Tools for Analyzing Lexical Variation & Social Meaning in Twitter Data

Gyula Zsombok

Assistant Professor, French & Francophone Studies

Joseph Roy

Director, Institutional Research & Analytics

Angela Erdiaw-Kwasie

Senior Data Analyst

Middlebury

ASEE | INSTITUTIONAL RESEARCH + ANALYTICS

Acknowledgements

This work is part of an on-going project on the effects of COVID-19 on Engineering Education funded by NSF-EEC Award # 1748840.

ASEE IRA Staff who contributed to the graphics, data presented and provided initial feedback for this workshop:
Timothy King, Data Collections Support Specialist
Charles Stuppard, Data Analyst II

ASEE Data Collection National Advisory Board: Robert Cassidy (UC Irvine), Timothy Chow (Rose-Hulman Institute of Tech.), Patreena Parsons (Johns Hopkins), M.K. Poindexter (Morgan State), Kathryn Flack Potts (Stanford), Judith Stoddard (Rensselaer), Letticia Ramlal-Lamble(Virginia Tech), David Robledo (Purdue) and JoAnne Valdenegro (Arizona State).

A warm thank you to the NWAV 49 organizing committee for hosting the first virtual NWAV!

Workshop Agenda

- Background
- Twitter Data in R
- Regular Expressions
  - Coding linguistic variables
  - Coding social variables
- Social Media Metrics (Twitter)
- Finding Meaning
  - Topic Models
  - Sentence embeddings

To get the most out of this workshop, we assume you have a working knowledge of the following:

- R, Rstudio
- Tidyverse: ggplot2, dplyr, etc
- Setting up an R project

Sociolinguistics & Twitter

- Jones, T. (2015). Toward a description of African American vernacular english dialect regions using “Black Twitter”. *American Speech*, 90(4), 403-440.
- Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. Frontiers in Artificial Intelligence, 2, 11.
- Gonçalves, B., & Sánchez, D. (2016). Learning about Spanish dialects through Twitter. Revista Internacional de Lingüística Iberoamericana, 65-75.
- Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Computers, environment and urban systems*, 59, 244-255.
- Ilbury, C. (2020). “Sassy Queens”: Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, 24(2), 245-264.
- Zsombok, G. (2022, forthcoming). Language Ideologies in the Age of the Internet: Hashtag on French Twitter. *Journal of French Language Studies*, *Special Issue on French Variation in Digital Media*.

Panel: Twitter

15:20 - 17:20 Saturday, 23rd October, 2021

---

8 Twitter as a laboratory for language variation and change. New opportunities for social media-based sociolinguistic research

Chair

Stefan Grondelaers
Radboud University Nijmegen, Nijmegen, Netherlands

Discussant

Jane Stuart-Smith
University of Glasgow, Glasgow, United Kingdom

Session proposal abstract

The past decade has witnessed an upsurge in studies which use the microblogging service Twitter as a source of primary data (see Hinrichs 2015 for a general introduction on social media-based variation research, and Squires 2016 for a number of earlier studies based on Twitter data). This special session is dedicated to expanding the possibilities for sociolinguistic exploitation of this vast and valuable data source.

Tweets represent the only social media data which are freely available in enormous quantities, and on account of their “conceptual orality” character (Androutsopoulos 2011: 149), tweets contain casual speech features, and manifest a standardness bandwidth which is much wider than prescriptively partial print materials. Tweets are littered with non-standard orthography which is the result of error, or expressive or indexical resourcefulness (Coats 2016: 188): Twitter shares with authentic speech the presence of phonetic, lexical, and morphosyntactic cues which systematically reveal identities and stances of tweeters.

In fact, Twitter distributions of investigated variants align so well with traditionally observed distributions, that Twitter is eminently suited to probing patterns which would otherwise require an unfeasibly large data collection effort (see Grieve et al. 2019 for lexical evidence, Strelluf 2019 for syntactic confirmation, and Van Halteren et al. 2018 for a dialect-geographical application). In addition, Twitter has proven beneficial for the investigation of non-standard or emergent features which are (still) marginal in print (Bohmann 2016), but also – given its larger toolbox of expressive possibilities – for the investigation of specific stylizations, like the “Sassy Queen” (Ilbury 2020).

What we hope you get out of this

- Beginner to Intermediate in R
  - Templates for handling text data in R
  - Advanced ways of modeling text data (topic modeling, USE)
  - Regular expressions
- Intermediate to Advanced in R
  - Template code
    - Topic Modeling
    - Social Media Analysis
  - Introduction to Sentence Embedding Models + Tools

Use of online platforms, apps varies – sometimes widely – by demographic group

% of U.S. adults in each demographic group who say they ever use ...

0% 20 40 60 80 100

	YouTube	Facebook	Instagram	Pinterest	LinkedIn	Snapcha	Twitter	WhatsApp	TikTok	Reddit	Nextdoor
Total	81	69	40	31	28	25	23	23	21	18	13
Men	82	61	36	16	31	22	25	26	17	23	10
Women	80	77	44	46	26	28	22	21	24	12	16
White	79	67	35	34	29	23	22	16	18	17	15
Black	84	74	49	35	27	26	29	23	30	17	10
Hispanic	85	72	52	18	19	31	23	46	31	14	8
Ages 18-29	95	70	71	32	30	65	42	24	48	36	5
30-49	91	77	48	34	36	24	27	30	22	22	17
50-64	83	73	29	38	33	12	18	23	14	10	16
65+	49	50	13	18	11	2	7	10	4	3	8
<$30K	75	70	35	21	12	25	12	23	22	10	6
$30K-$49,999	83	76	45	33	21	27	29	20	29	17	11
$50K-$74,999	79	61	39	29	21	29	22	19	20	20	12
$75K+	90	70	47	40	50	28	34	29	20	26	20
HS or less	70	64	30	22	10	21	14	20	21	9	4
Some college	86	71	44	36	28	32	26	16	24	20	12
College+	89	73	49	37	51	23	33	33	19	26	24
Urban	84	70	45	30	30	28	27	28	24	18	17
Suburban	81	70	41	32	33	25	23	23	20	21	14
Rural	74	67	25	34	15	18	18	9	16	10	2

Note: White and Black adults include those who report being only one race and are not Hispanic. Hispanics are of any race. Not all numerical differences between groups shown are statistically significant (e.g., there are no statistically significant differences between the shares of White, Black or Hispanic Americans who say the use Facebook). Respondents who did not give an answer are not shown.

Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.

"Social Media Use in 2021"

PEW RESEARCH CENTER

Each type of social media platform is used by different demographic groups.

Any data drawn from any from a social media platform will be biased (statistically) by its user base. This is also true for non-randomized interview techniques employed in sociolinguistics (e.g. recruiting from friend-of-friends, or a central location introduces statistical biases into the sample).

Twitter User Profiles:

Black > {White, Hispanic}

{30-50k;75k+} > 50-75k > {less than 30k}

Urban > Suburban >Rural

College + > Some College >> HS

The percent of people who ever use Twitter for the US in Jan-Feb 2021 is 23%.

Data in this workshop

- Objective
  - Understanding institutional communication strategies in response to COVID-19 and subsequent national events after March 2020 in comparison with the engineering education communities' response
- Example Data Set (provided to you): August 2020
- Full data set: May 2020 to present
  - Original Dataset through twitter public API
    - Every week pulling data that referenced engineering and engineering adjacent terms with higher education terms
  - Access through twitter research API to March 2020 and prior

Groups Tracked

- Institutions
  - Profiles (Participating in ASEE’s annual survey)
  - Verified
  - Unverified
- Community
  - Verified
  - Unverified

Finding Meaning in Text

Objectives

- Give a high-level overview of techniques to extract (or find) meaning in textual data
- Point to libraries and show examples of this with twitter data
- Highlight adjacent work in computational historical linguistics with similar techniques

Topic Modeling

- “Topic”: A group of words that share the same context and are likely to co-occur together within one document.
- From Blei (2012: 77) *Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents.* Large is greater than 100 documents or texts. The number of topics, k, should be much less than the number of documents in your corpus.
- Bag of words approach: no structure accounted for and order not considered

Latent Dirichlet Allocation

- http://papers.nips.cc/paper/2070-latent-dirichlet-allocation.pdf
- Set of documents that contain different topics.
- We want to estimate these topics in each document, but each topic is made up of words associated with that topic at some probability.
- The topic structure (i.e. the topics in each document and the word probabilities associated with each topic) are hidden [not observed]
- The documents are observed.
- Blei (2012)
  - Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. *http://dl.acm.org/citation.cfm?id=2133826*
  - 17,000 articles from last 50 years in journal *Science*

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

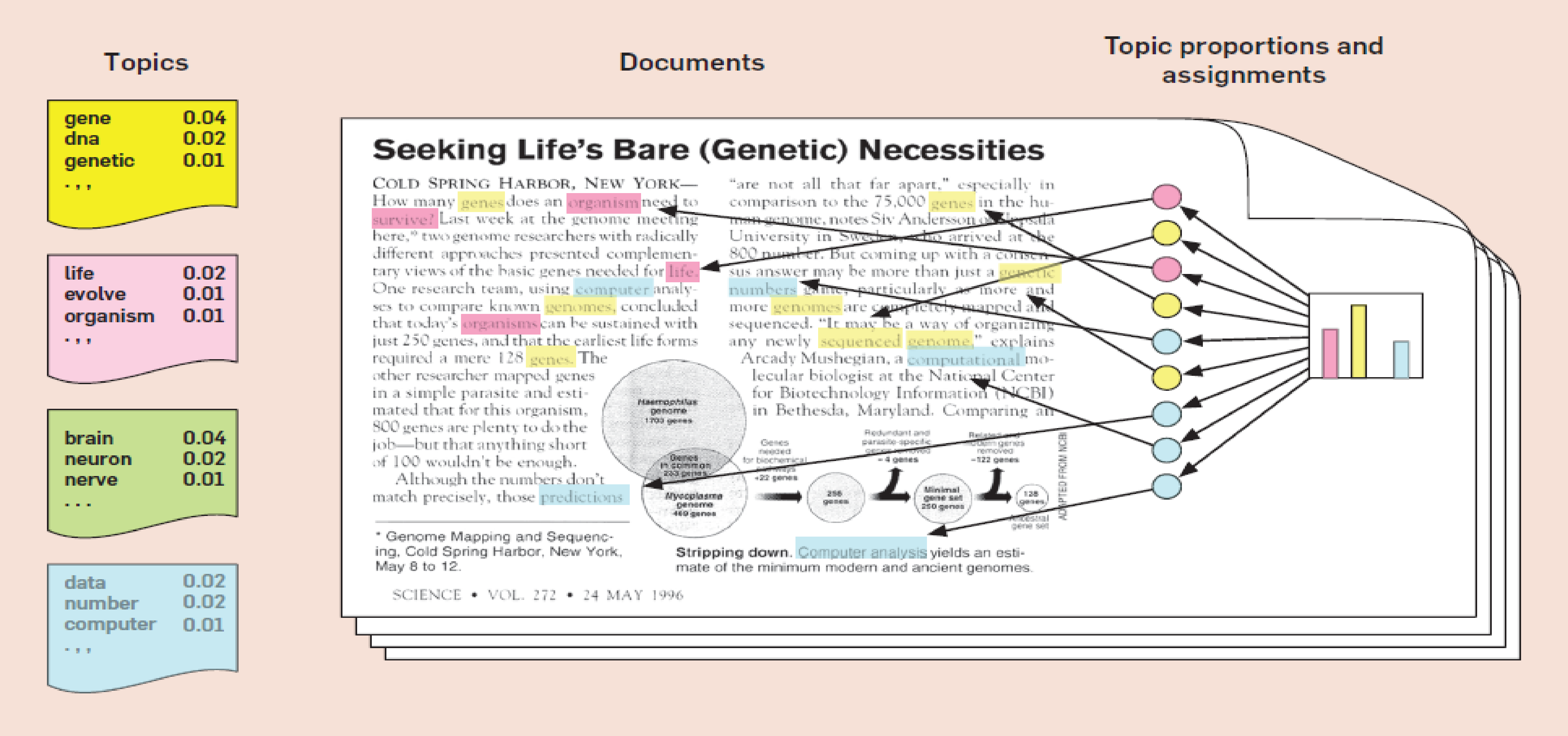
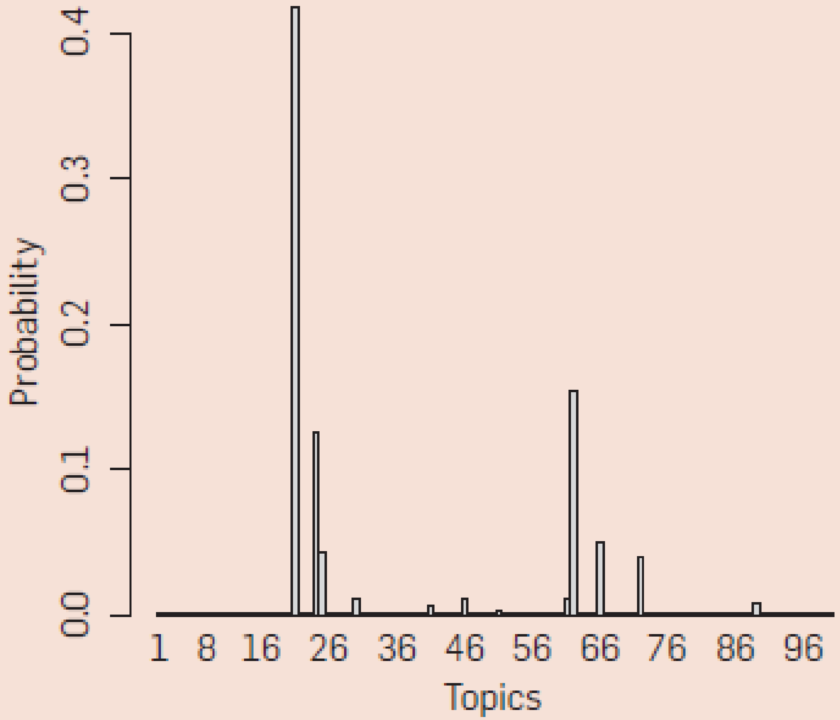


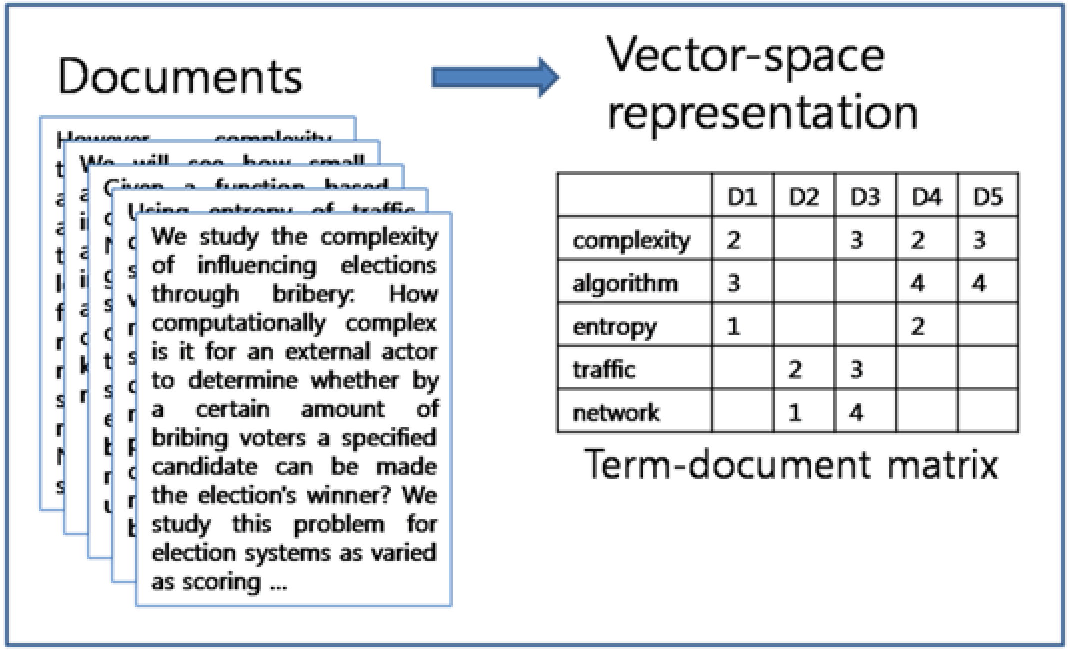
Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Tuning parameters

- K (number of topics): https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html
- Alpha, Beta (sparsity of topic model): a low alpha value places more weight on having each document composed of only a few dominant topics (whereas a high value will return many more relatively dominant topics). Similarly, a low beta value places more weight on having each topic composed of only a few dominant words.



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

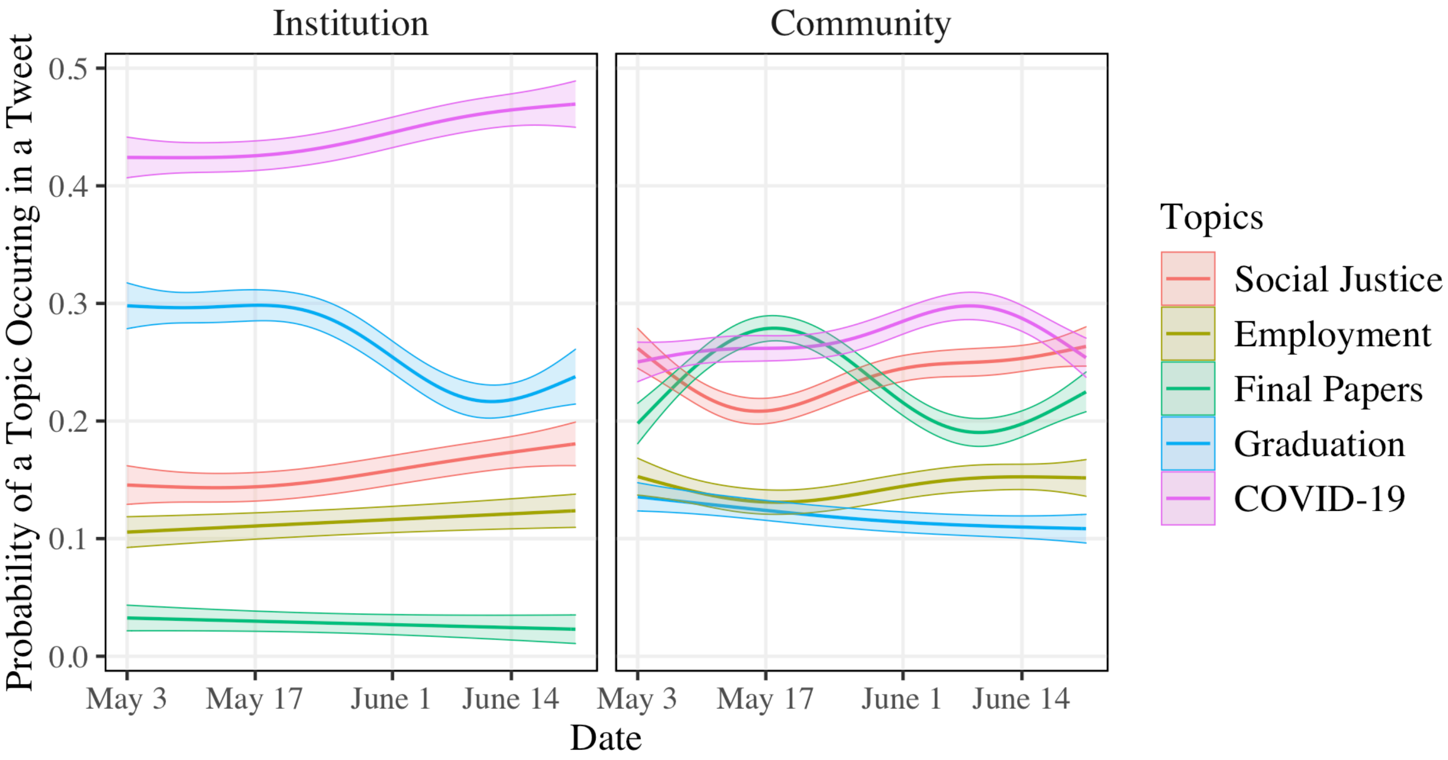
Pre-processing before the TD Matrix

1. Remove stop & other common words
2. STEM words

From: DSA, Hyderabad https://www.quora.com/profile/Data-Science-Authority-1

For our project, our documents are individual tweets, but this can also represent sociolinguistic interviews, responses to open-ended survey questions, etc. The counts in each cell represent the number of times a word appears in that document. Terms are the rows (i.e. unique lexical items in the corpus) and documents are the columns.

(1) [Employment]: **"Permanent employment opportunity for #UAlbertaENG alumni from Computer Engineering - Data Engineer - PlacePro# 56893 - Closing July 11"** (2020-06-12)
(2) [Final Projects/Papers]: **"all my friends r getting their finals canceled but i have 4 lesson of the day is don't study engineering"** (2020-06-07)
(3) [Social Justice] : **"black lives matter in STEM. we need more black scientists, engineers, computer programmers, and researchers. i want to dedicate my life as an educator to help uplift marginalized communities, especially my black students."** (2020-06-04)



Topic Model of the Engineering Education Community versus Engineering Institution's tweets between May 3, 2020 and June 15, 2020

Example Code for Topic Modeling

Modeling Meaning with Tensorflow

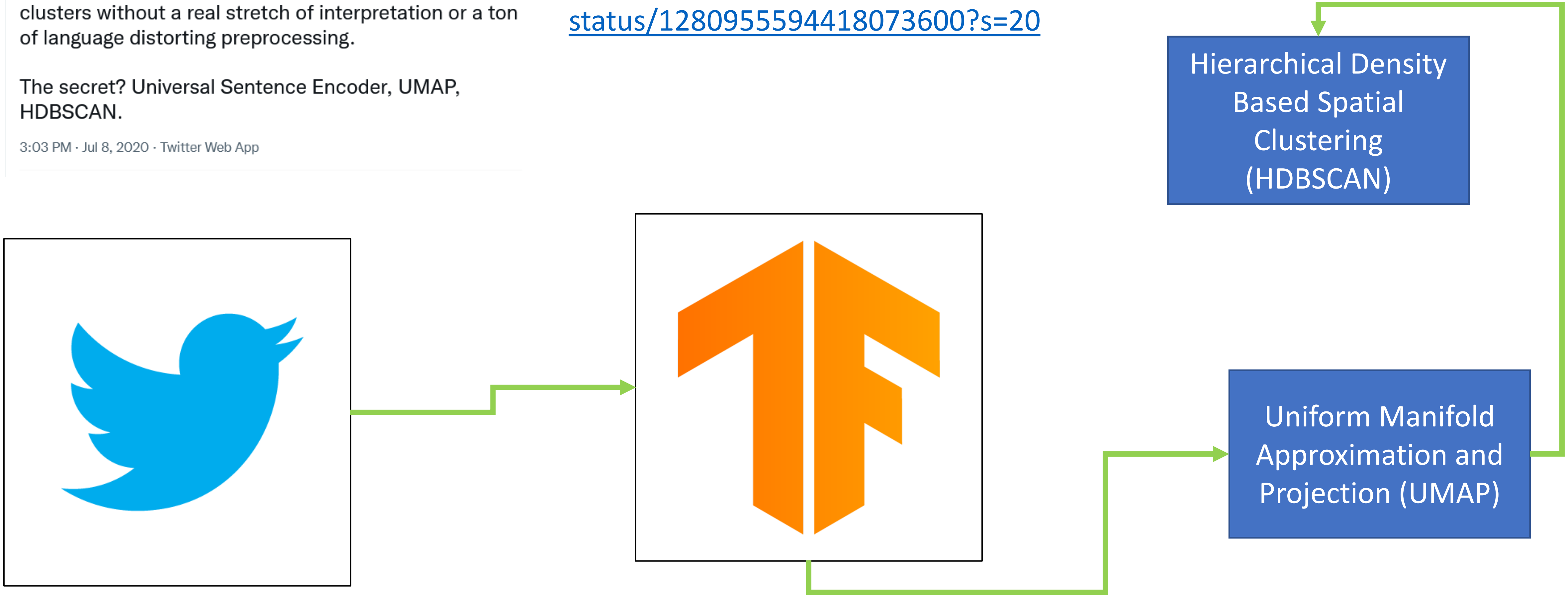
Peter Baumgartner
@pmbaumgartner

Today marks the first time I've seen unsupervised learning actually work on text and give meaningful clusters without a real stretch of interpretation or a ton of language distorting preprocessing.

The secret? Universal Sentence Encoder, UMAP, HDBSCAN.

3:03 PM · Jul 8, 2020 · Twitter Web App

Thread here:
https://twitter.com/pmbaumgartner/status/1280955594418073600?s=20



(1) Universal Sentence Encoder

- Word Embeddings
  - Hu, H., Amaral, P., & Kübler, S. (2021). Word embeddings and semantic shifts in historical Spanish: Methodological considerations. *Digital Scholarship in the Humanities*.
- Sentence embedding expands this idea to setences (https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder)
- Multilingual Support (cross-linguistic comparison):
- https://www.tensorflow.org/hub/tutorials/cross_lingual_similarity_with_tf_hub_multilingual_universal_encoder

Words as Vectors

Word Embedding Algorithms: Takes a corpus input and produces a word embedding for each lexical item (a vector of 100+ length). Words that share common contexts are closest in the vector space.

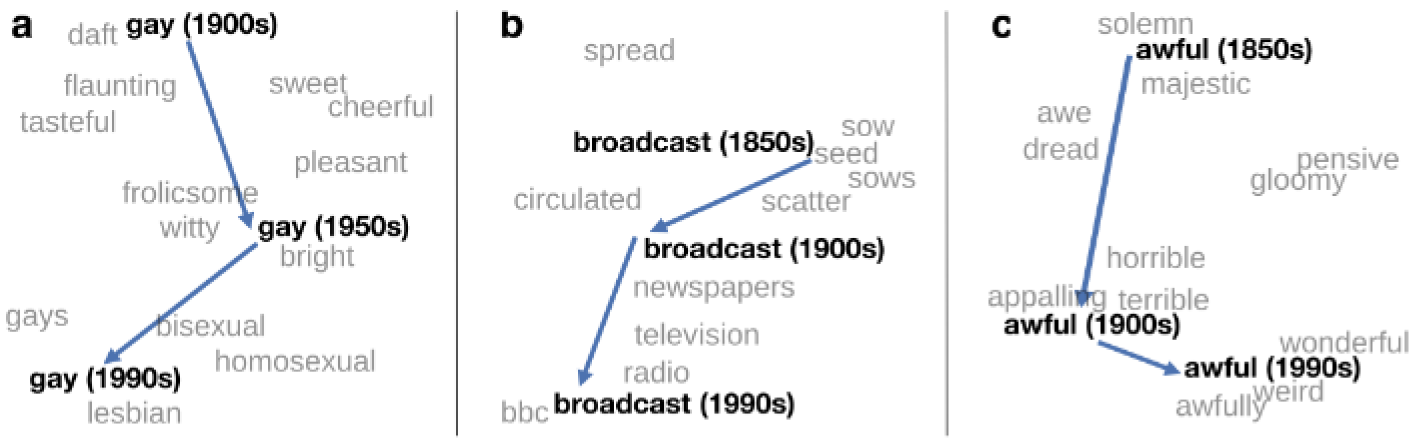


Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

From: Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change." *arXiv preprint arXiv:1605.09096* (2016).

Sentences as Vectors

- Similarly, models can be built to map sentences to a vector representing context or meaning.
- In TensorFlow: the Universal Sentence Encoder
  - https://www.tensorflow.org/install/
  - https://tensorflow.rstudio.com/
- Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant et al. "Universal sentence encoder for English." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169-174. 2018.

Example Code for USE + UMAP + HDBSCAN

What value do these techniques have for sociolinguistic studies?

- What value does sociolinguistic knowledge have to these automated process of language on which these AI based language models are built?
  - Racial Bias in Automated Speech Recognition:
    - Wassink, Alicia Beckford. "Uneven Success: Automatic Speech Recognition and Ethnicity-related Dialects." In *2020 Annual Meeting*. AAAS, 2020.
    - Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. "Racial disparities in automated speech recognition." *Proceedings of the National Academy of Sciences* 117, no. 14 (2020): 7684-7689.
  - Social Media: Blodgett, Su Lin, Lisa Green, and Brendan O'Connor. "Demographic Dialectal Variation in Social Media: A Case Study of African-American English." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119-1130. 2016.
  - Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610-623. 2021.

Selected References

Bail, Christopher A. "The cultural environment: Measuring culture with big data." *Theory and Society* 43, no. 3-4 (2014): 465-482.

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, no. 6 (2013): 570-606.

Fligstein, Neil, Jonah S. Brundage, and Michael Schultz. "Why the Federal Reserve Failed to See the Financial Crisis of 2008: The Role of "Macroeconomics" as a Sense making and Cultural Frame." (2014).

Gentzkow, Matthew, and Jesse M. Shapiro. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78, no. 1 (2010): 35-71.

Grimmer, Justin. "Appropriators not position takers: The distorting effects of electoral incentives on congressional representation." *American Journal of Political Science* 57, no. 3 (2013): 624-642.

Hargittai, Eszter. "Is bigger always better? Potential biases of big data derived from social network sites." *The ANNALS of the American Academy of Political and Social Science* 659, no. 1 (2015): 63-76.

Ignatow, Gabe. "Theoretical foundations for digital text analysis." *Journal for the Theory of Social Behaviour* (2015).

Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers. "Quantitative analysis of large amounts of journalistic texts using topic modelling." *Digital Journalism* 4, no. 1 (2016): 89-106.

Jelveh, Zubin, Bruce Kogut, and Suresh Naidu. "Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics." In *EMNLP*, pp. 1804-1809. 2014.

Lee, Monica, and John Levi Martin. "Coding, counting and cultural cartography." American Journal of Cultural Sociology 3, no. 1 (2015): 1-33.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the dynamics of the news cycle." In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497-506. ACM, 2009.

Loughran, Tim, and Bill McDonald. "When is a Liability not a Liability?." *Journal of Finance, forthcoming* (2009).

Mohr, John W., and Petko Bogdanov. "Introduction—Topic models: What they are and why they matter." *Poetics* 41, no. 6 (2013): 545-569.

Mohr, John W., Robin Wagner-Pacifici, Ronald L. Breiger, and Petko Bogdanov. "Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics." *Poetics* 41, no. 6 (2013): 670-700.

Niculae, Vlad, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. "Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game." *arXiv preprint arXiv:1506.04744* (2015).

Resnik, Philip, Anderson Garron, and Rebecca Resnik. "Using topic modeling to improve prediction of neuroticism and depression." In *Proceedings of the 2013 Conference on Empirical Methods in Natural*, pp. 1348-1353. Association for Computational Linguistics}, 2013.

Saavedra, Serguei, Kathleen Hagerty, and Brian Uzzi. "Synchronicity, instant messaging, and performance among financial traders." *Proceedings of the National Academy of Sciences* 108, no. 13 (2011): 5296-5301.

Yu, Dian, Yulia Tyshchuk, Heng Ji, and W. A. Wallace. "Detecting deceptive groups using conversations and network analysis." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pp. 26-31. 2015.

Extra Topics (Not included in Workshop)

Advanced Topic Modeling

Structured, Hierarchical and Dynamic Models

THIS IS BIOSTAT
@THISISBIOSTAT

Follow

All models are wrong but some are accompanied with well-documented R packages so I dunno just use those I guess. The ones with R packages.

3:01 PM - 5 Jul 2017

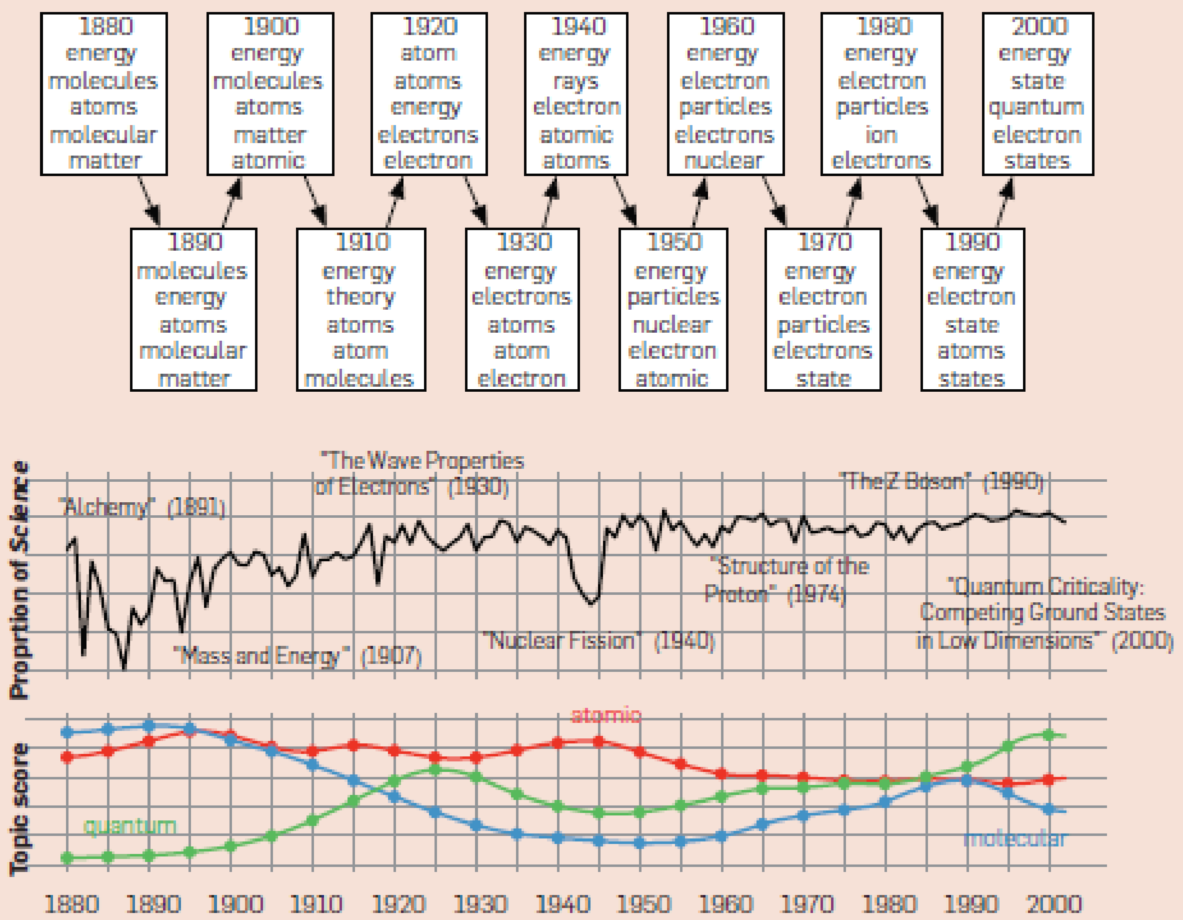
111 Retweets **248** Likes

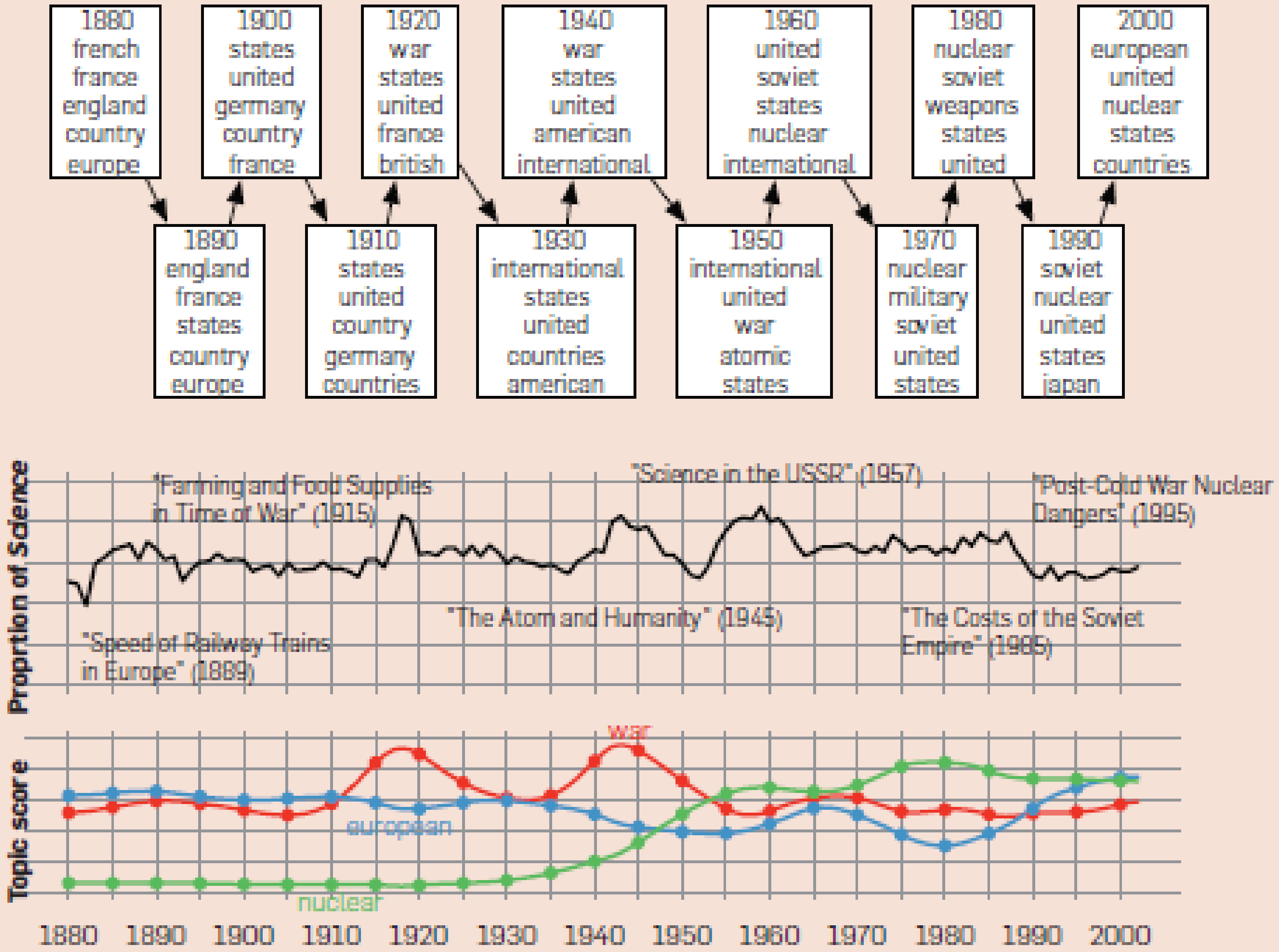
111 248

Original quote: “All models are wrong, some are useful.”
George Box

Dynamic Topic Modeling (i.e. Time)

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.




1880 french france england country europe
1890 england france states country europe
1900 states united germany country france
1910 states united country germany countries
1920 war states united france british
1930 international states united countries american
1940 war states united american international
1950 international united war atomic states
1960 united soviet states nuclear international
1970 nuclear military soviet united states
1980 nuclear soviet weapons states united
1990 soviet nuclear united states japan
2000 european united nuclear states countries
Proprtion of Science
"Farming and Food Supplies in Time of War" (1915)
"Science in the USSR" (1957)
"Post-Cold War Nuclear Dangers" (1995)
"The Atom and Humanity" (1945)
"The Costs of the Soviet Empire" (1985)
"Speed of Railway Trains in Europe" (1889)
Topic score
war
european
nuclear
1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000

Hierarchical Topic Modeling

- Topics are in a hierarchy of topics
- E.g. food -> {vegetables, meat, dairy}
  - Dairy → {cheese, cream}
- Python implementation.

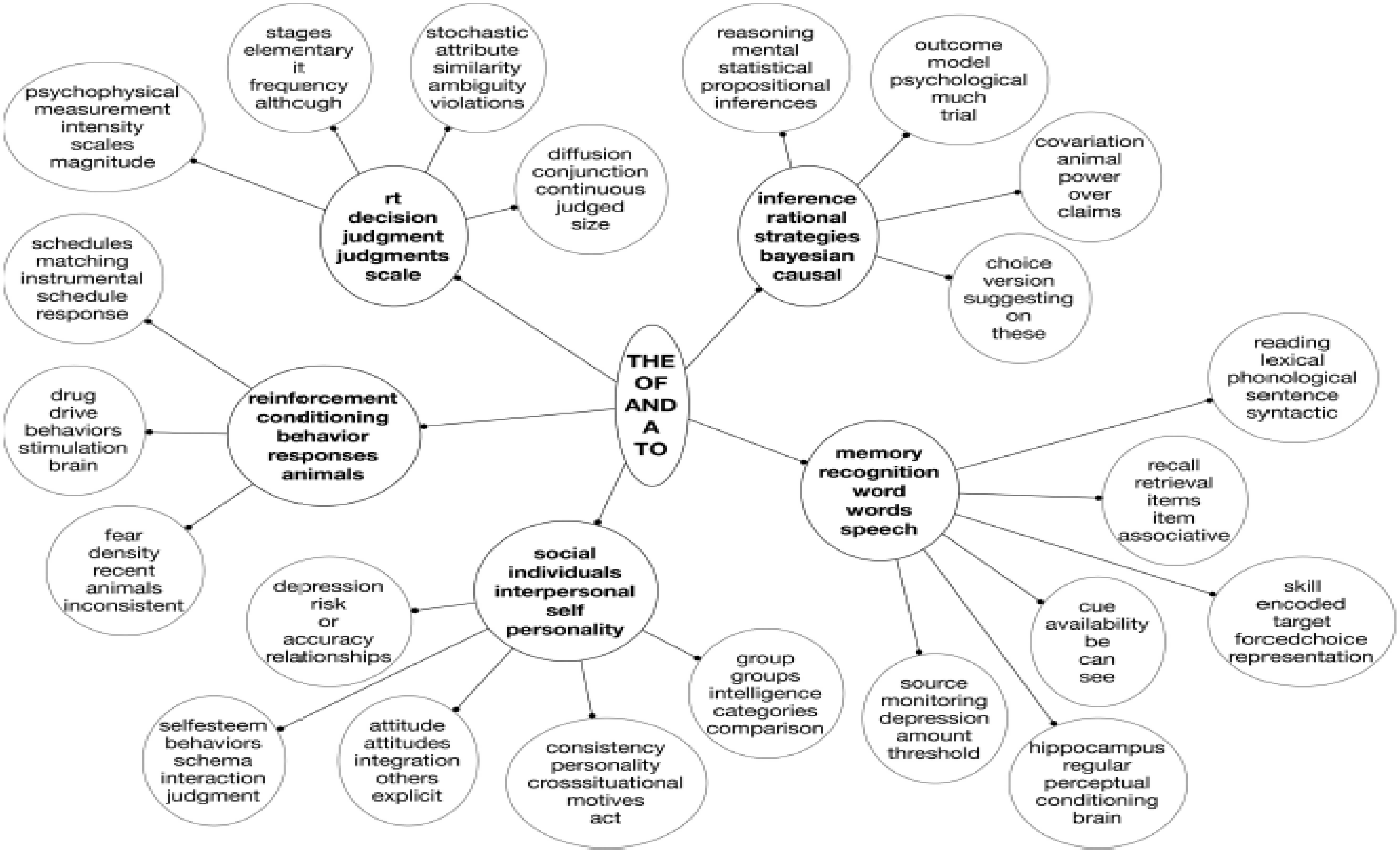


FIG. 7. A portion of the hierarchy learned from the 1,272 abstracts of *Psychological Review* from 1967–2003. The vocabulary was restricted to the 1,971 terms that occurred in more than five documents, yielding a corpus of 136K words. The learned hierarchy, of which only a portion is illustrated, contains 52 topics.

Structured Topic Modeling

- Topics are conditional on predictors
  - Historical Linguistics: Social Characteristics, Time, Variants.
- Topic is now dependent on both the words in a document and the features associated with a document.

Word2Vec

Word Embedding Models – taking into account context.

So Far

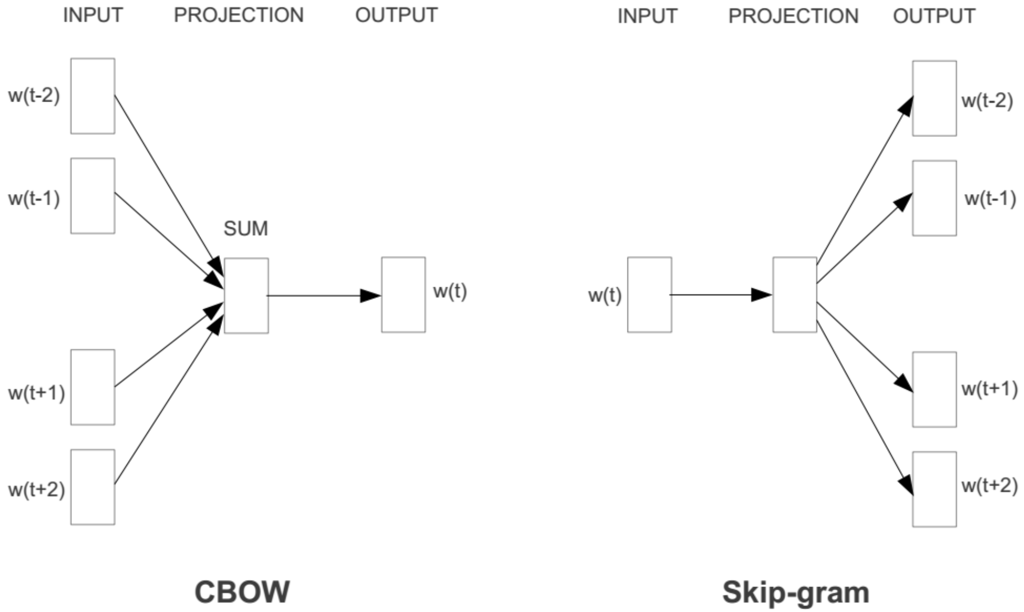
- Topic modeling requires a bag of words approach.
- Syntactic Topic Modeling is a possibility (Blei has some published work)
  - Problems: Too computationally intensive – requires parsing data as a pre-processing step.
  - No widely available implementation (that I know of…)

word2vec Approach to represent the meaning of word

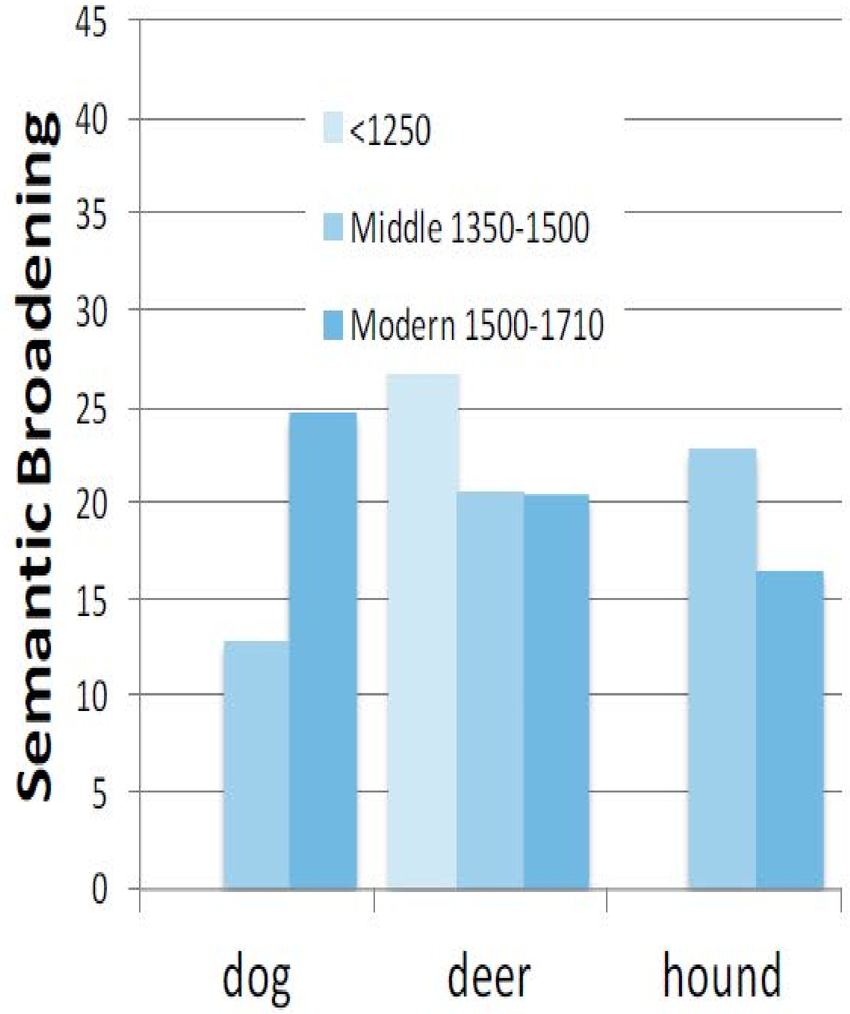
- Represent each word with a low-dimensional vector
- Word similarity = vector similarity
- Key idea: Predict surrounding words of every word
- Faster and can easily incorporate a new sentence/document or add a word to the vocabulary
- Allows context (i.e. surrounding words) to matter in output.

Represent the meaning of **word** – word2vec

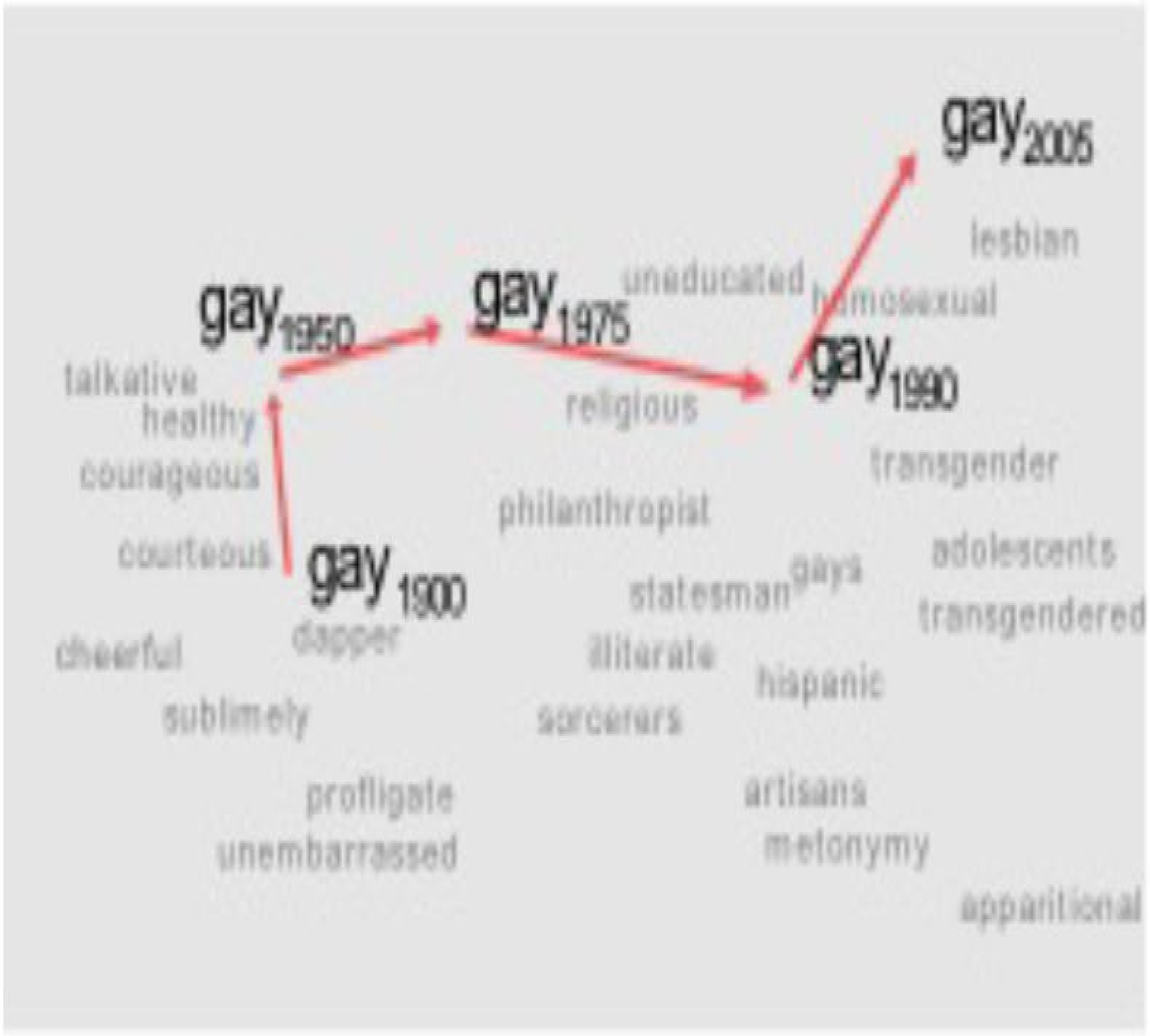
- 2 basic neural network models:
  - Continuous Bag of Word (CBOW): use a window of word to predict the middle word
  - Skip-gram (SG): use a word to predict the surrounding ones in window.



Sagi, Kaufmann Clark 2013



Kulkarni, Al-Rfou, Perozzi, Skiena 2015



Some interesting results

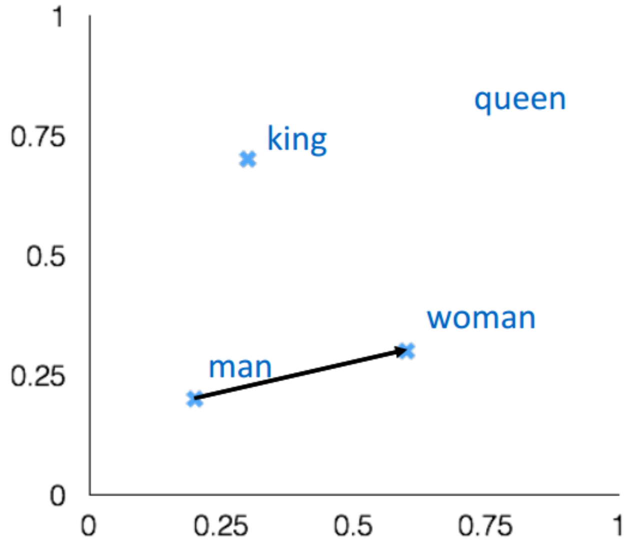
Word Analogies

Test for linear relationships, examined by Mikolov et al. (2014)

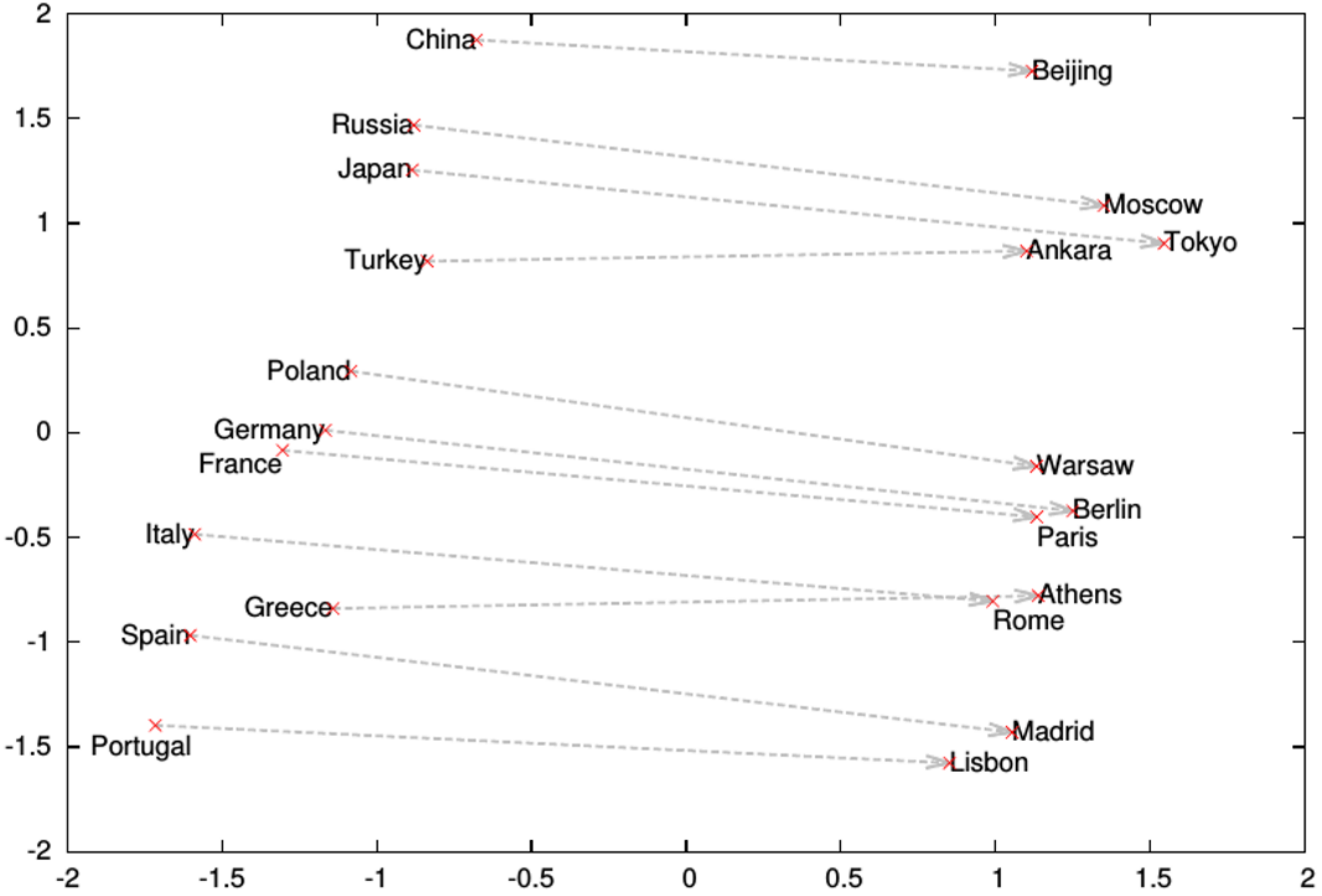
a:b :: c:? → $$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]

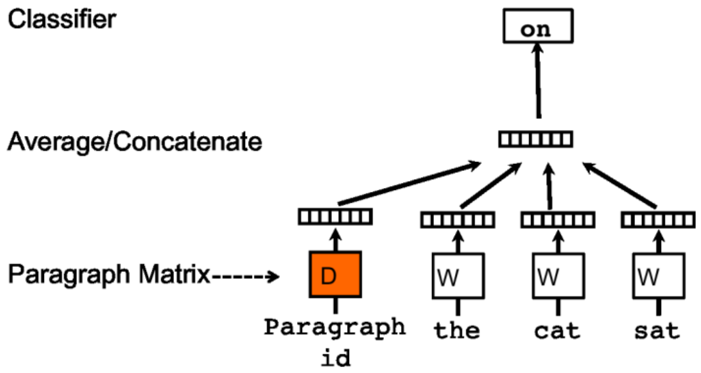


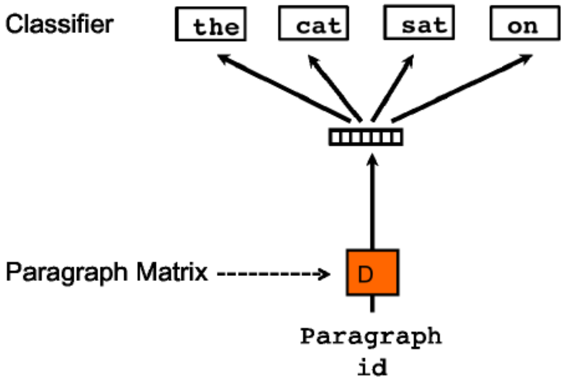
Word analogies



Represent the meaning of **sentence/text**

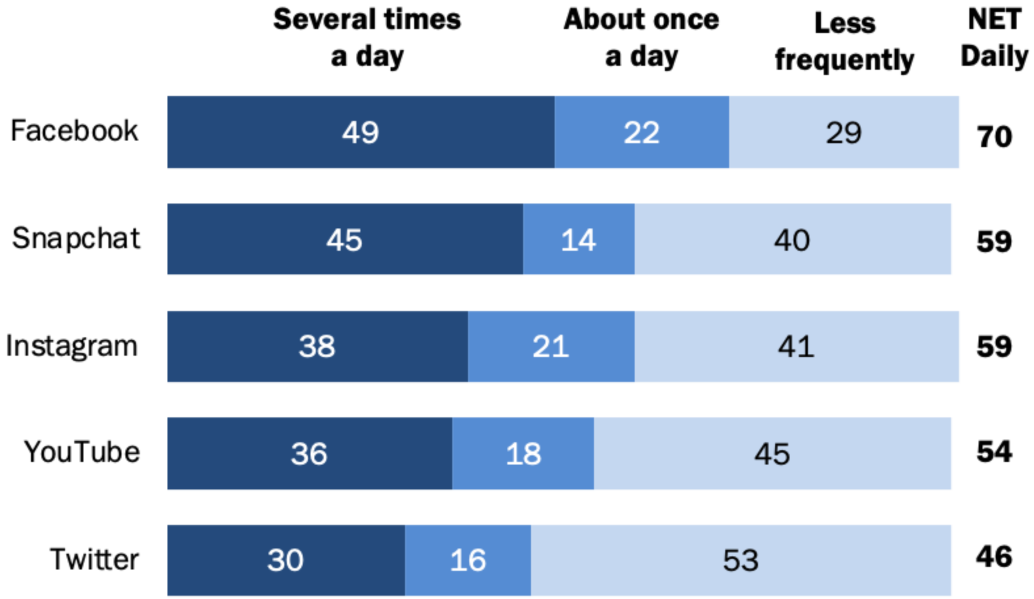
- Paragraph vector (2014, Quoc Le, Mikolov)
  - Extend word2vec to text level
  - Also two models: add paragraph vector as the input





Seven-in-ten Facebook users say they visit site daily

Among U.S. adults who say they use ___, % who use that site …



Note: Respondents who did not give an answer are not shown. “Less frequently” category includes users who visit these sites a few times a week, every few weeks or less often.
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.
“Social Media Use in 2021”

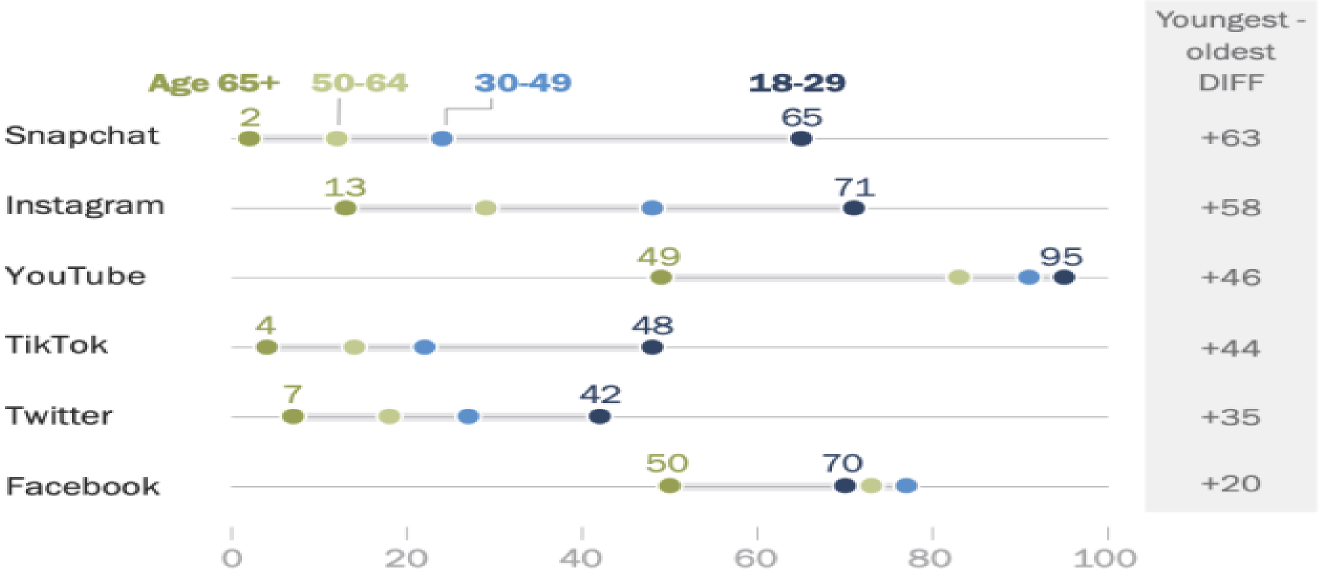
PEW RESEARCH CENTER

Twitter has the least daily engagement (only 46%) of major social media platforms.

Social Media Population

Age gaps in Snapchat, Instagram use are particularly wide, less so for Facebook

% of U.S. adults in each age group who say they ever use ...



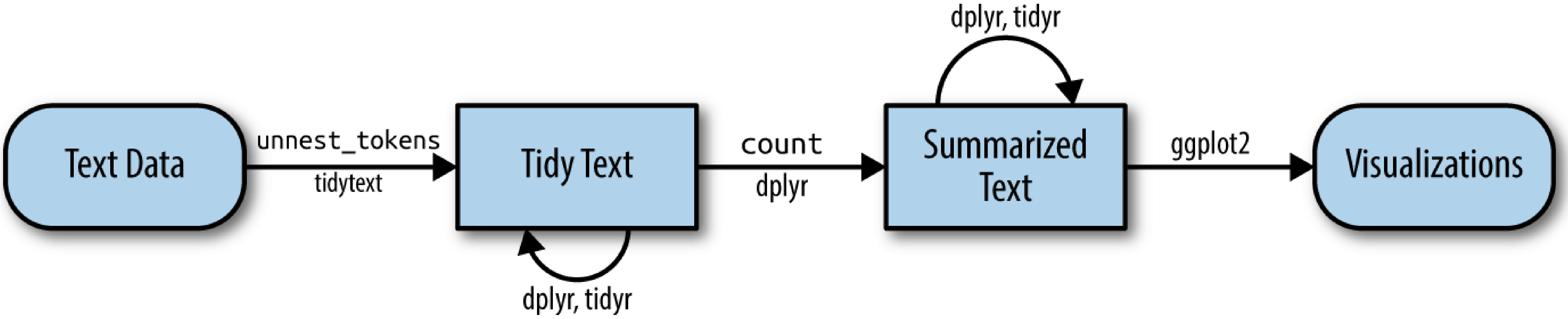
Note: All differences shown in DIFF column are statistically significant. The DIFF values shown are based on subtracting the rounded values in the chart. Respondents who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.
"Social Media Use in 2021"

PEW RESEARCH CENTER

Each type of social media platform is used by different demographic groups.

Any data drawn from any from a social media platform will be biased (statistically) by its user base. This is also true for non-randomized interview techniques employed in sociolinguistics (e.g. recruiting from friend-of-friends, or a central location introduces statistical biases into the sample).

Twitter is used most by 18-29 cohort, but less than other platforms.



From Text Mining with R: a Tidy Approach, by Julia Silge and David Robinson.
https://www.tidytextmining.com/tidytext.html

Code + example data available online:
https://github.com/joe-roy/nwav49_workshop

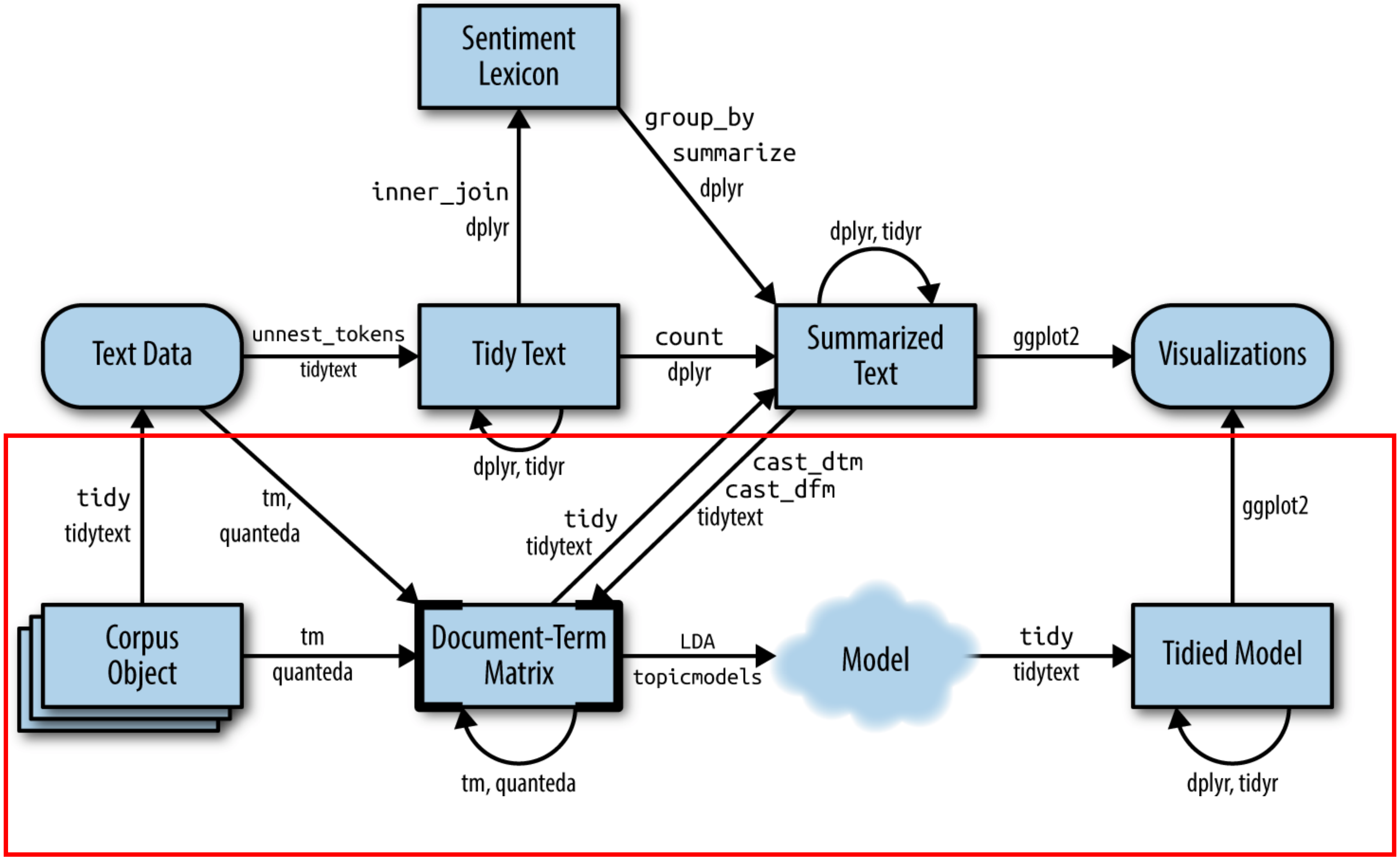


Figure 6.1, Text Mining with R: a tidy approach. Julia Silge and David Robinson. https://www.tidytextmining.com/index.html