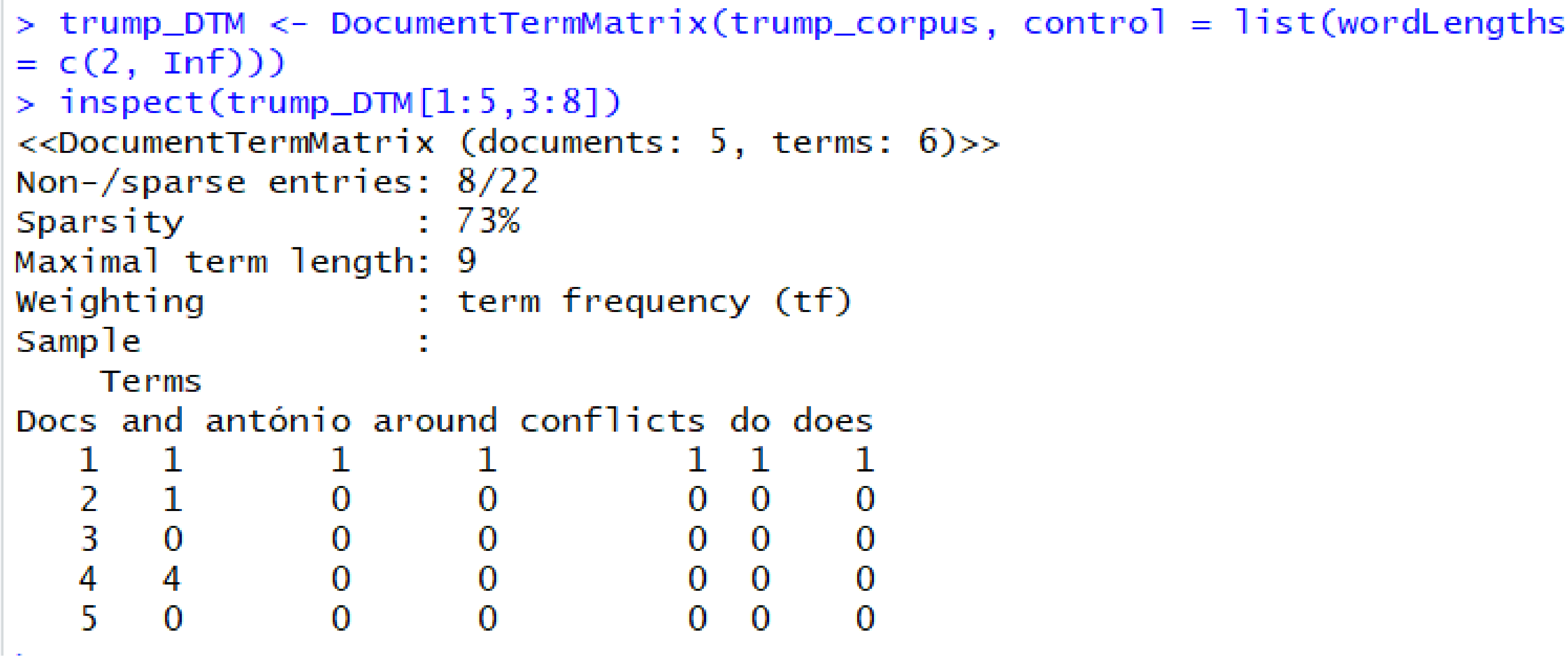
Document Term Matrix

	A	**All**	**Apple**	**As**	**…**	**Zebra**
Doc1	40	10	0	18		60
Doc2	19	20	1	13		0
Doc3	47	29	10	19		0
Doc3	22	13	0	29		0
….						.
Last Doc	15	19	1	40		0

Document Term Matrix

	A	**All**	**Apple**	**As**	**…**	**Zebra**
Doc1	40	10	0	18		60
Doc2	19	20	1	13		0
Doc3	47	29	10	19		0
Doc3	22	13	0	29		0
….						.
Doc100	15	19	1	40		0

Document-Term Matrix



```
> trump_DTM <- DocumentTermMatrix(trump_corpus, control = list(wordLengths
= c(2, Inf)))
> inspect(trump_DTM[1:5,3:8])
<<DocumentTermMatrix (documents: 5, terms: 6)>>
Non-/sparse entries: 8/22
Sparsity           : 73%
Maximal term length: 9
Weighting          : term frequency (tf)
Sample             :
    Terms
Docs and antónio around conflicts do does
   1   1       1      1         1  1    1
   2   1       0      0         0  0    0
   3   0       0      0         0  0    0
   4   4       0      0         0  0    0
   5   0       0      0         0  0    0
```

**This slide image is adapted from Chris Bail's SICSS-Duke Workshop R tutorial*

Quantifying word meaning

“You shall know a word by the company it keeps” (Firth 1953)

- Meaning comes from the distribution of a words’ context words
- Meaning as *relational*
- Meaning as a *system*

Word Meaning

You shall know a word by the company it keeps (Firth, J. R. 1957)

- Word meaning comes from distribution of context words
    - What about ambiguity, polysemy?
    - Extralinguistic context (e.g., liberal vs conservative contexts)?
    - The interpretant?
    - Referents?
    - Conversational data?
    - Order of word/sequences?

Word Meaning

- Word Sense Induction/Disambiguation
  - E.g., river bank or $ bank? fat?
  - “WordNets” in R

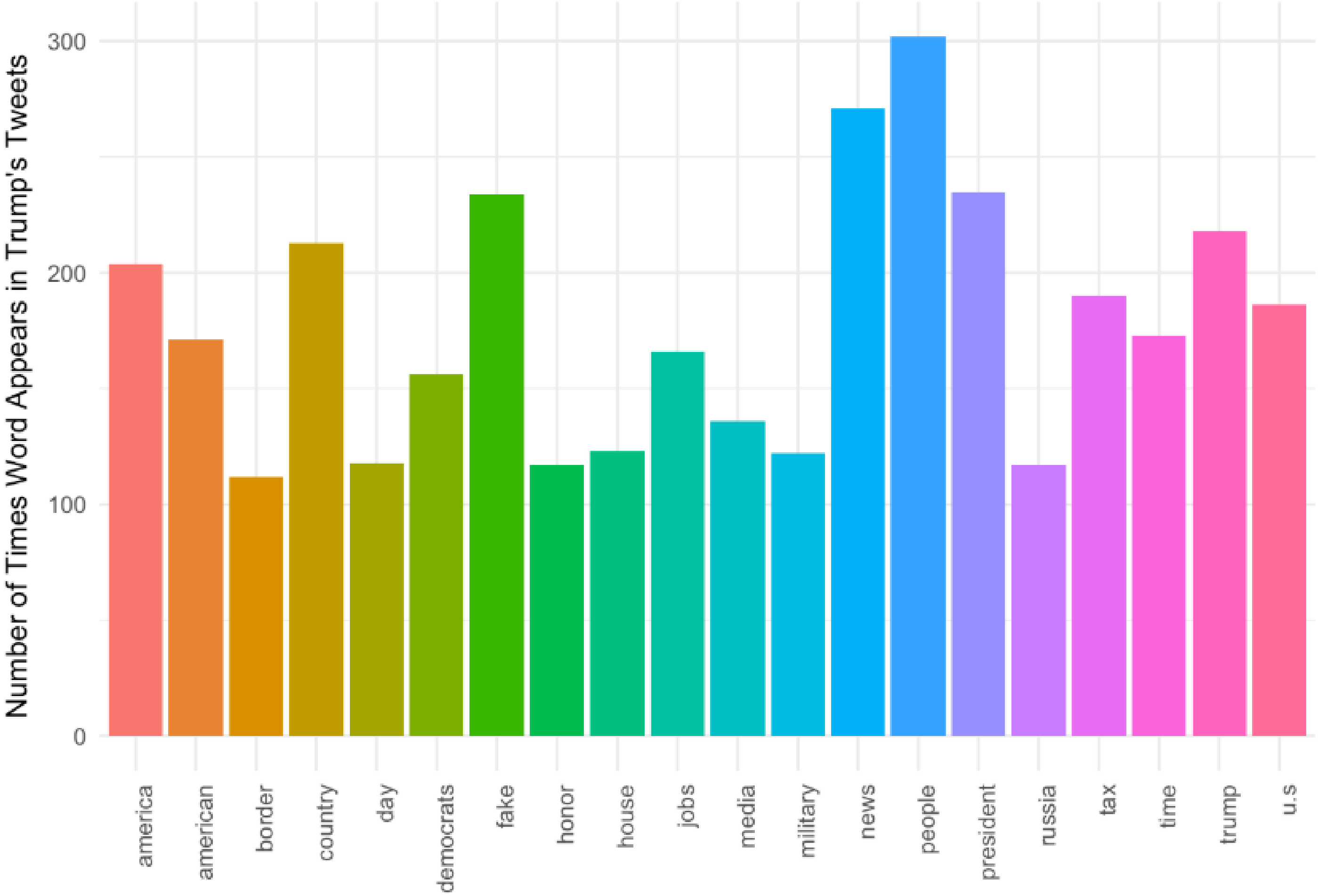
- Part of Speech Tagging

- Co-reference resolution
  - E.g., who is ‘he’ in “The quick brown fox jumped over the dog. He was scared.”

Questions

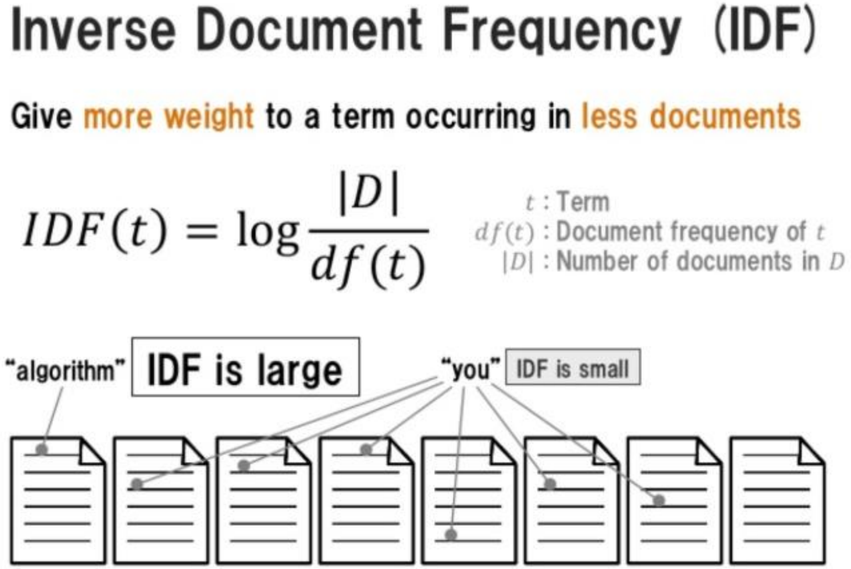
3. Dictionary Methods

Counting


Number of Times Word Appears in Trump's Tweets
300
200
100
0
america
american
border
country
day
democrats
fake
honor
house
jobs
media
military
news
people
president
russia
tax
time
trump
u.s

TF-IDF

Intuition: Weight the frequency (tf) of a word in a document based on how many documents the word occurs in (idf), in the overall corpus



Inverse Document Frequency (IDF)

Give more weight to a term occurring in less documents

$$IDF(t) = \log \frac{|D|}{df(t)}$$

t : Term

$df(t)$: Document frequency of t

$|D|$: Number of documents in D

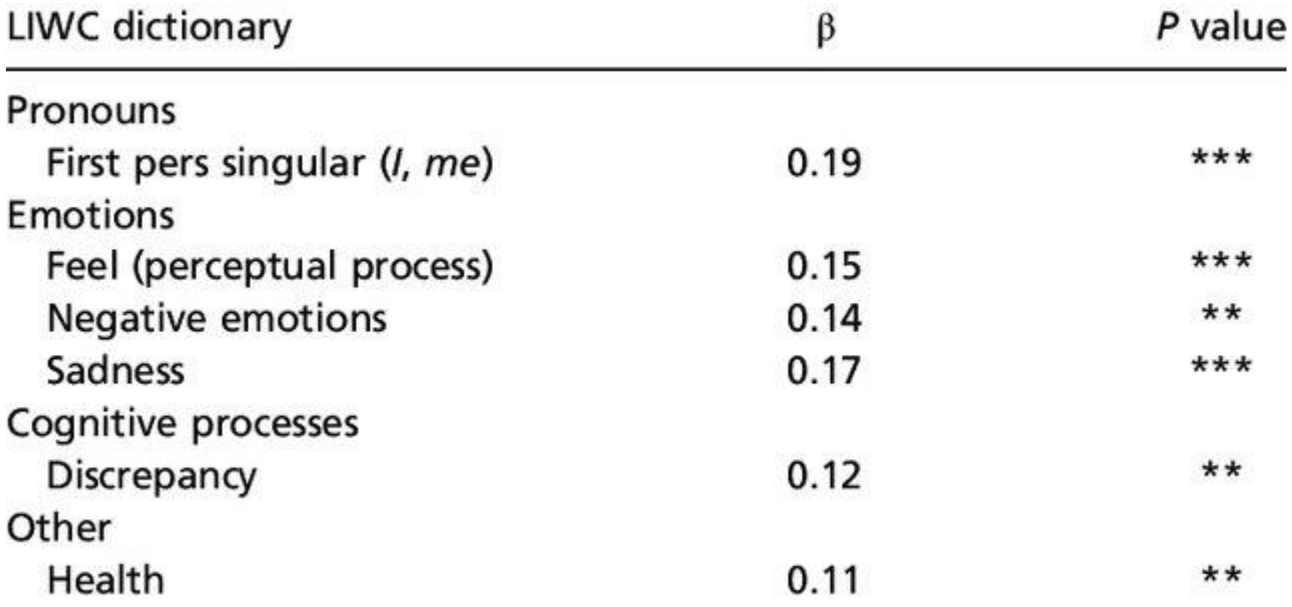
Dictionary Based Methods

- E.g., count frequency of "he" vs "she" in the news, to see if women and men are equally represented, or look if "they" is used more across time compared to "he" and "she"

- E.g., is "I" vs "we" an indicator for clinical depression?

- How do we come up with words in the dictionary?

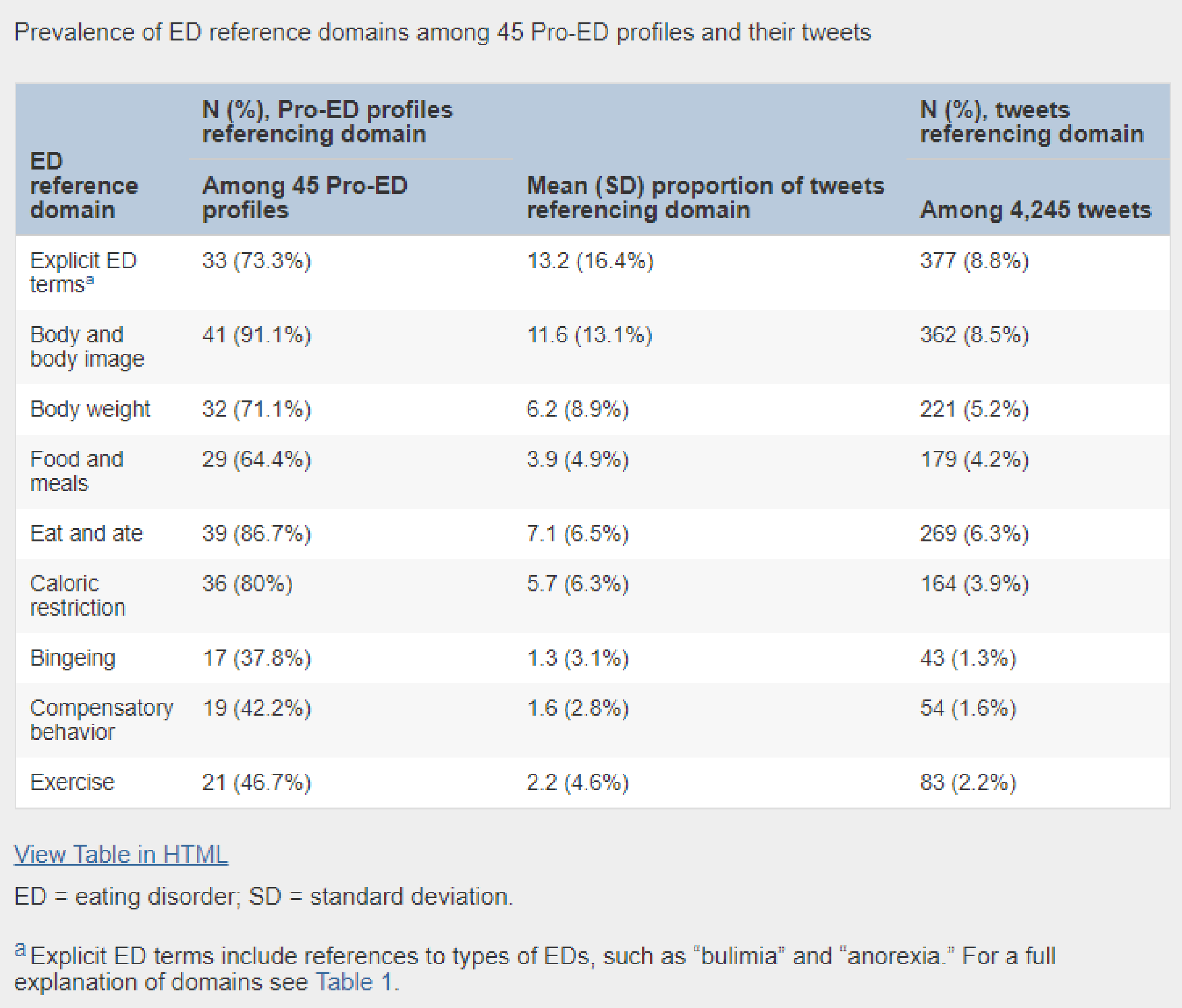
Existing Lexicon/Dictionaries, such as LIWC



LIWC dictionary	β	*P* value
Pronouns		
First pers singular (*I*, *me*)	0.19	***
Emotions		
Feel (perceptual process)	0.15	***
Negative emotions	0.14	**
Sadness	0.17	***
Cognitive processes		
Discrepancy	0.12	**
Other		
Health	0.11	**

Eichstaedt, Johannes C., et al. "Facebook language predicts depression in medical records." *Proceedings of the National Academy of Sciences* 115.44 (2018): 11203-11208.

Or, Create Your Own Dictionary



Prevalence of ED reference domains among 45 Pro-ED profiles and their tweets

ED reference domain	N (%), Pro-ED profiles referencing domain	Mean (SD) proportion of tweets referencing domain	N (%), tweets referencing domain
	Among 45 Pro-ED profiles		Among 4,245 tweets
Explicit ED terms[a]	33 (73.3%)	13.2 (16.4%)	377 (8.8%)
Body and body image	41 (91.1%)	11.6 (13.1%)	362 (8.5%)
Body weight	32 (71.1%)	6.2 (8.9%)	221 (5.2%)
Food and meals	29 (64.4%)	3.9 (4.9%)	179 (4.2%)
Eat and ate	39 (86.7%)	7.1 (6.5%)	269 (6.3%)
Caloric restriction	36 (80%)	5.7 (6.3%)	164 (3.9%)
Bingeing	17 (37.8%)	1.3 (3.1%)	43 (1.3%)
Compensatory behavior	19 (42.2%)	1.6 (2.8%)	54 (1.6%)
Exercise	21 (46.7%)	2.2 (4.6%)	83 (2.2%)

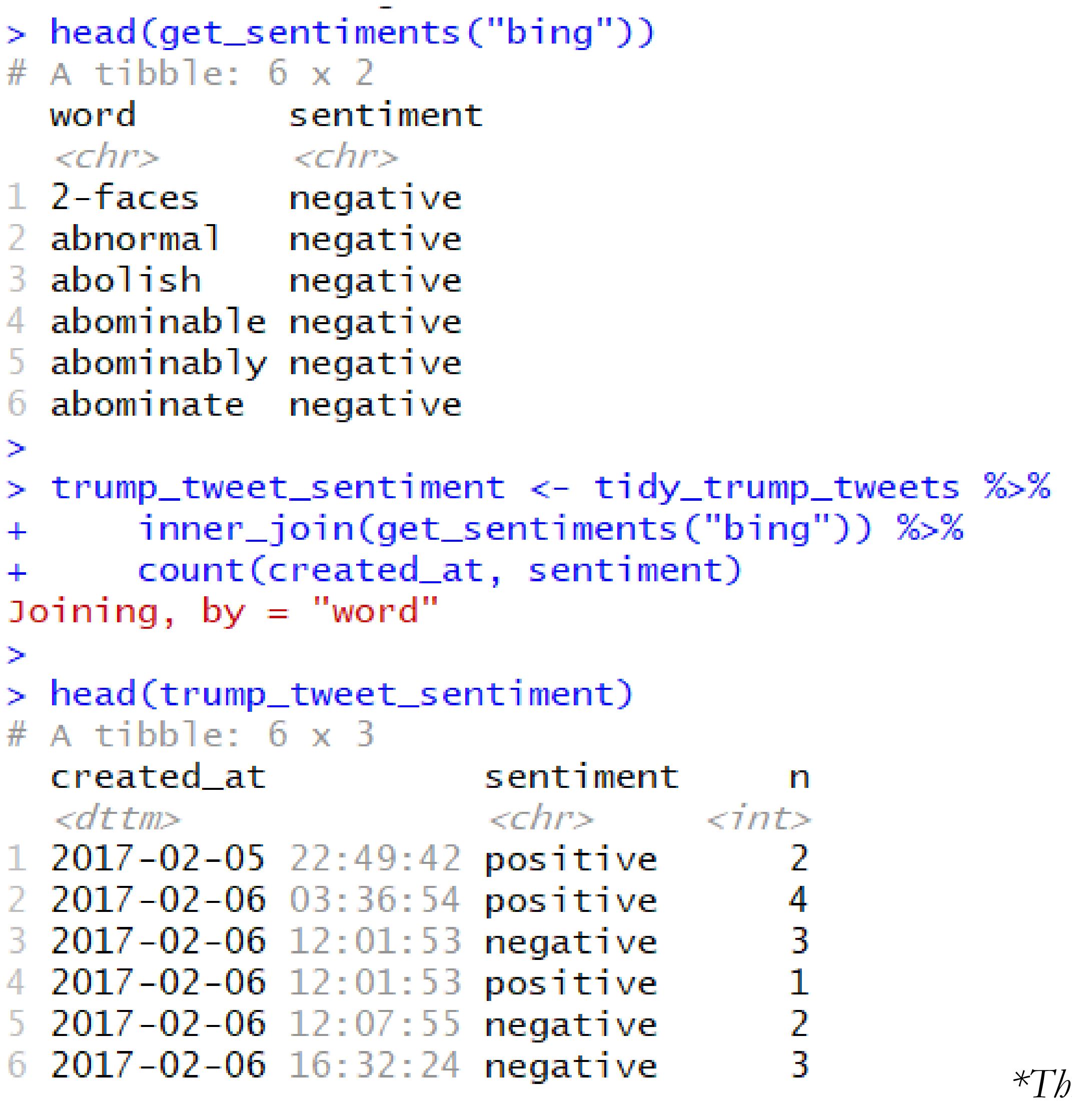
View Table in HTML

ED = eating disorder; SD = standard deviation.

[a] Explicit ED terms include references to types of EDs, such as "bulimia" and "anorexia." For a full explanation of domains see Table 1.

Arseniev-Koehler, Alina, et al. "# Proana: pro-eating disorder socialization on Twitter." *Journal of Adolescent Health* 58.6 (2016): 659-664.

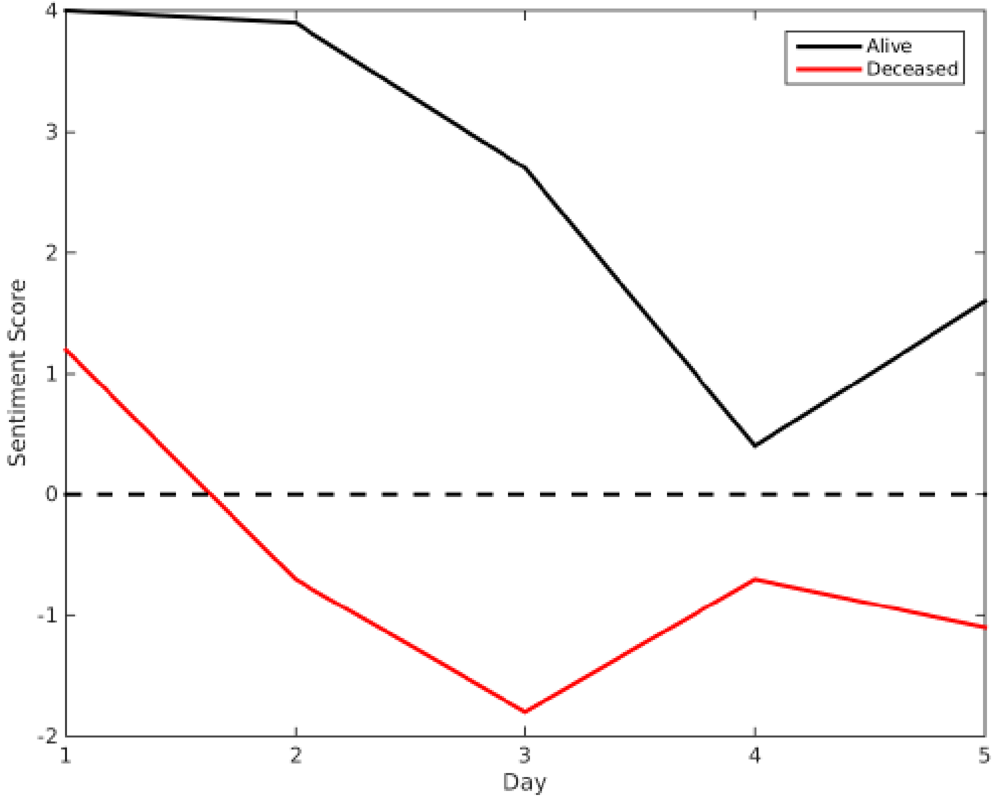
Sentiment Analysis

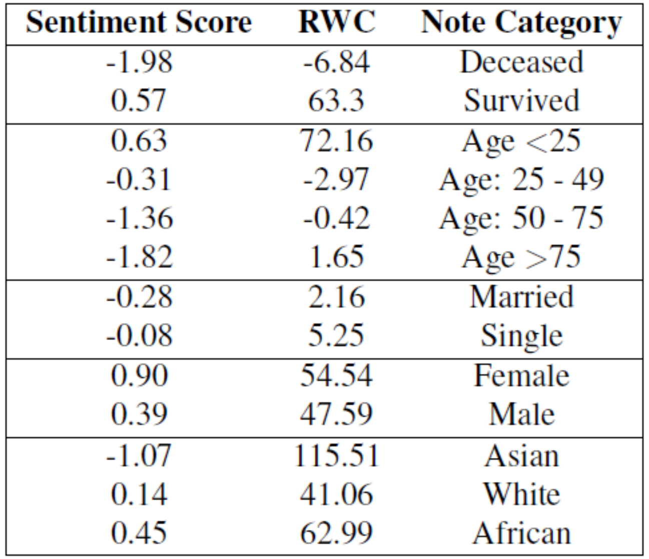


```
> head(get_sentiments("bing"))
# A tibble: 6 x 2
  word       sentiment
  <chr>      <chr>
1 2-faces    negative
2 abnormal   negative
3 abolish    negative
4 abominable negative
5 abominably negative
6 abominate  negative
>
> trump_tweet_sentiment <- tidy_trump_tweets %>%
+     inner_join(get_sentiments("bing")) %>%
+     count(created_at, sentiment)
Joining, by = "word"
>
> head(trump_tweet_sentiment)
# A tibble: 6 x 3
  created_at          sentiment     n
  <dttm>              <chr>     <int>
1 2017-02-05 22:49:42 positive      2
2 2017-02-06 03:36:54 positive      4
3 2017-02-06 12:01:53 negative      3
4 2017-02-06 12:01:53 positive      1
5 2017-02-06 12:07:55 negative      2
6 2017-02-06 16:32:24 negative      3
```

**This slide image is adapted from Chris Bail's SICSS-Duke Workshop R tutorial*

Sentiment Analysis



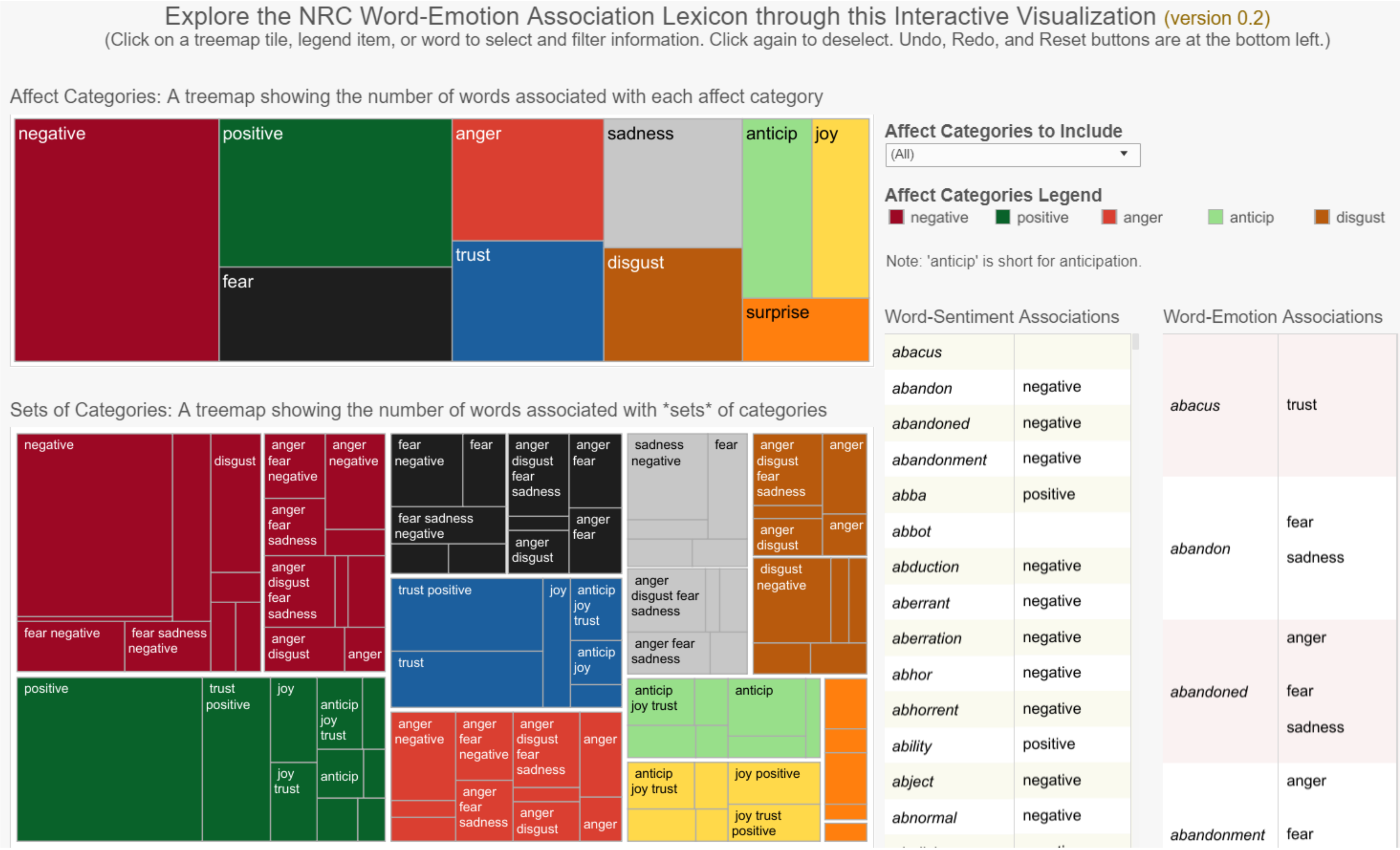


Sentiment Score	RWC	Note Category
-1.98	-6.84	Deceased
0.57	63.3	Survived
0.63	72.16	Age <25
-0.31	-2.97	Age: 25 - 49
-1.36	-0.42	Age: 50 - 75
-1.82	1.65	Age >75
-0.28	2.16	Married
-0.08	5.25	Single
0.90	54.54	Female
0.39	47.59	Male
-1.07	115.51	Asian
0.14	41.06	White
0.45	62.99	African

Table 2. *A comparison of our sentiment score and an alternative score using the ratio of positive to negative terms (RWC) for each of the note categories.*

Ghassemi, Mohammad M., Roger G. Mark, and Shamim Nemati. "A visualization of evolving clinical sentiment using vector representations of clinical notes." *2015 Computing in cardiology conference (CinC)*. IEEE, 2015.

Sentiment & Emotion Analysis



Explore the NRC Word-Emotion Association Lexicon through this Interactive Visualization (version 0.2)

(Click on a treemap tile, legend item, or word to select and filter information. Click again to deselect. Undo, Redo, and Reset buttons are at the bottom left.)

Affect Categories: A treemap showing the number of words associated with each affect category

Affect Categories to Include

(All) ▾

Affect Categories Legend

🟥 negative 🟩 positive 🟥 anger 🟩 anticip 🟧 disgust

Note: 'anticip' is short for anticipation.

Sets of Categories: A treemap showing the number of words associated with *sets* of categories

Word-Sentiment Associations

abacus	
abandon	negative
abandoned	negative
abandonment	negative
abba	positive
abbot	
abduction	negative
aberrant	negative
aberration	negative
abhor	negative
abhorrent	negative
ability	positive
abject	negative
abnormal	negative

Word-Emotion Associations

abacus	trust
abandon	fear sadness
abandoned	anger fear sadness
abandonment	anger fear

Sample lexicon:

- How generated?
- Granularity of sentiment?
- Validated?

- NRC Emotion (and Twitter Sentiment Lexicons)
  - https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
- ANEW Emotion Lexicon
- LIWC (not free, but the most standard)
- General Inquirer (various lexicon):
  - http://www.wjh.harvard.edu/~inquirer/homecat.htm
- Loughran (various lexicon):
  - https://sraf.nd.edu/textual-analysis/resources/
- WordNet Affect (emotion and related lexicon)
  - http://wndomains.fbk.eu/wnaffect.html
- MPQA Lexicon (arguing, sentiment)
  - https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

Questions

4. A Teaser on other Methods

Other methods for analyzing text data

- Topic Modeling
- Word Networks
- Unsupervised Machine Learning
- Word Embeddings
- Supervised Machine Learning

Nancy

"To stop **coronavirus** we need **testing** & **contact tracing**."

Adam

"The **Chinese** made **COVID-19** in a **lab** to kill our **economy**"

Mike

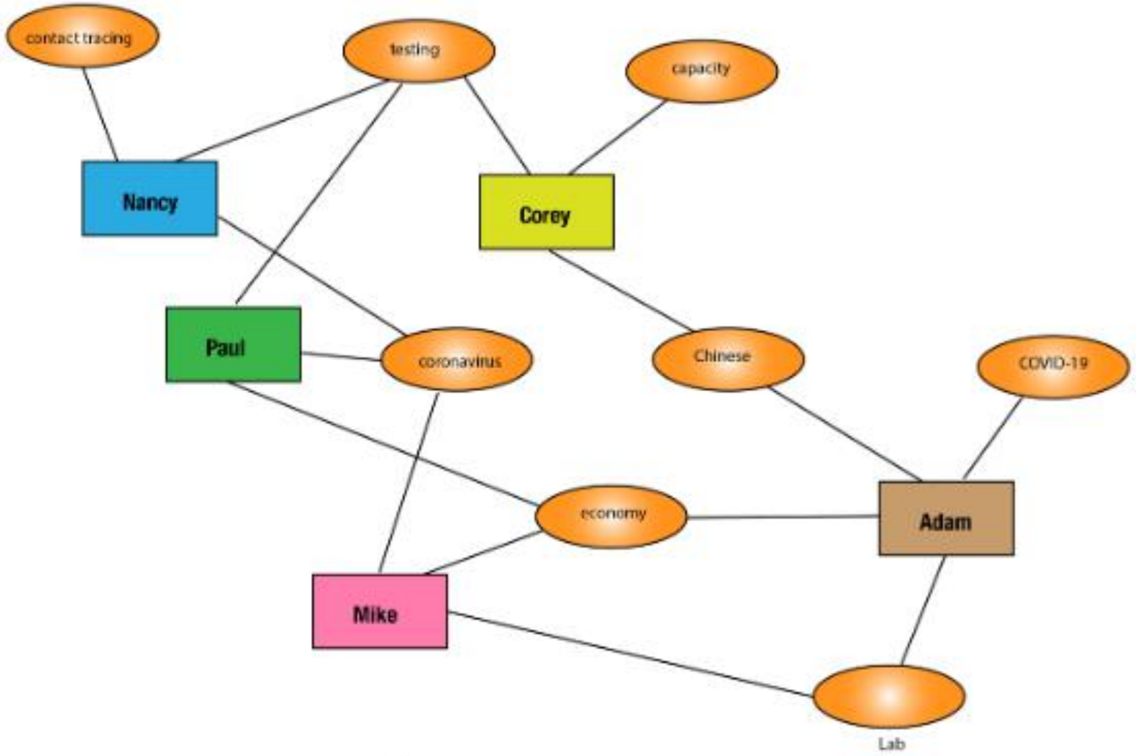
"**Coronavirus** is killing our **economy**, whether it was made in a **lab** or not"

Paul

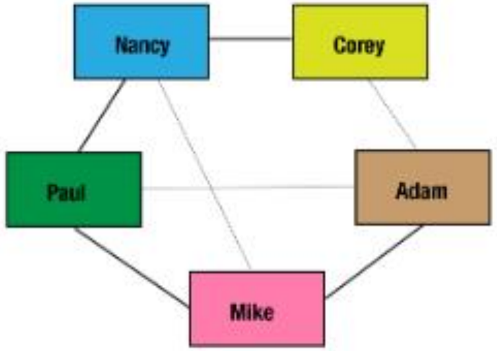
"Our **economy** won't get going again without **testing** for **Coronavirus**"

Corey

"Instead of blaming the **Chinese** we should admire their **testing capacity**"

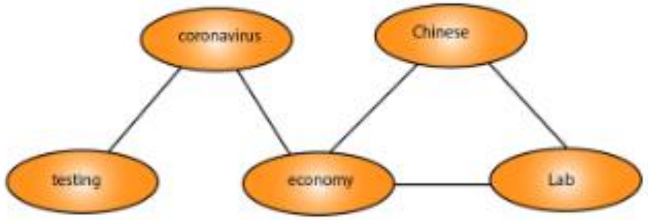


Author Network



(Represents Similarities Between Authors in Terms of their Word Usage)

Word Network

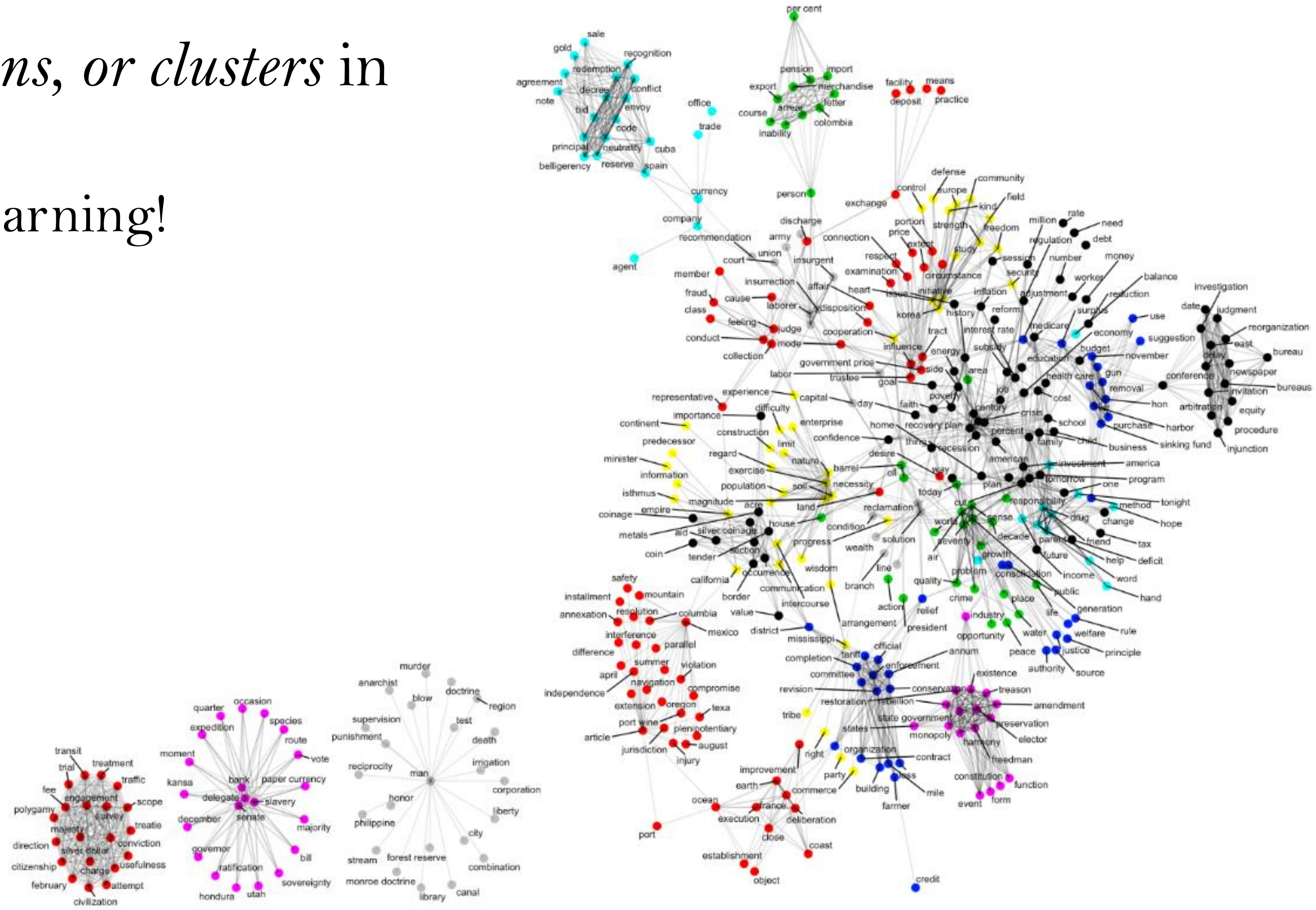


(Represents Similarities between Words used across Authors)

https://compsocialscience.github.io/summer-institute/2020/materials/day3-text-analysis/text-networks/rmarkdown/Text_Networks.html

Text Networks

- Then, you might find *patterns, or clusters* in this data
  - i.e., unsupervised machine learning!



https://compsocialscience.github.io/summer-institute/2020/materials/day3-text-analysis/text-networks/rmarkdown/Text_Networks.html

Topic Modeling

- Find “themes” in your data
- Needle in a haystack…vs the haystack itself
- Close reading + computational analysis
- How many topics?

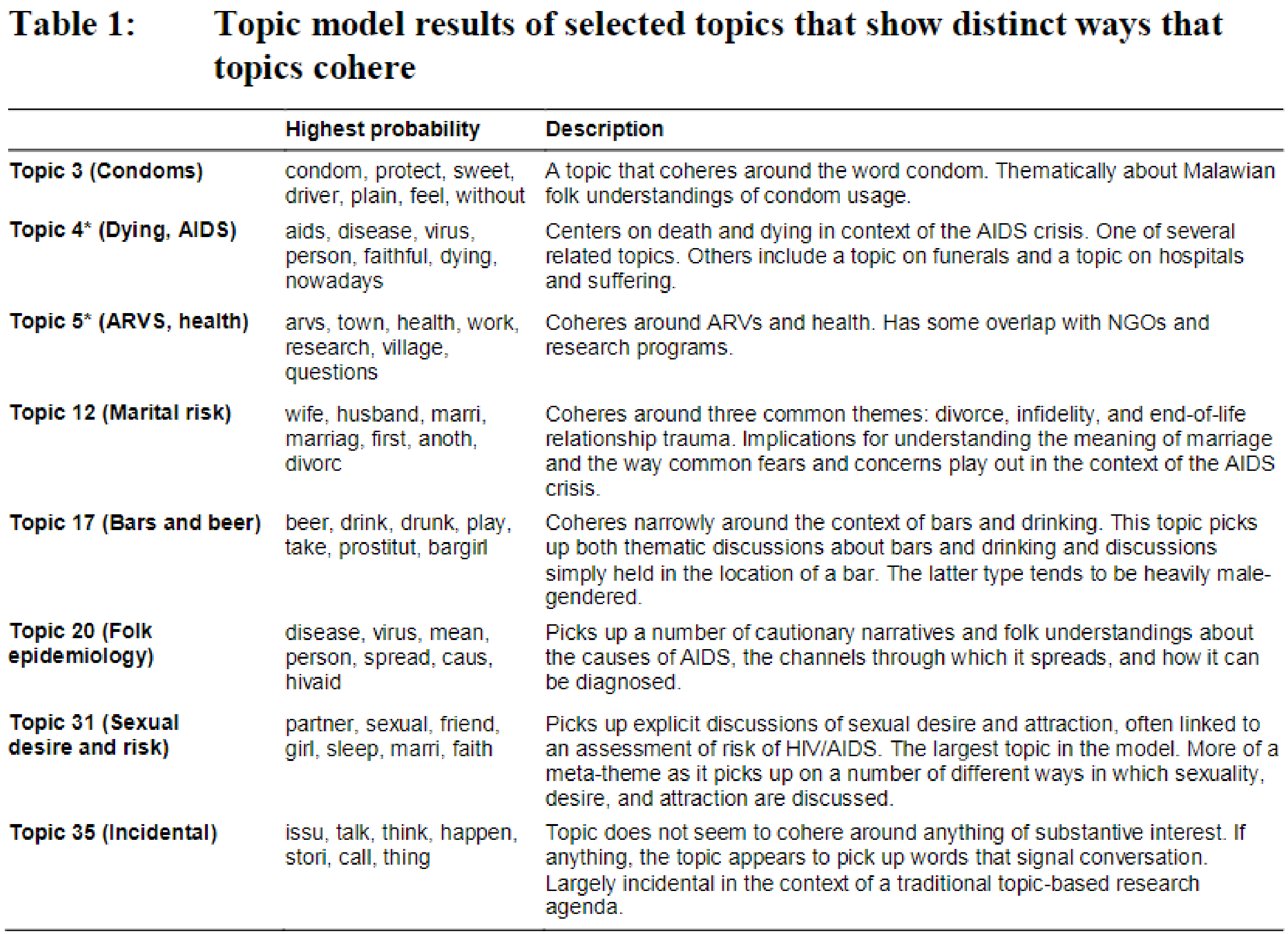
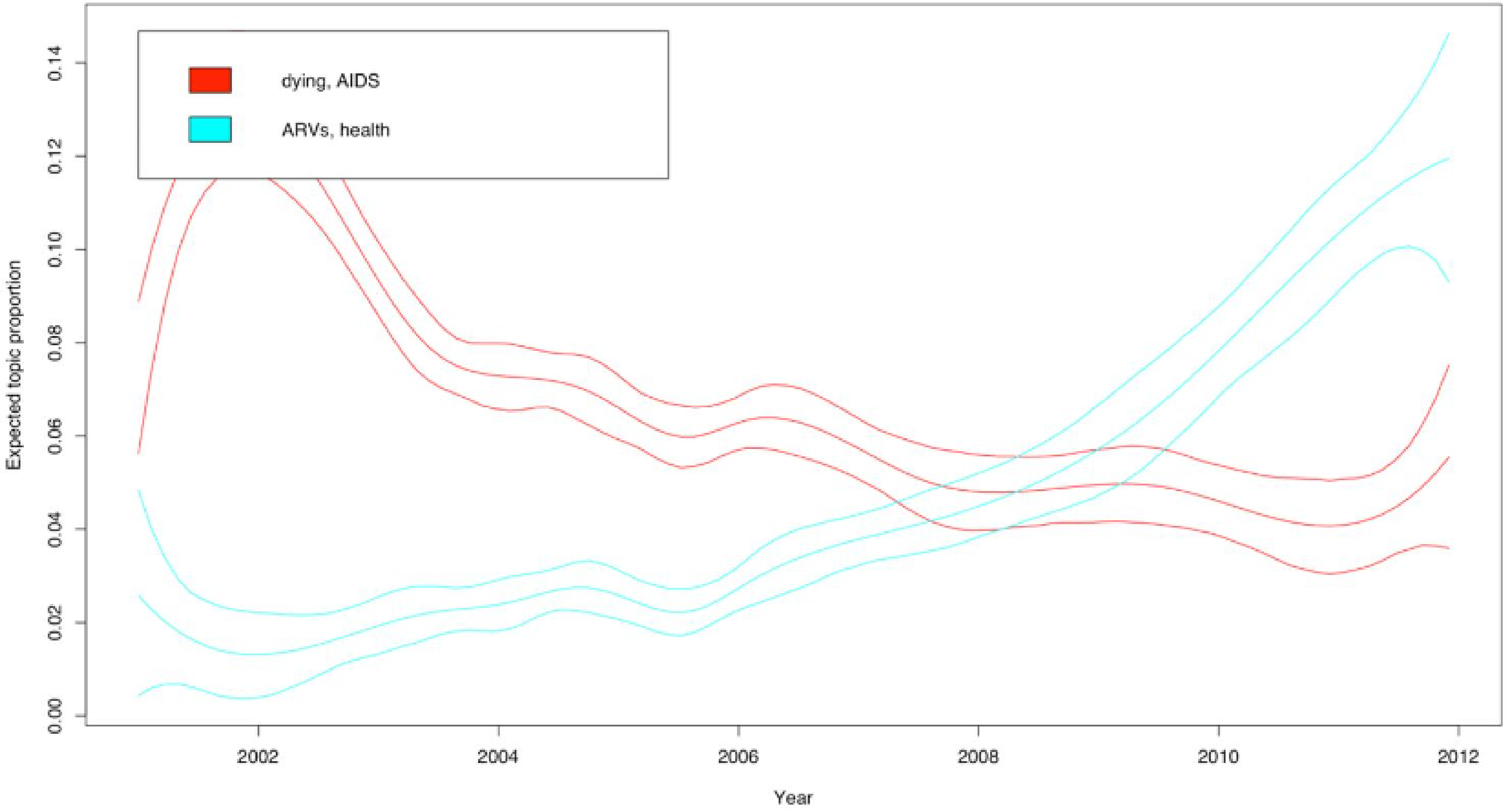


Table 1: Topic model results of selected topics that show distinct ways that topics cohere

	Highest probability	Description
Topic 3 (Condoms)	condom, protect, sweet, driver, plain, feel, without	A topic that coheres around the word condom. Thematically about Malawian folk understandings of condom usage.
Topic 4* (Dying, AIDS)	aids, disease, virus, person, faithful, dying, nowadays	Centers on death and dying in context of the AIDS crisis. One of several related topics. Others include a topic on funerals and a topic on hospitals and suffering.
Topic 5* (ARVS, health)	arvs, town, health, work, research, village, questions	Coheres around ARVs and health. Has some overlap with NGOs and research programs.
Topic 12 (Marital risk)	wife, husband, marri, marriag, first, anoth, divorc	Coheres around three common themes: divorce, infidelity, and end-of-life relationship trauma. Implications for understanding the meaning of marriage and the way common fears and concerns play out in the context of the AIDS crisis.
Topic 17 (Bars and beer)	beer, drink, drunk, play, take, prostitut, bargirl	Coheres narrowly around the context of bars and drinking. This topic picks up both thematic discussions about bars and drinking and discussions simply held in the location of a bar. The latter type tends to be heavily male-gendered.
Topic 20 (Folk epidemiology)	disease, virus, mean, person, spread, caus, hivaid	Picks up a number of cautionary narratives and folk understandings about the causes of AIDS, the channels through which it spreads, and how it can be diagnosed.
Topic 31 (Sexual desire and risk)	partner, sexual, friend, girl, sleep, marri, faith	Picks up explicit discussions of sexual desire and attraction, often linked to an assessment of risk of HIV/AIDS. The largest topic in the model. More of a meta-theme as it picks up on a number of different ways in which sexuality, desire, and attraction are discussed.
Topic 35 (Incidental)	issu, talk, think, happen, stori, call, thing	Topic does not seem to cohere around anything of substantive interest. If anything, the topic appears to pick up words that signal conversation. Largely incidental in the context of a traditional topic-based research agenda.

Chakrabarti, Parijat, and Margaret Frye. "A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography." *Demographic Research* 37 (2017): 1351-1382.

Figure 3: Proportion of the corpus related to topics over time, ‘dying, AIDS’ versus ‘ARVs, health’



ARV= antiretroviral

Notes: The lines represent smoothed trends, with 95% confidence intervals represented by the outer lines.

The Y-axis represents the proportion of the corpus that relates to this topic at a particular point in time. The X-axis represents the date on which the journalist wrote about the conversational incident – within a few days of when the incident itself occurred.

Among the most common words in the ‘dying, AIDS’ topic are: AIDS, disease, dying.

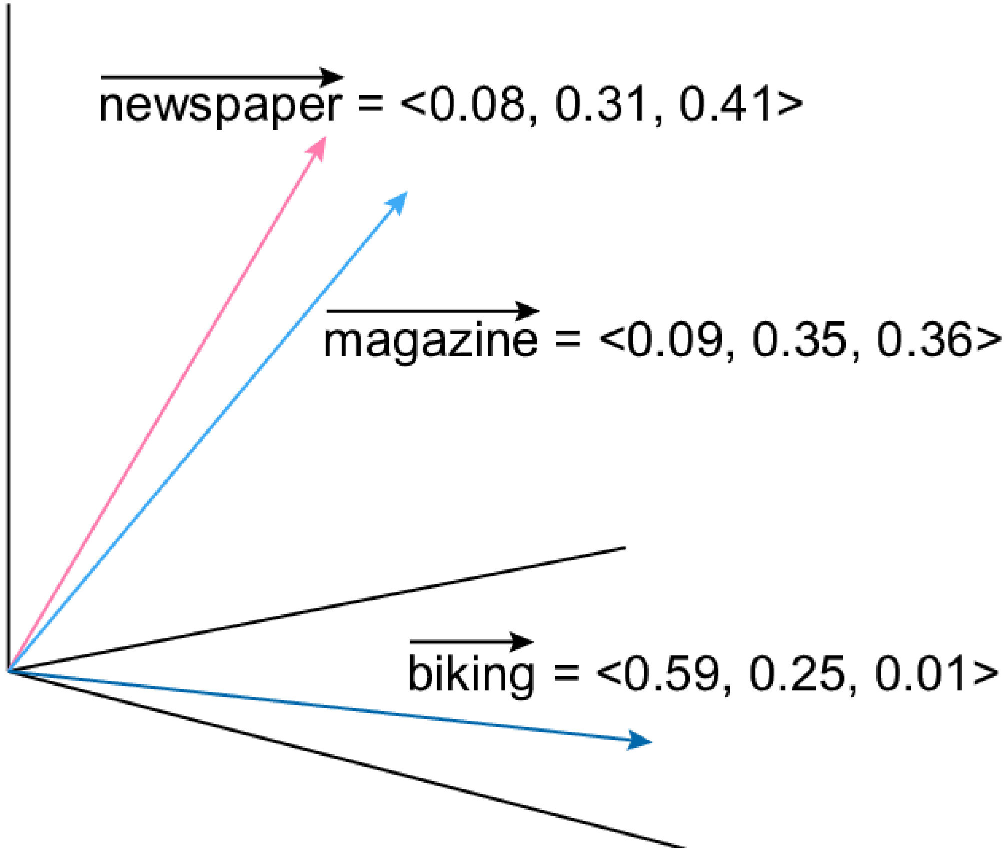
Among the most common words in the ‘ARVs, health’ topic are: ARVs, town, health, work, village, research.

Chakrabarti, Parijat, and Margaret Frye. *Demographic Research* 37 (2017): 1351-1382.

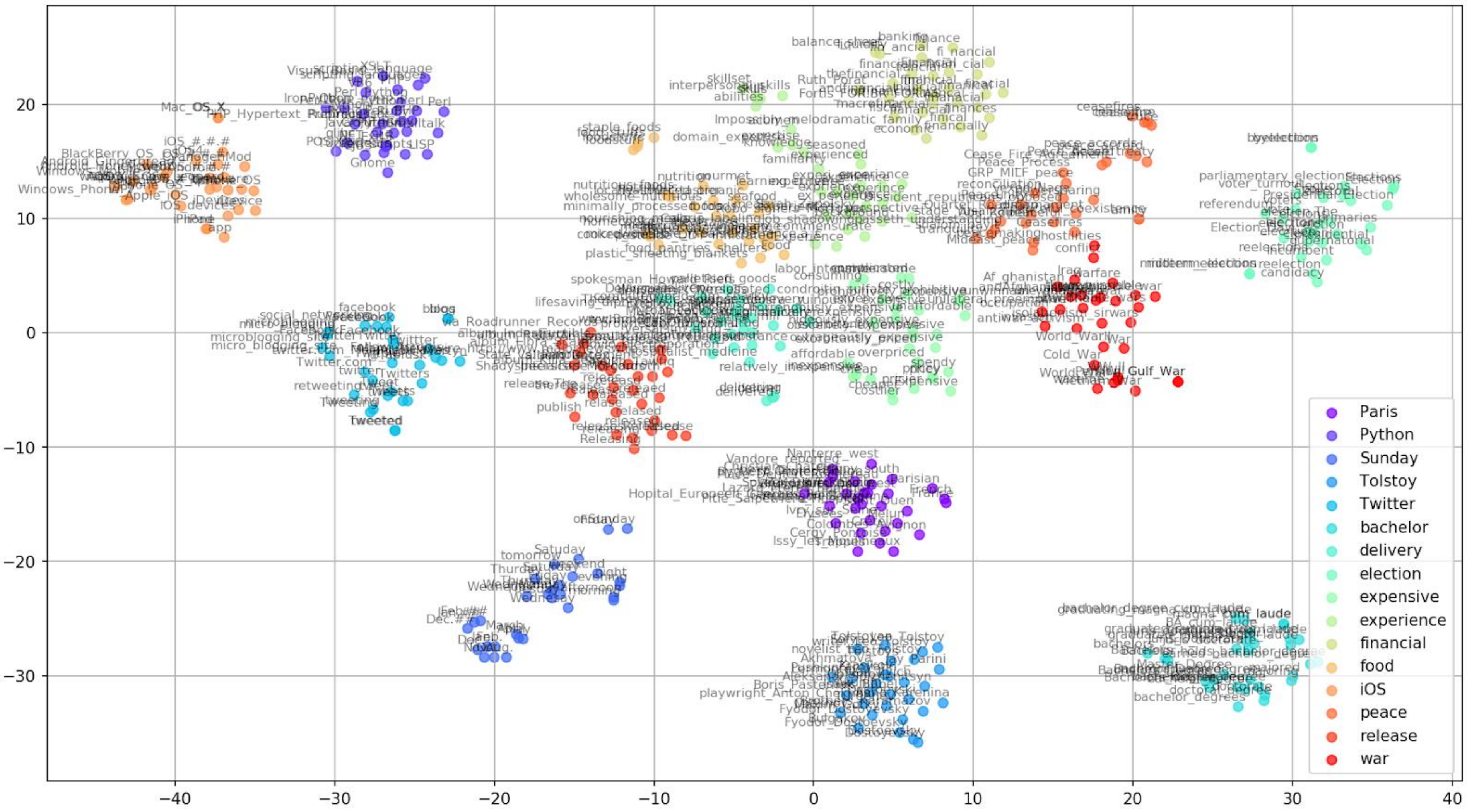
Have you tried topic modeling? What did you do?

Word Embeddings

- Represent each word in a corpus as a vector which corresponds to the way the words is used (compared to other words) the corpus
- Words used in more similar ways, will have more similar vectors
  - measure "similarity" with cosine similarity


$\overrightarrow{\text{newspaper}}$ = <0.08, 0.31, 0.41>
$\overrightarrow{\text{magazine}}$ = <0.09, 0.35, 0.36>
$\overrightarrow{\text{biking}}$ = <0.59, 0.25, 0.01>

Visualizing Word Embeddings with t-SNE



https://towardsdatascience.com/google-news-and-leo-tolstoy-visualizing-word2vec-word-embeddings-with-t-sne-11558d8bd4d

Models to “learn” word embeddings

- Word2Vec (2 variants: SkipGram and CBOW)
- GlovE
- FastText
- BERT and ELMO
- ….

How does Word2Vec learn word-vectors?

Can you guess the missing word?

"…Americans have grown ______ over the last generation, inviting more heart disease, diabetes and premature deaths…"

Which vocabulary word has the vector with highest *cosine similarity* to the context words' vectors?

→We know what the missing word actually is in the NYT ("fatter")

How does Word2Vec learn word-vectors?

Can you guess the missing word?

"…Americans have grown ______ over the last generation, inviting more heart disease, diabetes and premature deaths…"

1. Start with random vectors for each vocab word

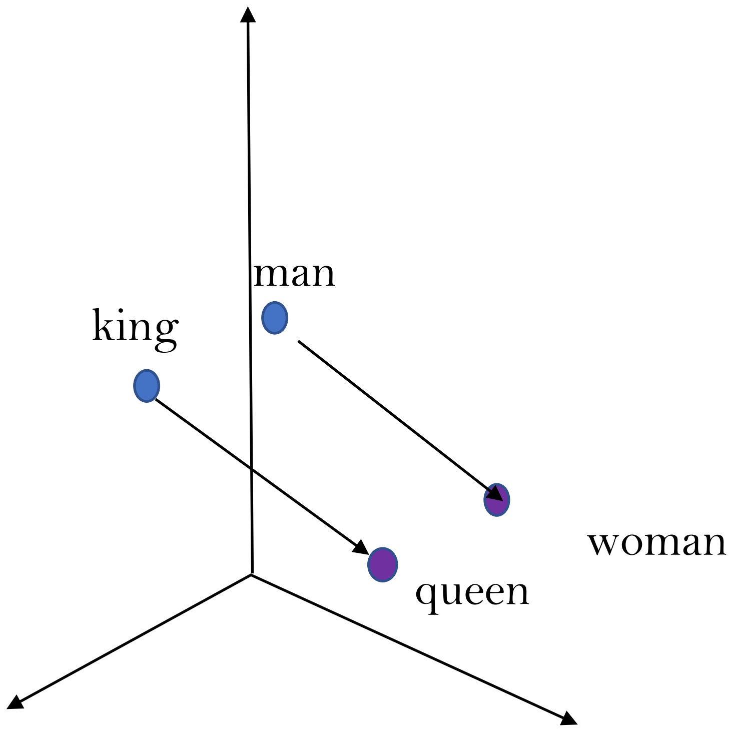
2. Pull out a set of context words from the data and remove one word ("target")
3. Average the vectors of these context words
4. Which word in the vocabulary has the highest *cosine similarity* to this context vector?
5. Correct answer ("fatter"?) Then we have good word-vectors
6. Wrong answer? Tweak the word-vectors so "fatter" is closer to the context word-vectors
7. Go back to step 2

Implementing Word2Vec

- Collect text data
- Clean the data
- Train a model
  - Decide hyperparameters (e.g., 100-d word-vectors? 10 context words?)
  - Validate with Google Analogy Test or WordSim Test

- OR, use a pre-trained model

Latent Dimensions and Analogies



a "gender" dimension

"man" is to "woman" as "king" is to _____?

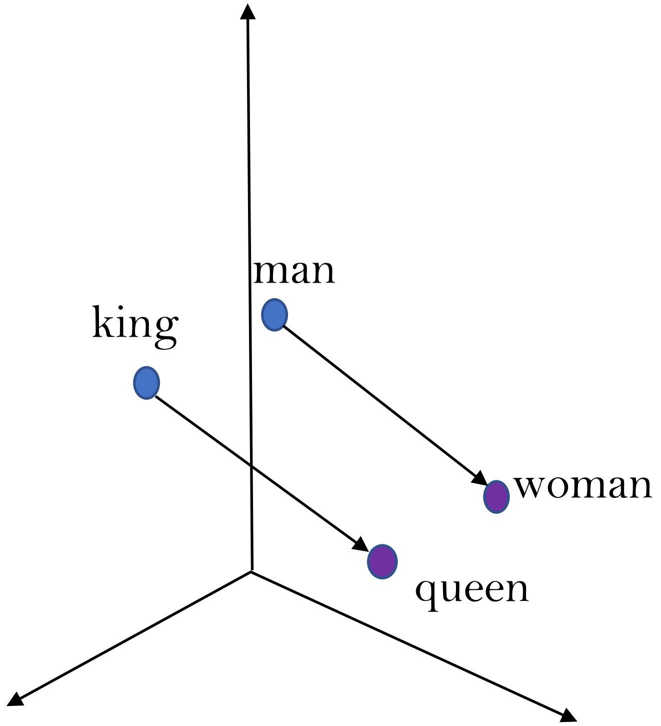
Analogy Task:
woman-man= queen – king

(woman-man) + king= queen

<u>Now, try to solve with word-vectors:</u>
What is the closest word-vector to:

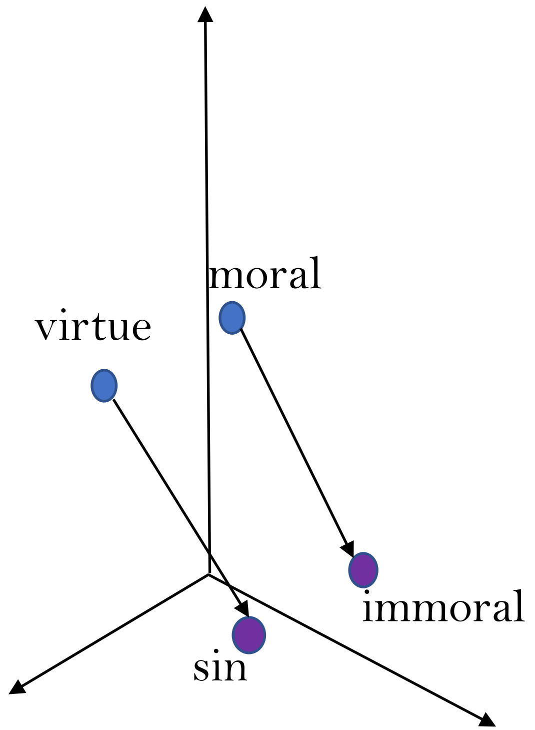
(woman-man) + king ?

How to Extract Latent Dimensions



a “gender” dimension

=AVG(feminine words) – AVG(masculine words)

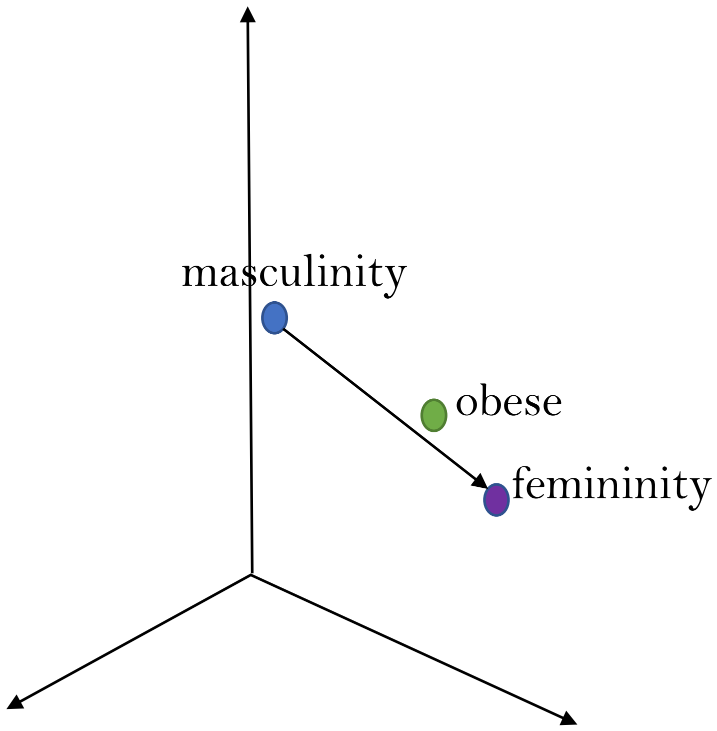


a “moral” dimension

=AVG(moral words) – AVG(immoral words)

*other methods, too

Using Latent Dimensions



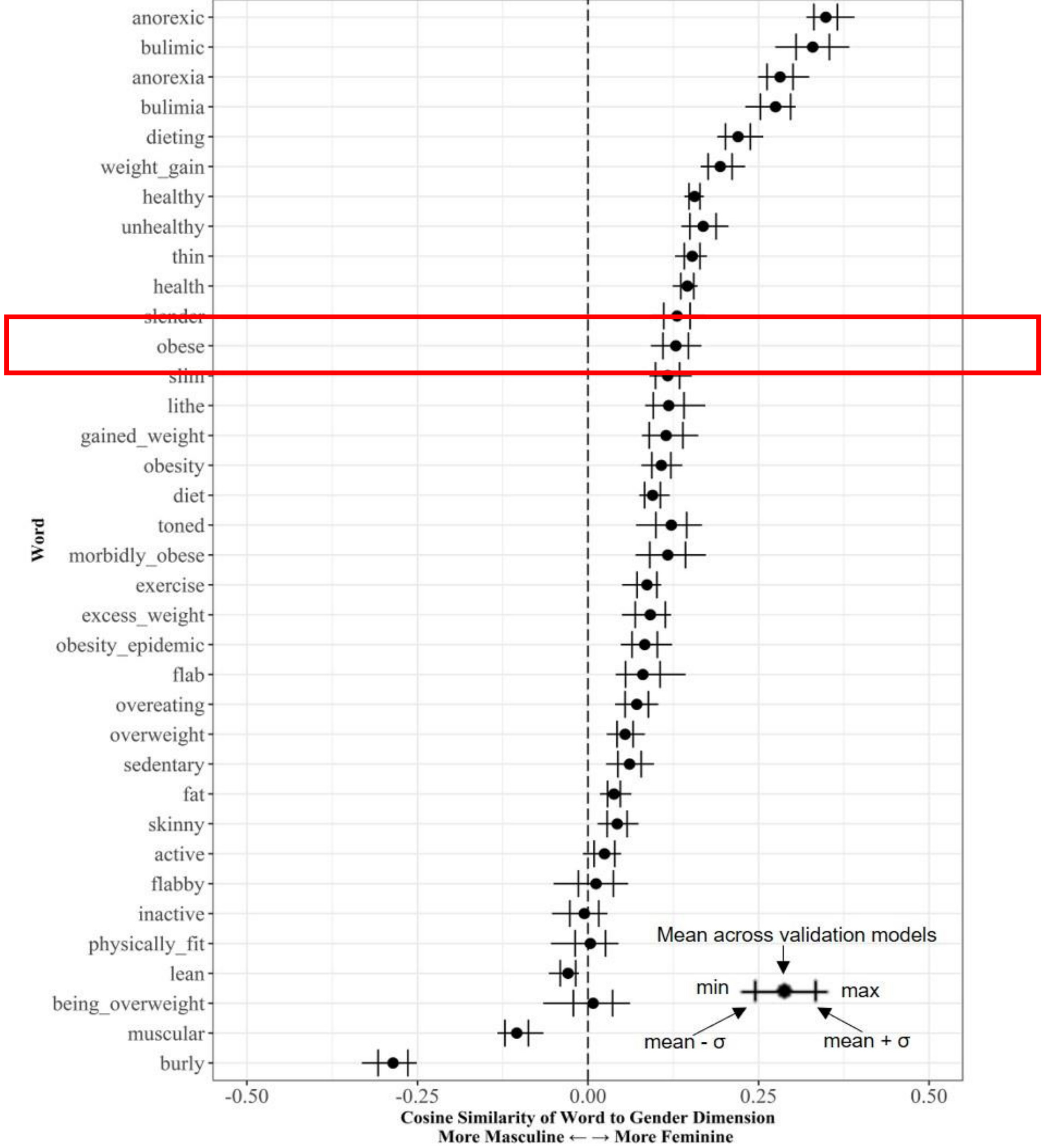
Cosine similarity (gender, overweight) = .4 (possible range: -1 to 1)

Tells you if "obese" is on the feminine or masculine side, and *how* feminine or masculine

a "gender" dimension

=AVG(feminine words) – AVG(masculine words)

Obesity and Gender



….but so is body, weight, and health discourse at large

Arseniev-Koehler, Alina and Jacob Foster. “Machine Learning as a Model for Cultural Learning: Teaching an Algorithm the Meaning of Fat”

Supervised Machine Learning and Text

- E.g., sentiment analysis
- E.g., predict missing variable
- E.g., detect hate speech
- E.g., gender connotations of words (another method)
  - Train a classifier to predict word-vectors as feminine or masculine
  - How does it predict the gender of "overweight," "thin," "slender," "potbelly," "chiseled," "voluptuous" or "pretty"?

Sample Text Datasets

- Social media, blogs
- Malawi narratives (Chaktrabati and Frye), "Becoming a Nazi" narratives (Bearman and Stovel)
- News, Magazines
  - (e.g., NYT Annotated Corpus, Reuters, GoogleNews pretrained word-vectors, LexisNexis)
- DocSouth Narratives: https://docsouth.unc.edu/neh/
- Movie scripts
- Corpus of Historical American English
- Yelp, Amazon Reviews
- Song lyrics
- Advertisements
- Scientific articles

When have you used text data and how did you use it?

Discussion:

- What are some text data sources you are interested in, or know of?
- What are possible research questions to ask with the text data you know about? Or, what are possible research questions that might use text data in general?
- What kinds of methods could you use to solve your question? Are there different methods you could use to answer the same question?

Some jargon on tools and methods

- “NLP,” “Computational Linguistics”
    - Part-of-Speech Tagging
    - Co-Reference Resolution (The woman told her daughter *she* needs to …)
    - Word Sense Disambiguation (river bank or $ bank?)
    - Named entity extraction, reconciliation
    - Sentiment and emotion analysis
- Dictionary Based Methods
    - LIWC, ANEW
- Topic Models
- Word Embeddings (especially, Neural Word Embeddings), Vector Space Models
- Semantic networks

5. Breakout room activity:

1. Go through the SICSS-Duke 2020 R tutorial on the basics of text analysis together: https://compsocialscience.github.io/summer-institute/2020/materials/day3-text-analysis/basic-text-analysis/rmarkdown/Basic_Text_Analysis_in_R.html

Then, either:

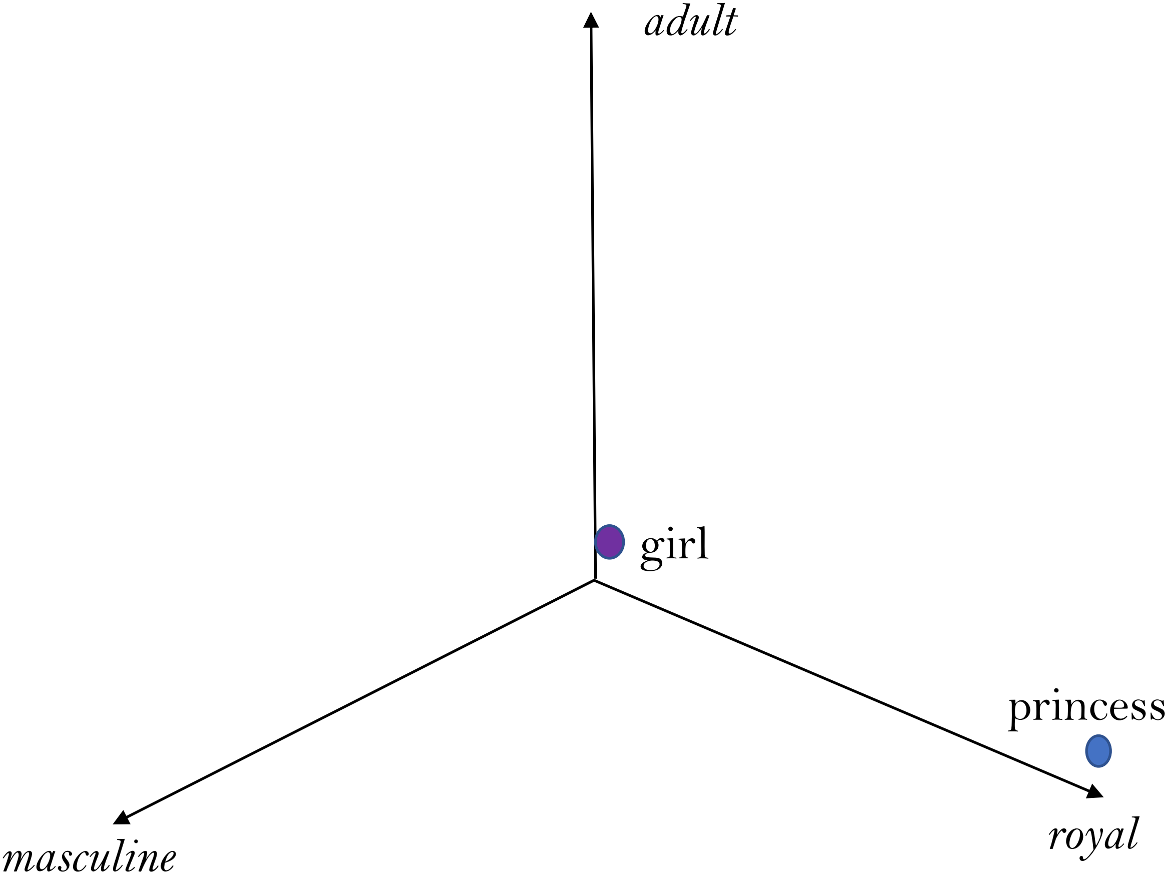
1. Choose one of the SICSS-Duke 2020 R tutorials (dictionary-based analysis, topic modeling, or text networks and try your best to replicate it together). (See annotated code on https://compsocialscience.github.io/summer-institute/curriculum#day_3)

2. (More challenging, but this could turn into your project next week): Identify text data that is interesting to you, and a research question. Then, think about how you would answer this question, and start trying out cleaning and analyses.

Questions

EXTRA SLIDES

Vector Word Spaces



Vocabulary Word	Gender	Royalty	Age
King	0	1	1
Queen	1	1	1
Man	0	0	1
Woman	1	0	1
Girl	1	0	0
Boy	0	0	0
Princess	1	1	0
Prince	0	1	0