Text as data

Amelia McNamara

flickr: Monika Hoinkis

Syllabus for STAT 490: Categor

https://stthomas.instructure.com/courses/53862/assignments/syllabus

Account

Dashboard

Courses

Groups

Calendar

Inbox

History

Commons

Search

10

Help

Categorical & Text Analysis › Syllabus

Search this course

Student View

Spring 2022

- Home
- Syllabus
- Announcements
- Discussions
- Modules
- Grades
- People
- Attendance
- Zoom Pro
- Library Help
- Search
- New Analytics
- Accessibility Report
- GitHub
- Files
- Assignments
- Quizzes
- Pages
- Outcomes
- Collaborations
- Rubrics
- Settings

Course Syllabus

STAT 490 (Topics: Categorical and Text Analysis) Syllabus

About the Course

Instructor

- Dr. Amelia McNamara (`amelia.mcnamara@stthomas.edu`, 651-962-5391).
- Student hours: Mondays 4-5 pm (Zoom), Fridays 8-9 am (Zoom), and by appointment (Calendly).

Course format

This course is in person. For the first few weeks of the course, there will also be a hybrid Zoom option. If anything changes regarding the format of the course, I'll keep you informed through course announcements.

Covid-19 circumstances: At St Thomas, we are committed to a culture of care for all. If you are spending time on campus, you are expected to abide by the campus preparedness plan. This includes wearing a mask in all common spaces (including our classroom) and maintaining a 6-foot distance from others. If you feel sick, please stay home and plan to attend via Zoom.

Course Description

Much of the data we focus on in statistics is quantitative, but there are rich methods for working with non-quantitative data as well. This course will cover elements of categorical data analysis (polytomous and ordinal logistic regression, visualization of Likert-scale data), as well as text analysis (including sentiment analysis and topic modeling). The course will emphasize reproducible research methods and include a strong computing component. Both R and Python will be employed as programming tools.

Prerequisites:

Prerequisites: Grades C- or higher in STAT 320 or 333, CISC 130 or 131.

Course Goals

- Gain a basic understanding of the field of text analysis
- Learn to apply models to non-quantitative data
- Develop skills in working with non-quantitative data in R and Python

Textbooks

There is no required textbook for this course. We will be referencing a variety of texts, including:

- R for Data Science, Garrett Grolemund and Hadley Wickham.
- Python Data Science Handbook, Jake VanderPlas.
- Advanced R, Hadley Wickham.

World Cafe

This semester our class will be participating in the World Café, hosted by Justice and Peace Studies. The World Café at UST is an interdisciplinary dialogue opportunity in which hundreds of students and faculty from across the university participate to analyze and offer perspectives on a critical social issue. Over the years, UST has hosted World Cafes on a range of topics from the HIV/AIDS pandemic to climate change to gun violence, to name a few.

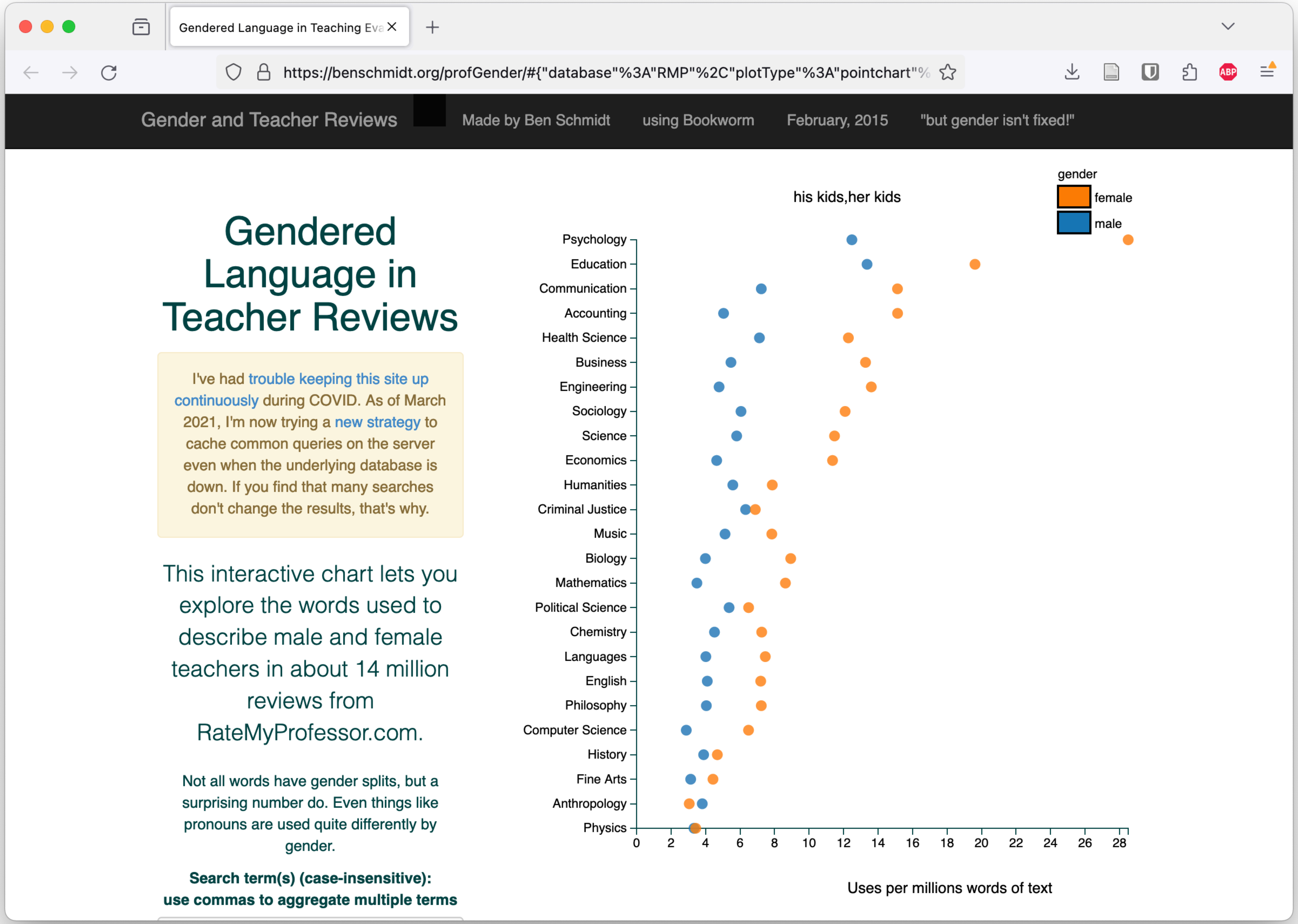
This year's focus is on settler colonialism and indigenous rights, building off the recent work of UST's land acknowledgment committee. As participants in this year's World Café, our class will attend a joint dialogue on the evening of **April 20 from 6-8p** with all the other participating classes and also do some pre-work ahead of time, including hearing from a guest faculty member and completing a set of common readings. The aim of this year's World Café is for us to really dig in and think about what we understand and believe about the realities of settler colonialism on the land in which we now live, study, work, play… and what this might lead us to do today and going forward. The purpose of our dialogue is not necessarily to come to agreement but rather to inquire to learn and to deepen understandings by listening with curiosity and by articulating one's own beliefs/learnings.

Please put the joint dialogue event on your calendar: Wednesday, April 20th from 6:00 – 8:00 pm. Because I am requiring attendance at this event, we will cancel a regularly scheduled class session at some point.

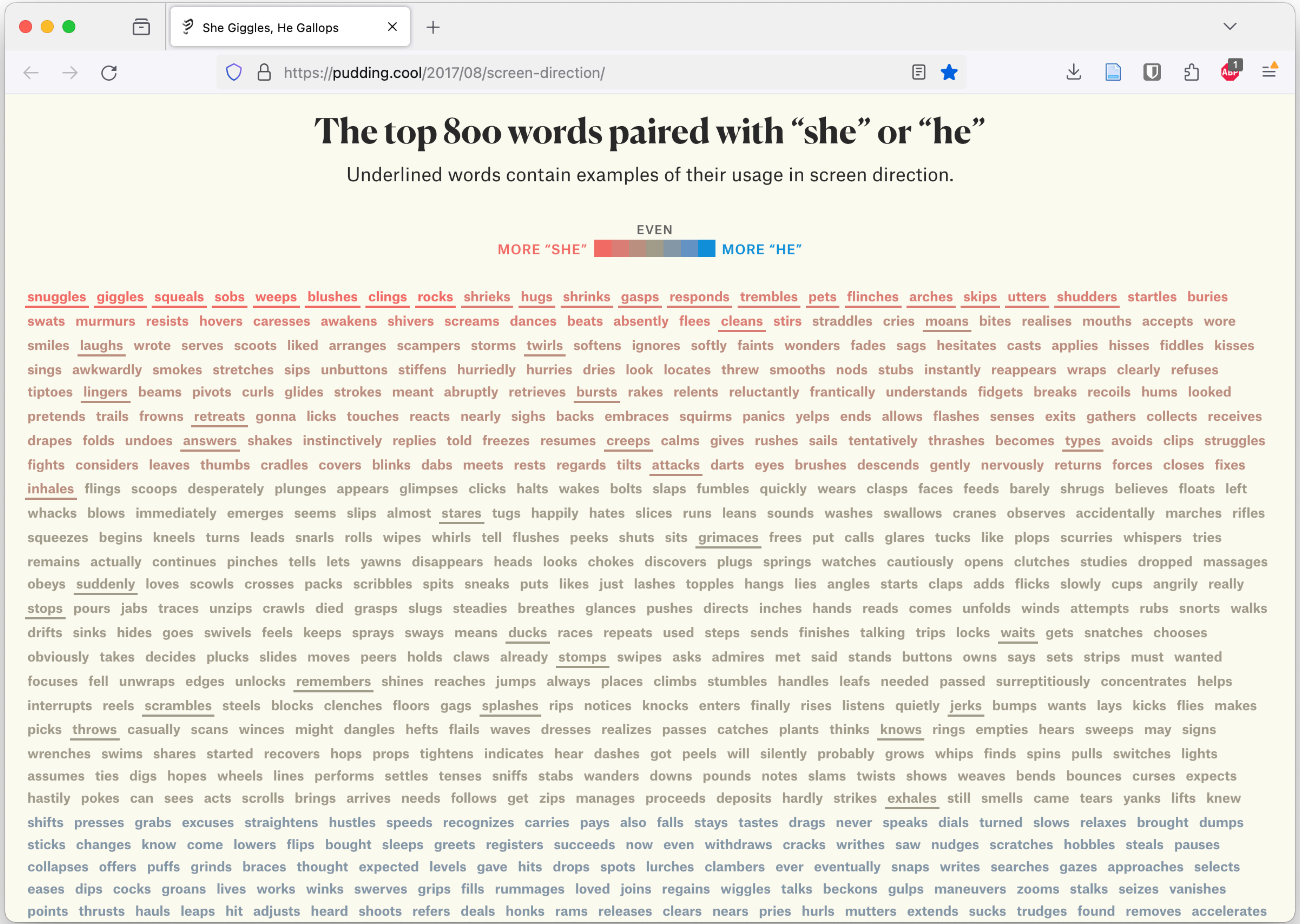
Tentative Schedule

The following is a brief outline of the course. Please refer to the course modules for more detailed information.

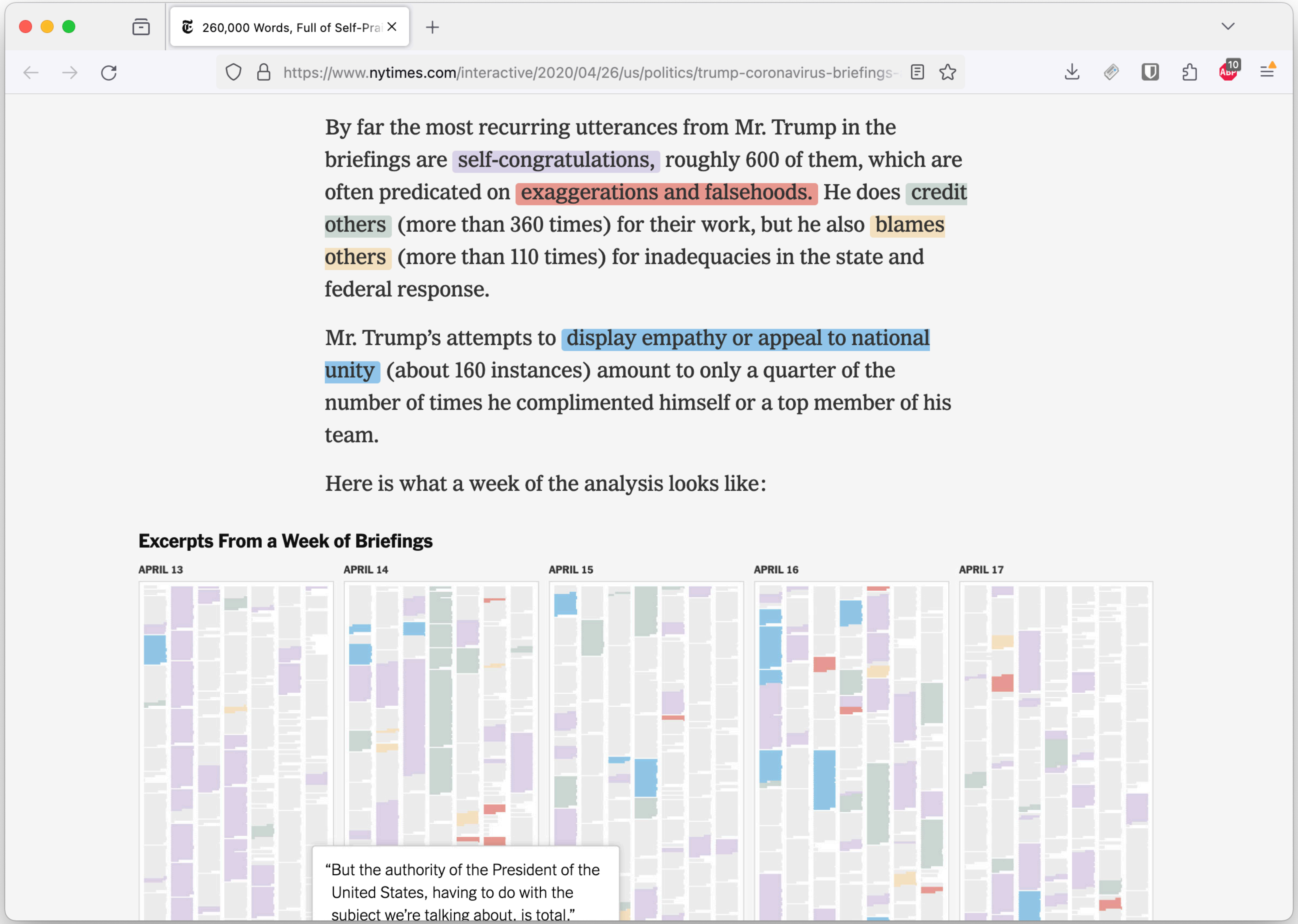
Week	Topic	Assignment(s)
1	Intro to class and non-quantitative data	
2	Review of data wrangling in R, intro to git and GitHub	Data wrangling mini-project
3	Review of logistic regression, intro to more complex versions	
4	Tidy and untidy data	Regression mini-project
5	Intro to text analysis	
6	Ethics and scraping	R text analysis mini-project
7	More complex text analysis in R	**Exam 1**
8	**Spring break**	
9	Intro to data wrangling in Python	Initial project proposal
10	Text analysis in Python	Revised project proposal
11	Other (human) languages	Python text analysis mini-project
12	**World Cafe**	World Cafe reflection
13	TBD	First draft and peer review
14	TBD	
15		Second draft, **Exam 2**



Gendered Language in Teacher Reviews. Ben Schmidt. https://benschmidt.org/profGender



She Giggles, He Gallops. Julia Silge https://pudding.cool/2017/08/screen-direction/



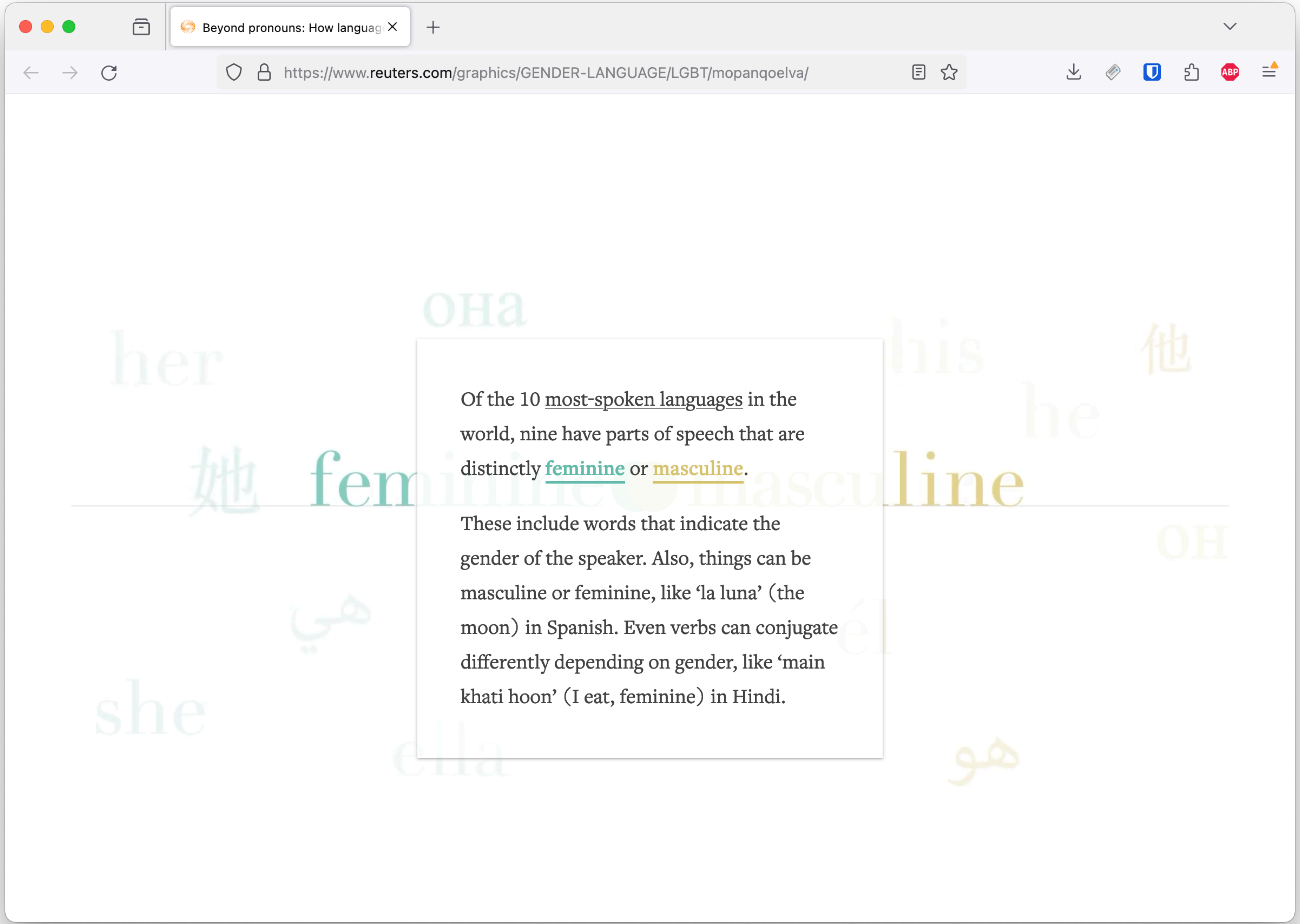
By far the most recurring utterances from Mr. Trump in the briefings are self-congratulations, roughly 600 of them, which are often predicated on exaggerations and falsehoods. He does credit others (more than 360 times) for their work, but he also blames others (more than 110 times) for inadequacies in the state and federal response.

Mr. Trump's attempts to display empathy or appeal to national unity (about 160 instances) amount to only a quarter of the number of times he complimented himself or a top member of his team.

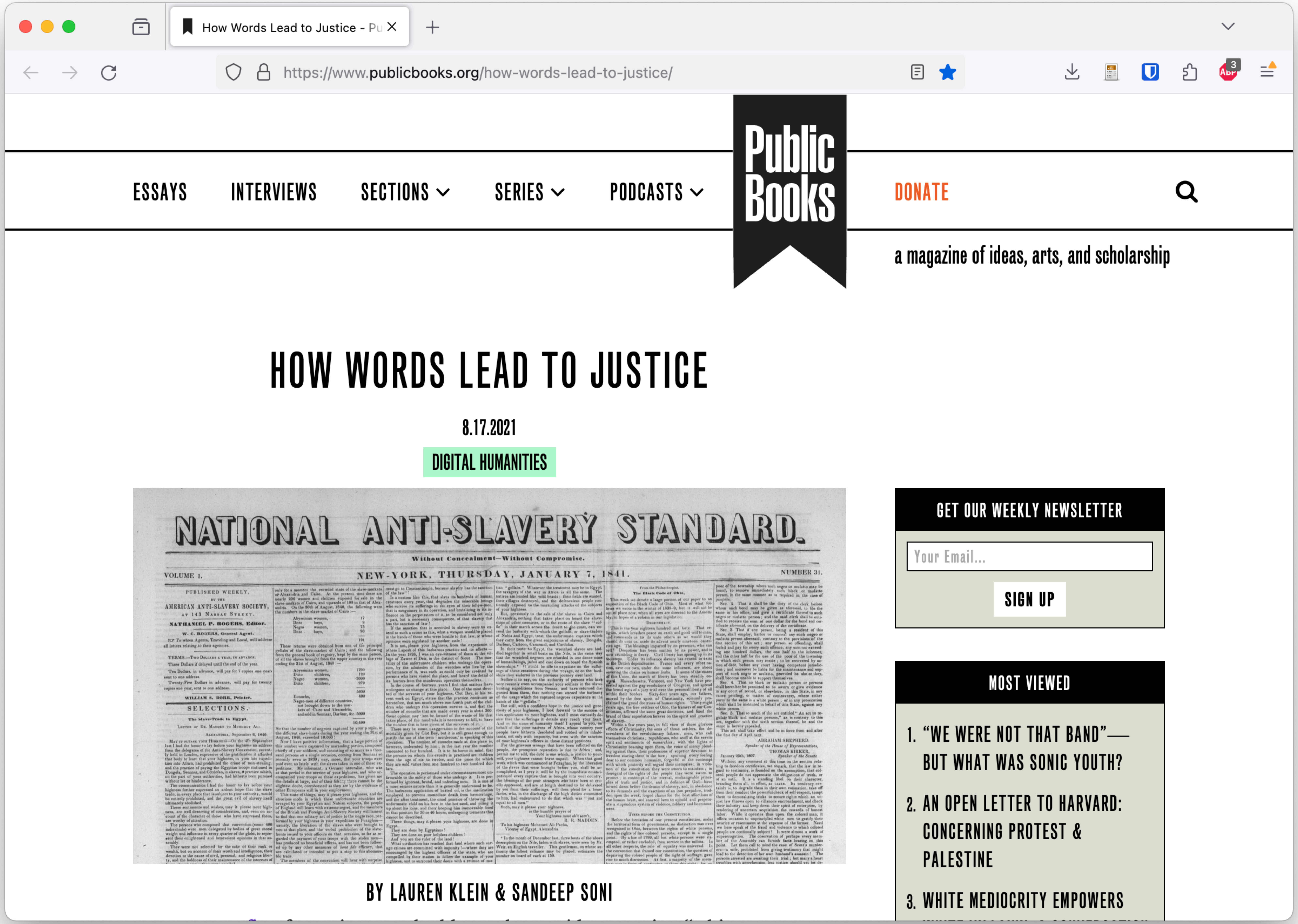
Here is what a week of the analysis looks like:

Excerpts From a Week of Briefings

260,000 Words, Full of Self-Praise, From Trump on the Virus. Jeremy W. Peters, Elaina Plott and Maggie Haberman
https://www.nytimes.com/interactive/2020/04/26/us/politics/trump-coronavirus-briefings-analyzed.html



Beyond Pronouns: How languages are reshaping to include nonbinary and gender-nonconforming people
Minami Funakoshi https://www.reuters.com/graphics/GENDER-LANGUAGE/LGBT/mopanqoelva/



How Words Lead to Justice. Lauren Klein & Sandeep Soni
https://www.publicbooks.org/how-words-lead-to-justice/

Copyright

- Copyrights no longer need to be registered to have legal effect; works just need to be "original works of authorship fixed in any tangible medium of expression" i.e. print or online text for written works, but also other media. [17 U.S. Code § 102].
- After Copyright protection ends, works are said to enter the "public domain", which just means that anyone can copy or use them for their own purposes.
- Many of the sources "safest" to use in terms of copyright risk are works in the public domain, findable in sources like Project Gutenberg, Hathi Trust, etc.

Libraries | University of St. Thomas

Library resources for STAT 490, John Heintz
UMN site https://www.lib.umn.edu/services/copyright/use

Fair Use: The Four Factor test

Excerpted from U of MN copyright site

- **Factor 1: purpose and character of the use**
  Purposes that favor fair use include education, scholarship, research, and news reporting, as well as criticism and commentary more generally. Non-profit purposes also favor fair use (especially when coupled with one of the other favored purposes.) Commercial or for-profit purposes weigh against fair use.

- **Factor 2: nature of the original work**
  Published or not: Using published material is more likely to be fair use, and using unpublished material is less likely to be fair use.
  "Factual" or "creative": Using a factual work is more likely to be fair use, using a creative work is less likely to be fair use. This is related to the fact that copyright does not protect facts and data.

Library resources for STAT 490, John Heintz
UMN site https://www.lib.umn.edu/services/copyright/use

Fair Use: The Four Factor test

Excerpted from U of MN copyright site

- **Factor 3: amount and substantiality of the portion used**
  amount: Using a smaller amount of the source work is more likely to be fair use, and using a larger amount is less likely to be fair use. But courts have been very clear that "amount" here is proportional.
  Substantiality: It is less likely to be fair use to use central parts of the work, and more likely to be fair use if you use a more peripheral part of the work.
- **Factor 4: effect of the use on the potential market for, or value of, the source work**
  is the use in question substituting for a sale the source's owner would otherwise make - either to the person making the proposed use, or to others?

Library resources for STAT 490, John Heintz
UMN site https://www.lib.umn.edu/services/copyright/use

Newer interpretations: “transformative” uses

Excerpted from U of MN copyright site

- Raised in Supreme Court decision (Campbell v. Acuff-Rose Music, 510 U.S. 569 (1994.)
- A new work based on an old one work is transformative if it uses the source work in completely new or unexpected ways. Importantly, a work may be transformative, and thus a fair use, even when all four of the statutory factors would traditionally weigh against fair use!
- Examples:
  - Parody
  - Criticism/commentary
  - New technologies: search engine copies, Google Books

Library resources for STAT 490, John Heintz
UMN site https://www.lib.umn.edu/services/copyright/use

GIDA
Global Indigenous
Data Alliance

Indigenous data sovereignty

https://www.gida-global.org/

Finding text data

- Project Gutenberg (Agatha Christi novels, Winnie the Pooh, many more)
- Hathi Trust
- Things you can copy-paste, access using APIs, or scrape (remember laws/ethics!)
-

Tokenization, stop words, stemming

The first step in most text analysis projects is called tokenization. If you have a long string of text, you need to turn it in to “tokens” that can be analyzed. The most common token is a word, but people also analyze bigrams (two-word pairs) and n-grams (a generalization of the same idea, n words in a group).

Once you have tokens, you can study things like the most common tokens.

…they will usually be things like “the” “of” and other common English words. So, we often have lists of “stop words” that we remove from the list.

…after that, you might notice that “cook” “cooked” “cooking” and “cooks” are all being counted separately, but they capture the same concept. So you can “stem” the words by removing common suffixes. (Another alternative is called lemmatization.)

“Bag of words” model

- The simplest way to analyze data is with the “bag of words” model. This thinks of words as individual units, without much consideration for context.
- But of course, context is really important!

Word clouds

A word cloud shows the frequency of words based on the size of the word. Mapping: size is the number of times it occurred. X and Y positions hold no meaning.

Often you will remove “stop words,” common words you are not that interested in. (Think prepositions and articles, like “the,” “and,” “of,” etc)

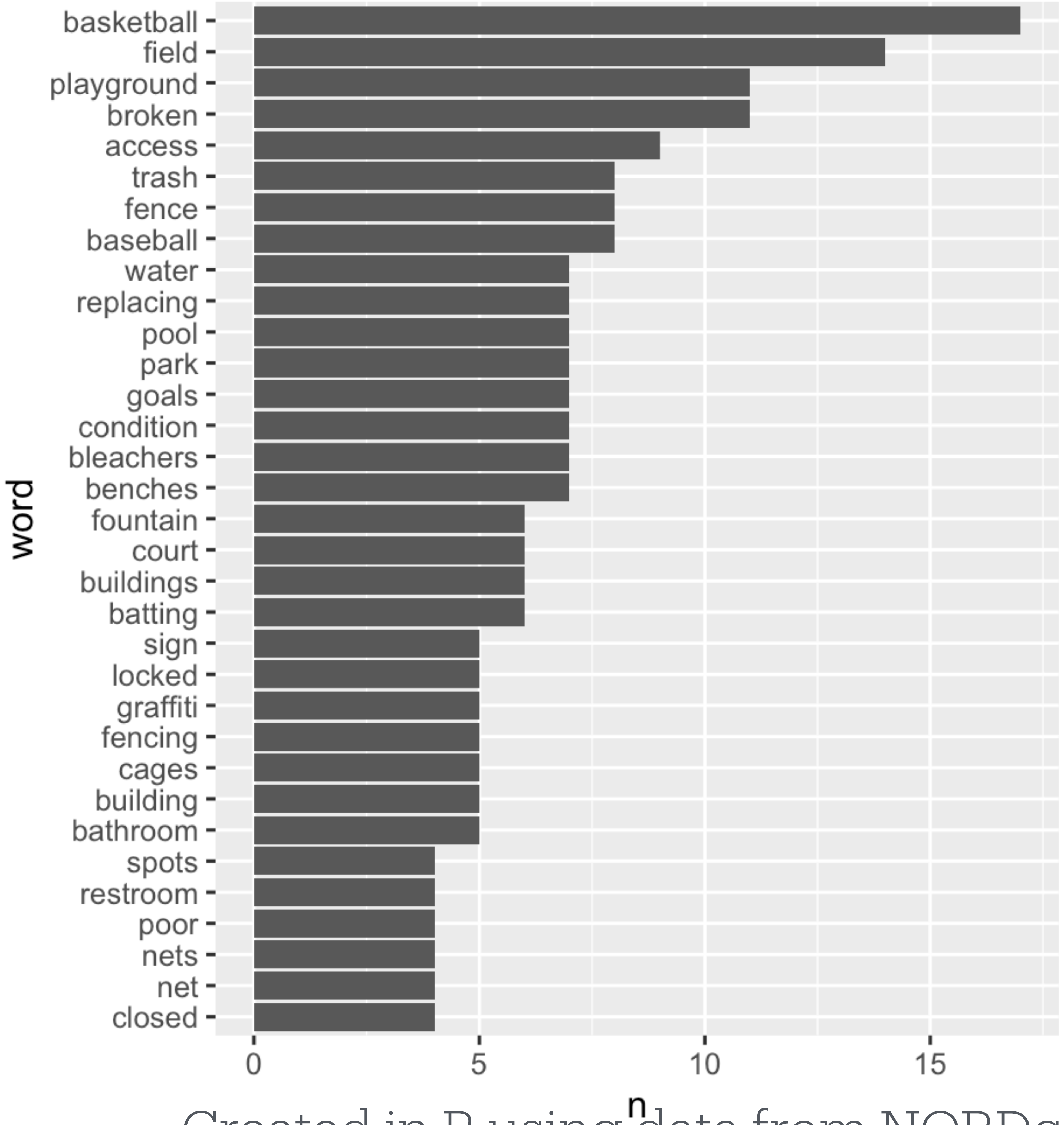
One problem with word clouds is they lose context, since they use the bag of words model. If someone is saying “not good” those words get separated and meaning can be lost.

Created in Voyant Tools using data from NORDc Facilities

Statisticians would prefer a barchart

There's a strand of the data viz world that argues that everything could be a bar chart. That's possibly true but also possibly a world without joy.
-Amanda Cox

Nathan Yau, *Data Points*. 2013



Created in R using data from NORDc Facilities (different stop word list!)

Adding context… with math

Vectors

Question 1. What space does each vector "live" (exist) in?

$$\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\0\\1\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}1\\1\\0\\0\end{pmatrix}$$

1-dimensional space? 2-dimensional space? 3-dimensional space?

How could we visualize or imagine these vectors?

From Chad Topaz's "An Ill-Advised Linear Algebra Tutorial"

Matrices

<u>Example.</u> $\underline{\underline{A}} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$ is a 3 x 2 matrix. We can conceive of it as two vectors, each of which lives in $\mathbb{R}^3$.

<u>Question 2</u>. Describe the sizes of these matrices. Be creative! ❤️

$$\begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -7 \end{pmatrix}, \begin{pmatrix} 1 & 10 & 3 \\ -2 & -10 & 16 \\ -1 & 0 & -11 \\ 4 & 0 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

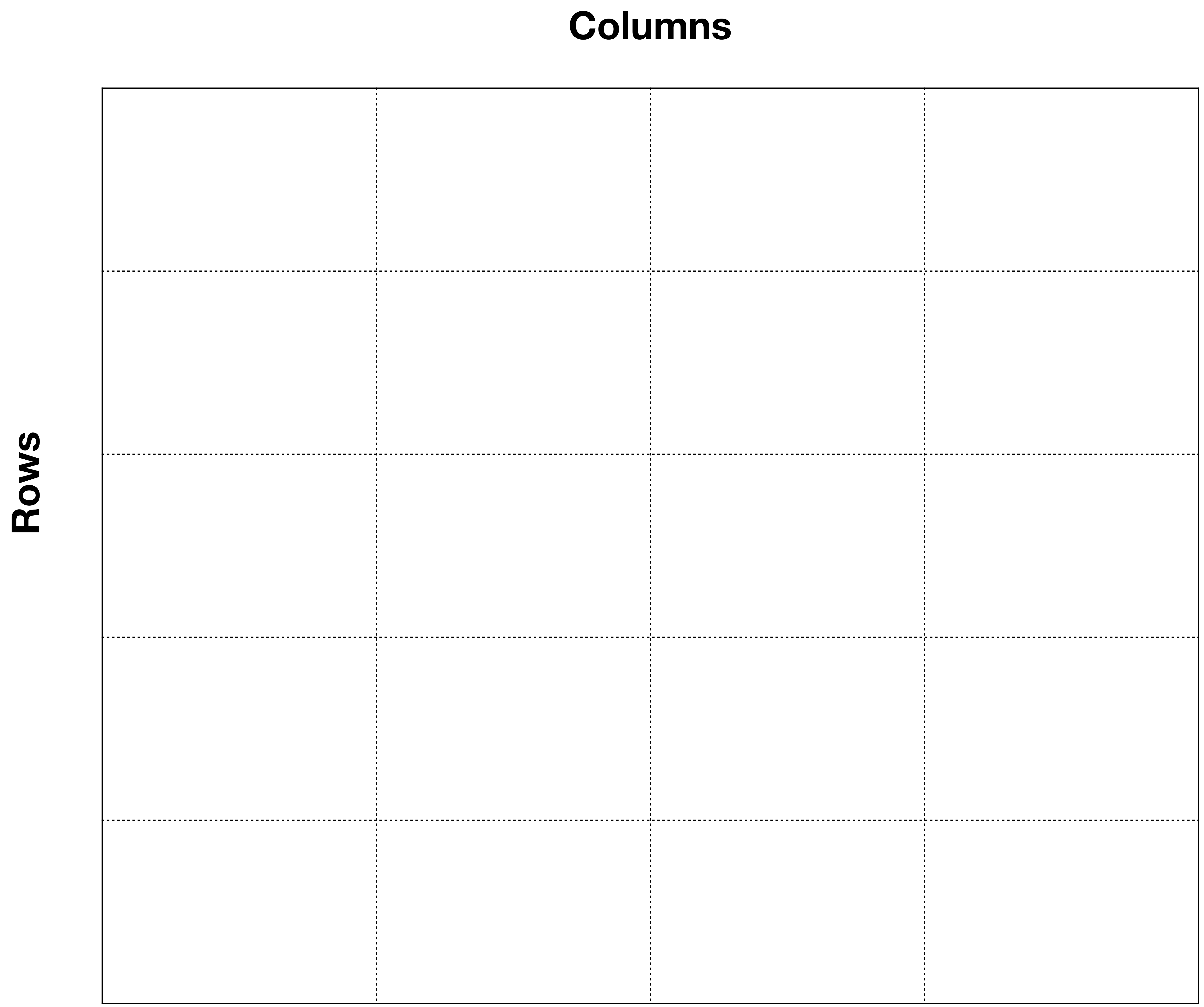
From Chad Topaz's "<u>An Ill-Advised Linear Algebra Tutorial</u>"

Recall: tidy data

A data set is **tidy** iff:

1. Each **variable** is in its own **column**
2. Each **case** is in its own **row**
3. Each **value** is in its own **cell**

One reason statisticians like tidy data is it is basically a matrix



Term-document matrix

A common type of matrix used in text analysis is a term-document matrix, sometimes known as a document-term matrix.

A term-document matrix has:

1. Each **term** in its own **column**
2. Each **document** in its own **row**
3. Each **count** in its own **cell**

Term-document matrix

Columns

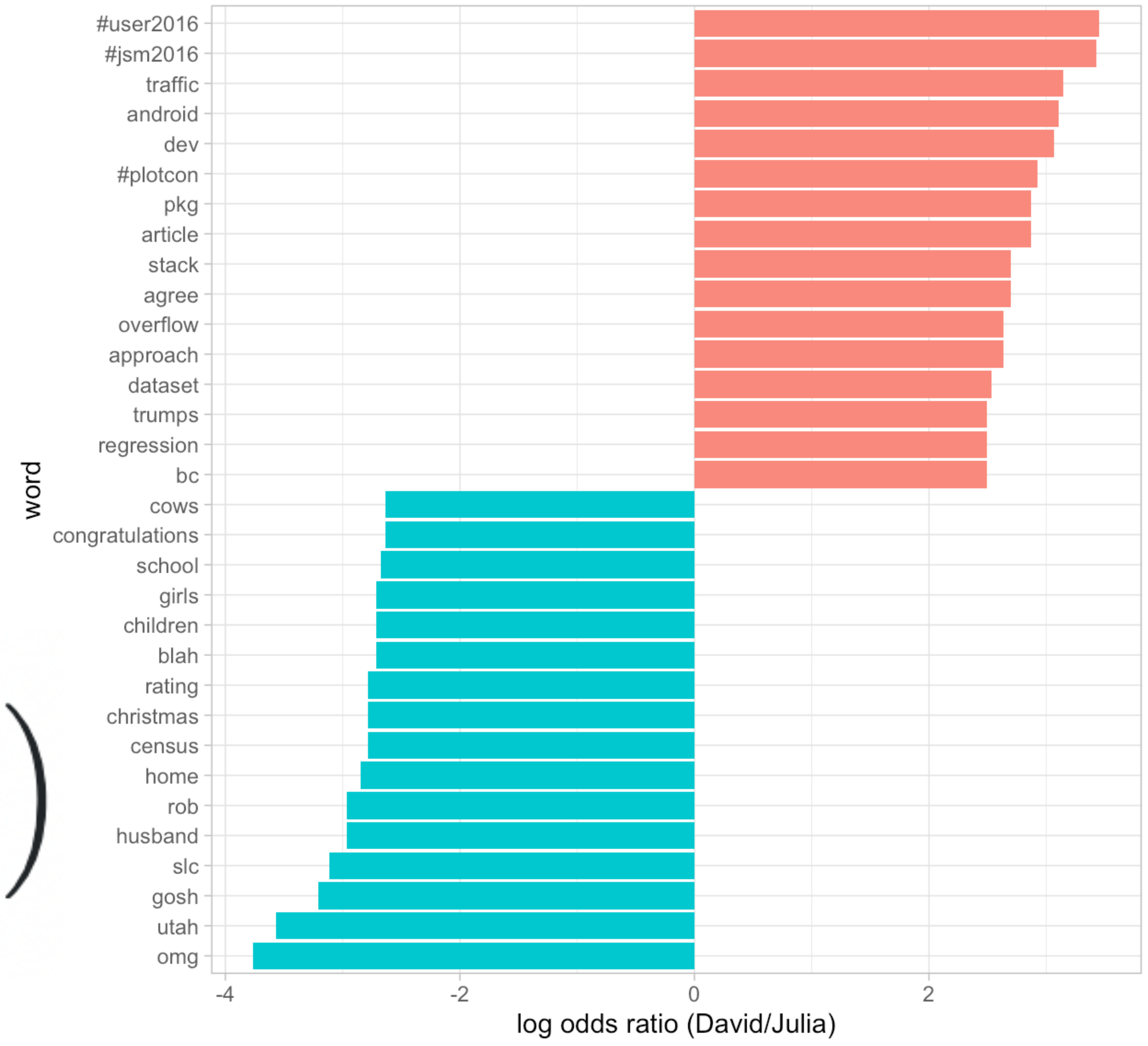
Rows

	The	And	He	To	Bambi	Pooh
Winnie the Pooh	762	999	639	570	0	424
Bambi	2209	1541	1413	1143	657	0

Tf-idf

One reason why term-document matrices can be useful is they allow you to compute the tf-idf— the Term Frequency Inverse Document Frequency

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$



Via Julia Silge and David Robinson's Text Mining with R

"A word is characterized by the company it keeps"

\- John Rupert Firth

Context is important

- Synonyms— words that mean the same thing
- Antonyms— words that mean the opposite thing
- Homonyms— words that sound the same, but have different meanings
  - Homophones— pronounced the same, spelled differently, different meanings
    - Pear/pair
    - Week/weak
    - Meet/meat
    - Sea/sea
  - Homographs — pronounced the same, spelled the same, different meanings
    - Bass
    - Buffet
    - Tear

Which do you think text analysis methods will have the easiest time with? The hardest?

Co-occurrence matrix

Another type of matrix is a co-occurrence matrix, which keeps track of words that co-occur (e.g. in documents or sentences).

A co-occurrence matrix has:

1. Each **term** in its own **column**
2. Each **term** in its own **row**
3. Each **count** in its own **cell**

Co-occurrence matrix

Co-occurrence matrices are good because they preserve context

	land	our	we	possessed
land	0			
our	2	0		
we	2	5	0	
possessed	0	0	1	0

Based on Robert Frost's "The Gift Outright"

What's different about these two types of matrices?

Think about shape, size, sparsity.

	The	And	He	To	Bambi	Pooh
Winnie the Pooh	762	999	639	570	0	424
Bambi	2209	1541	1413	1143	657	0

	land	our	we	possessed
land	0			
our	2	0		
we	2	5	0	
possessed	0	0	1	0

Matrix multiplication

<u>Example</u>. Let's calculate $\begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$. Well, let's first calculate the product for each column vector in the second matrix:

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 9 \\ 8 \\ 17 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 19 \\ 18 \\ 39 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 29 \\ 28 \\ 61 \end{pmatrix}.$$

Now it's smoosh time!

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 5 & 6 \end{pmatrix}\begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 9 & 19 & 29 \\ 8 & 18 & 28 \\ 17 & 39 & 61 \end{pmatrix}.$$

From Chad Topaz's "<u>An Ill-Advised Linear Algebra Tutorial</u>"

Matrix decomposition

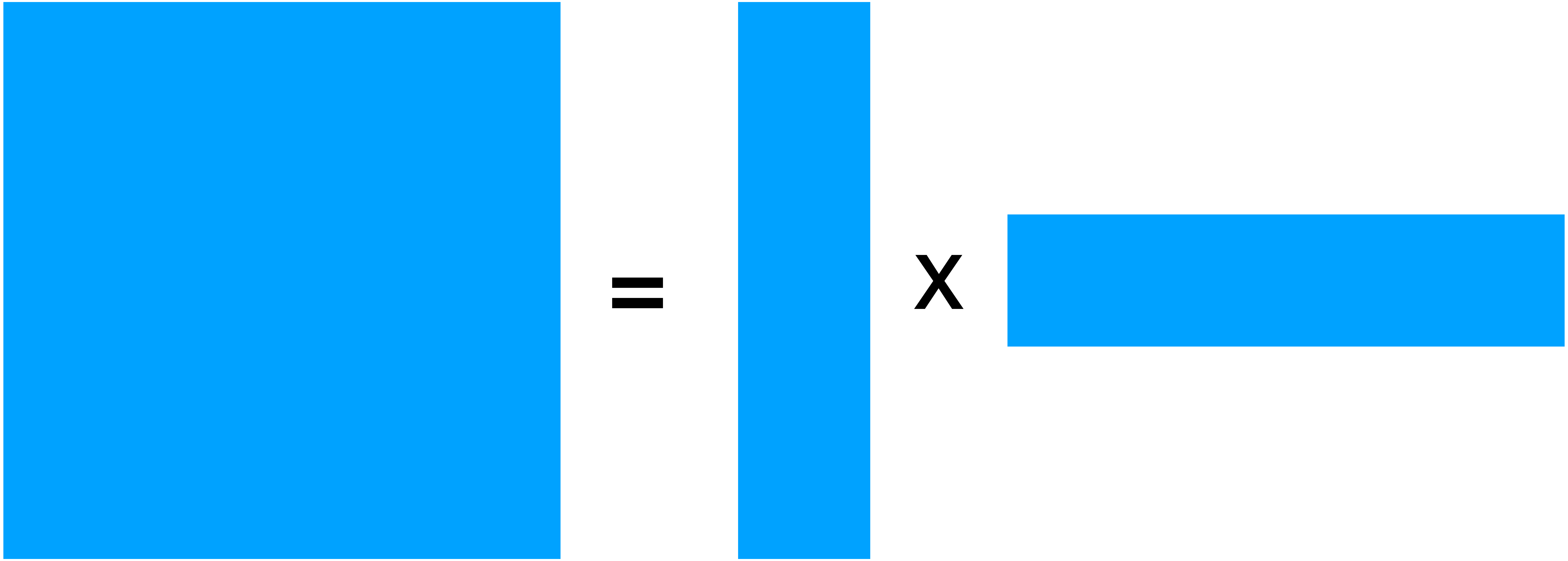
Way more complicated than multiplication, but there are lots of methods to go the other direction

- Eigen decomposition
- Singular value decomposition
- QR decomposition
- ...many more

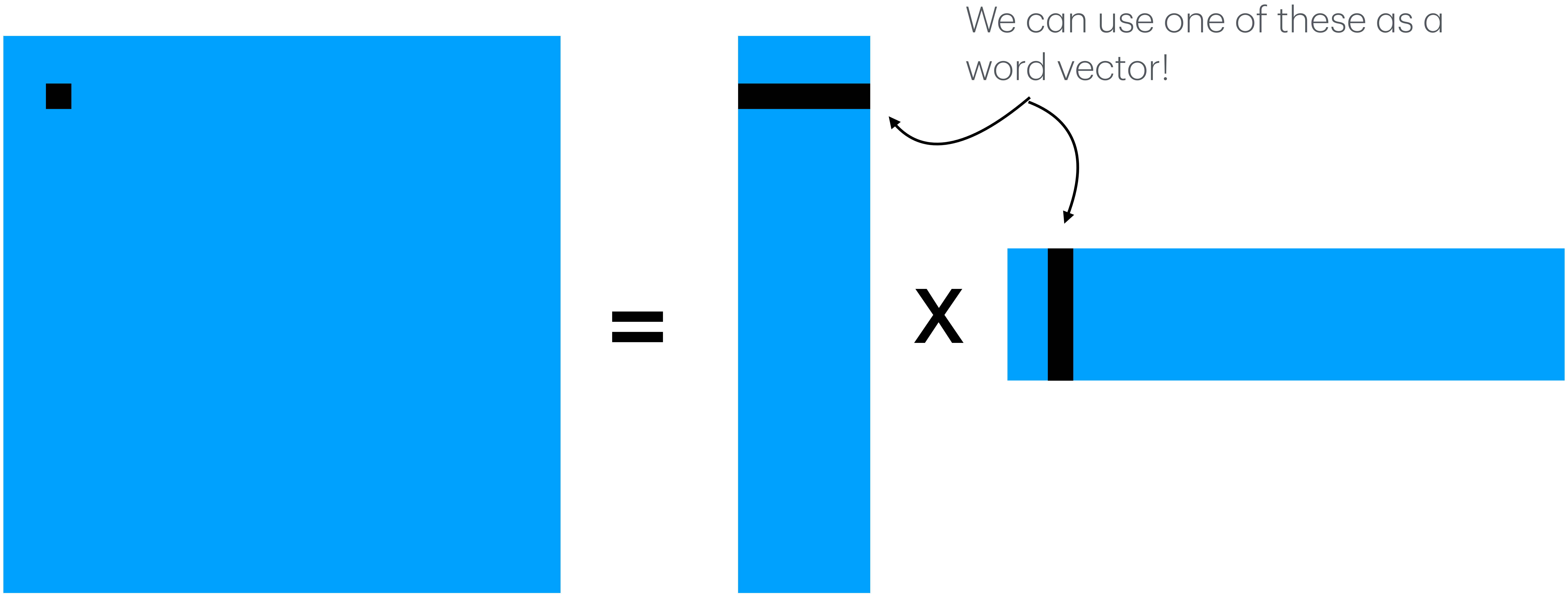
$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 9 & 19 & 29 \\ 8 & 18 & 28 \\ 17 & 39 & 61 \end{pmatrix}.$$

From Chad Topaz's "An Ill-Advised Linear Algebra Tutorial"

Decomposing a co-occurrence matrix



Decomposing a co-occurrence matrix



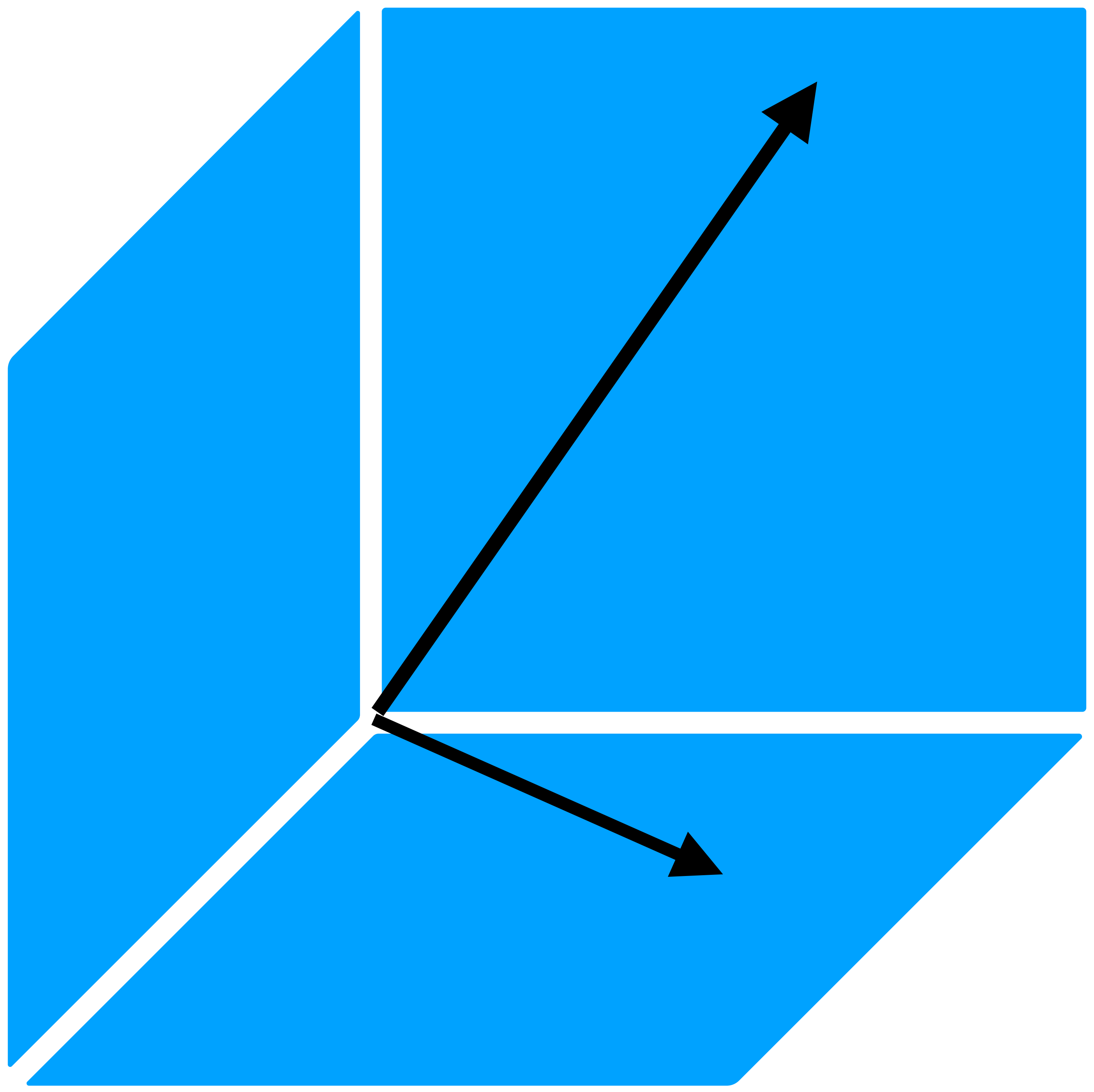
There are other ways to make words into vectors

- word2vec uses neural nets to find word vectors
- GloVe: Global Vectors for Word Representation
- spaCy has their own method!

Word vectors

Word vectors are vectors that represent words. You get to pick the dimensionality.

But… it's probably very high dimensional. Hard to imagine/visualize beyond 3D!



“I see well in many dimensions as long as the dimensions are around two”

- Martin Shubik

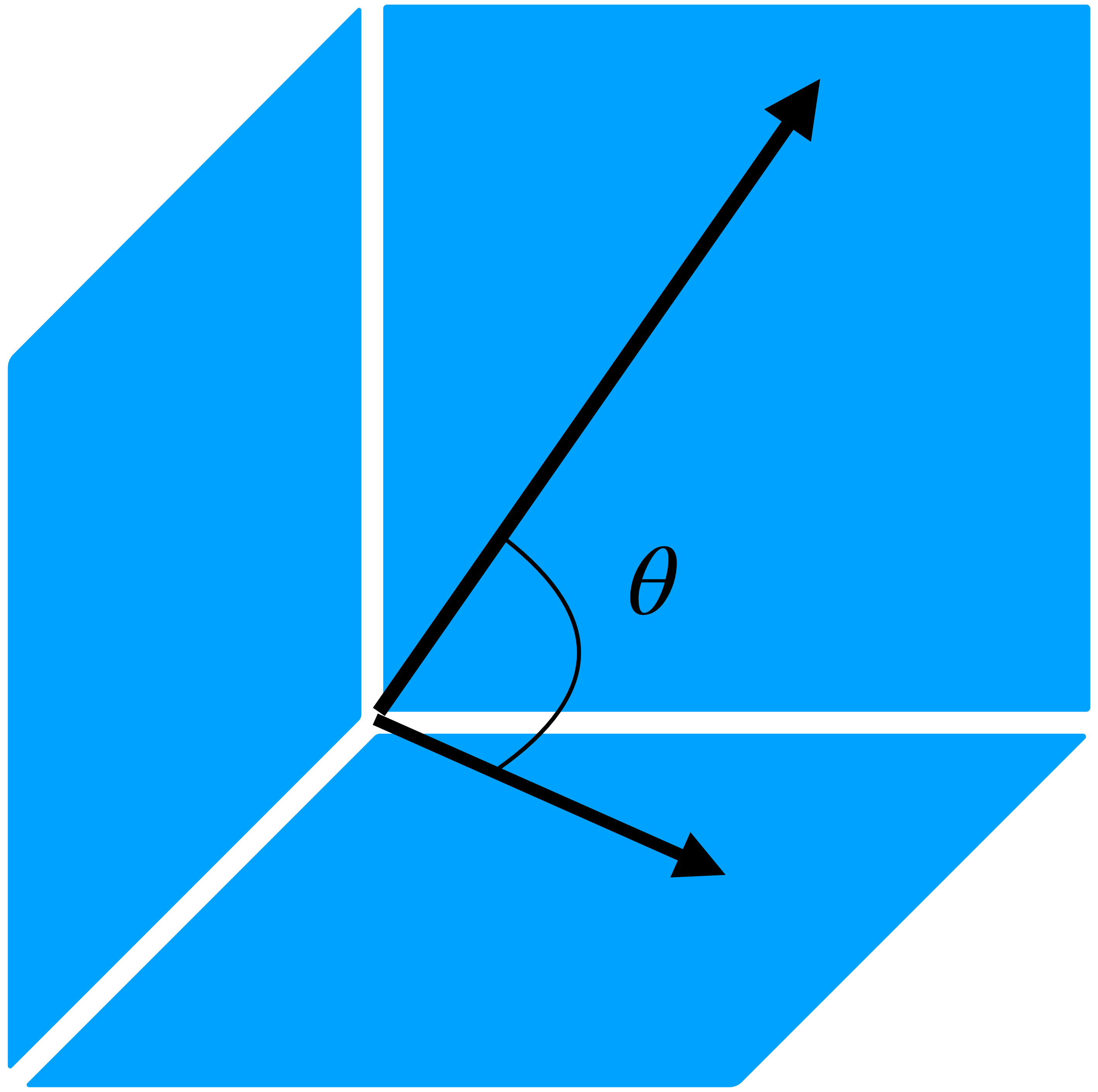
prim9 (John Tukey)

prim9 (John Tukey)

Cosine similarity

One way to measure the “distance” between two vectors is to find the angle between the two vectors (recall— all non-parallel vectors eventually intersect!) and then take the cosine of that angle.

Cosine similarity is always in [-1,1]



Cosine similarity close to 0— the two words are very different. They appear in really different linguistic context.

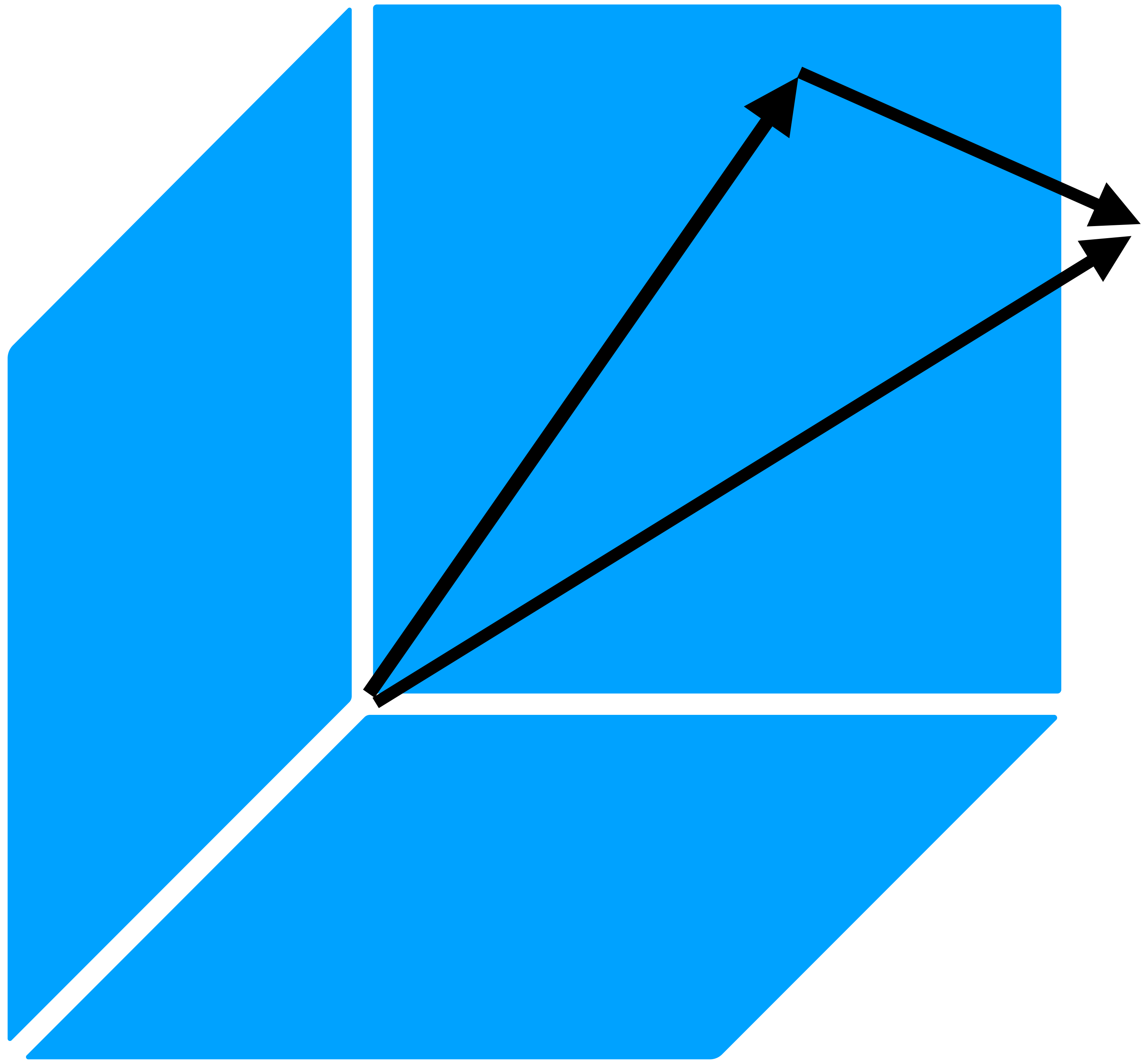
Cosine similarity close to 1— the words are very similar. They appear in similar linguistic contexts.

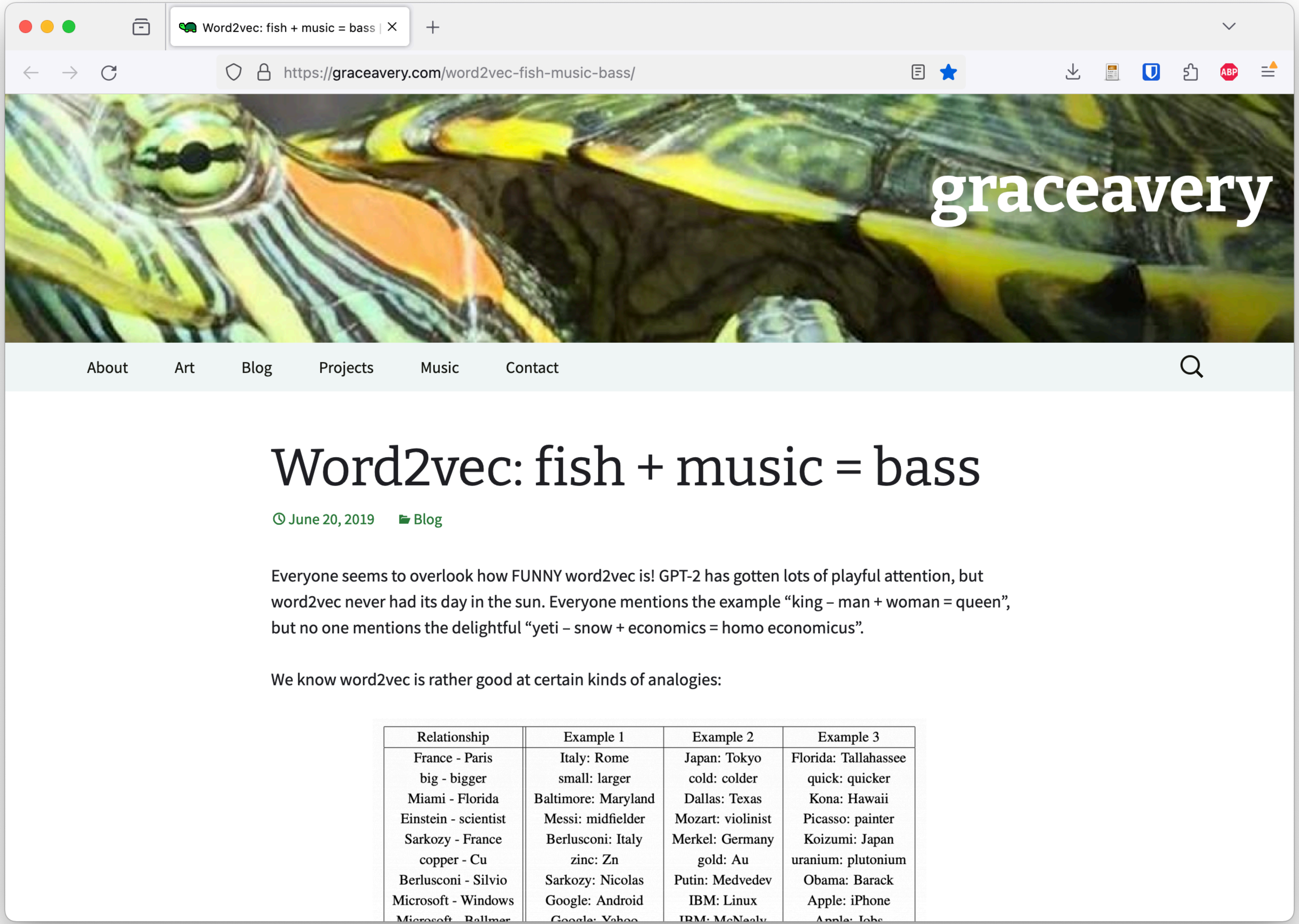
What do you think a cosine similarity of -1 would mean?

Adding and subtracting vectors

Famous example (apocryphal?)

king - man + woman = queen





Word2vec: fish + music = bass

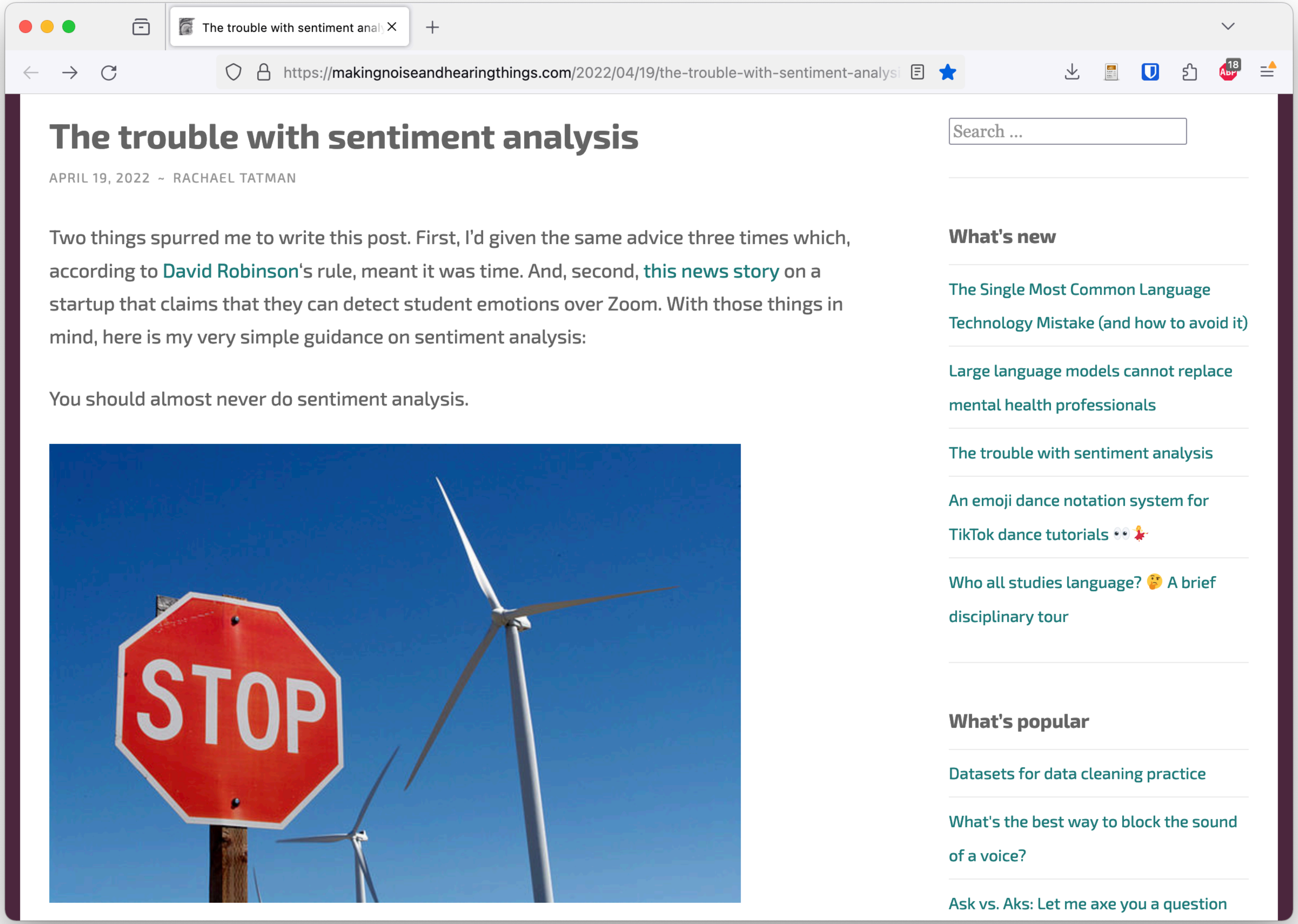
June 20, 2019 Blog

Everyone seems to overlook how FUNNY word2vec is! GPT-2 has gotten lots of playful attention, but word2vec never had its day in the sun. Everyone mentions the example "king – man + woman = queen", but no one mentions the delightful "yeti – snow + economics = homo economicus".

We know word2vec is rather good at certain kinds of analogies:

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs

Word2vec: fish + music = bass. Grace Avery
https://graceavery.com/word2vec-fish-music-bass/



The trouble with sentiment anal

https://makingnoiseandhearingthings.com/2022/04/19/the-trouble-with-sentiment-analysi

The trouble with sentiment analysis

APRIL 19, 2022 ~ RACHAEL TATMAN

Two things spurred me to write this post. First, I'd given the same advice three times which, according to David Robinson's rule, meant it was time. And, second, this news story on a startup that claims that they can detect student emotions over Zoom. With those things in mind, here is my very simple guidance on sentiment analysis:

You should almost never do sentiment analysis.

Search ...

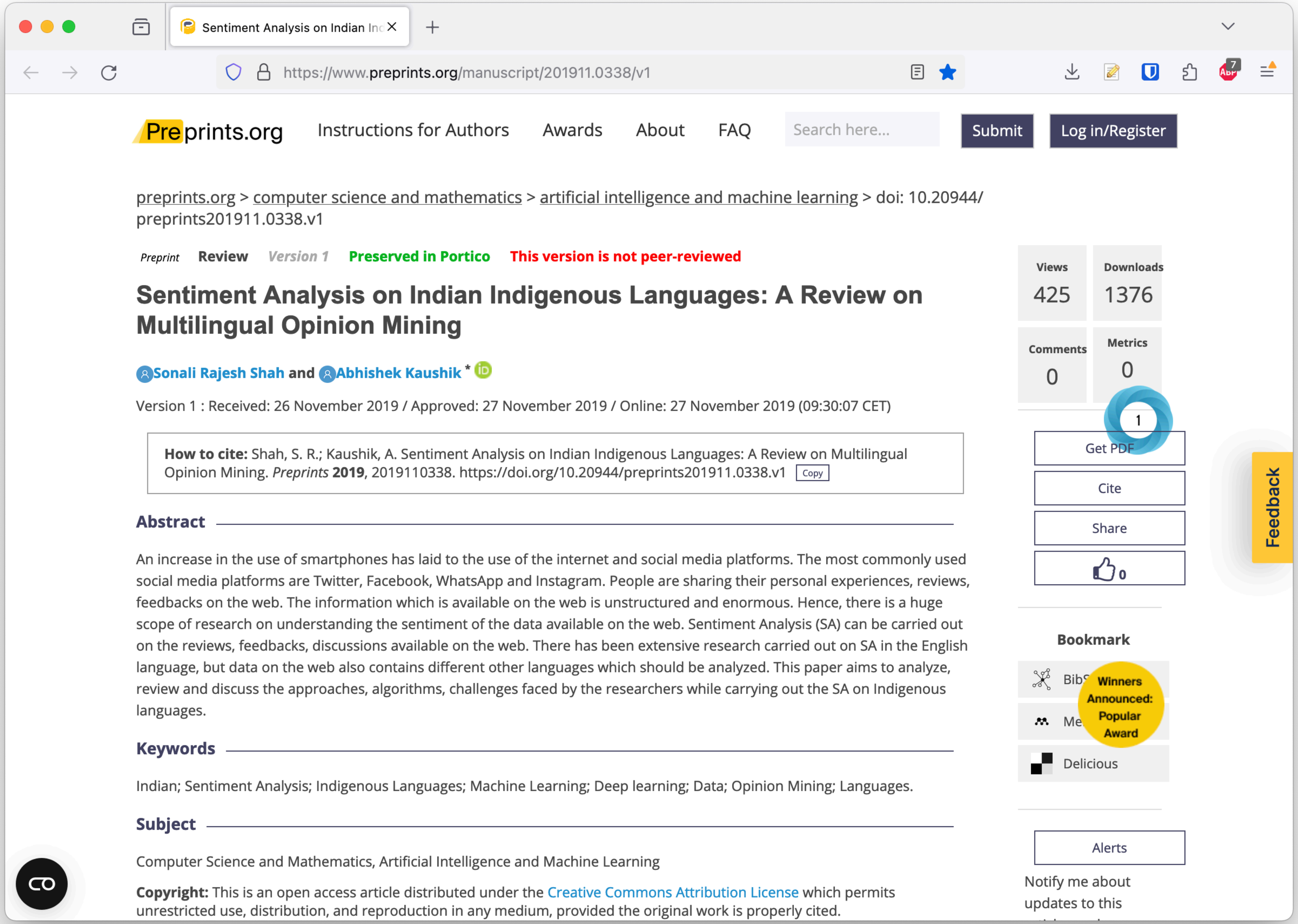
What's new

- The Single Most Common Language Technology Mistake (and how to avoid it)
- Large language models cannot replace mental health professionals
- The trouble with sentiment analysis
- An emoji dance notation system for TikTok dance tutorials 👀💃
- Who all studies language? 🤔 A brief disciplinary tour

What's popular

- Datasets for data cleaning practice
- What's the best way to block the sound of a voice?
- Ask vs. Aks: Let me axe you a question

The trouble with sentiment analysis. Rachel Tatman
https://makingnoiseandhearingthings.com/2022/04/19/the-trouble-with-sentiment-analysis/



preprints.org > computer science and mathematics > artificial intelligence and machine learning > doi: 10.20944/preprints201911.0338.v1

Preprint Review Version 1 Preserved in Portico This version is not peer-reviewed

Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining

Sonali Rajesh Shah and Abhishek Kaushik *

Version 1 : Received: 26 November 2019 / Approved: 27 November 2019 / Online: 27 November 2019 (09:30:07 CET)

How to cite: Shah, S. R.; Kaushik, A. Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining. *Preprints* **2019**, 2019110338. https://doi.org/10.20944/preprints201911.0338.v1 Copy

Abstract

An increase in the use of smartphones has laid to the use of the internet and social media platforms. The most commonly used social media platforms are Twitter, Facebook, WhatsApp and Instagram. People are sharing their personal experiences, reviews, feedbacks on the web. The information which is available on the web is unstructured and enormous. Hence, there is a huge scope of research on understanding the sentiment of the data available on the web. Sentiment Analysis (SA) can be carried out on the reviews, feedbacks, discussions available on the web. There has been extensive research carried out on SA in the English language, but data on the web also contains different other languages which should be analyzed. This paper aims to analyze, review and discuss the approaches, algorithms, challenges faced by the researchers while carrying out the SA on Indigenous languages.

Keywords

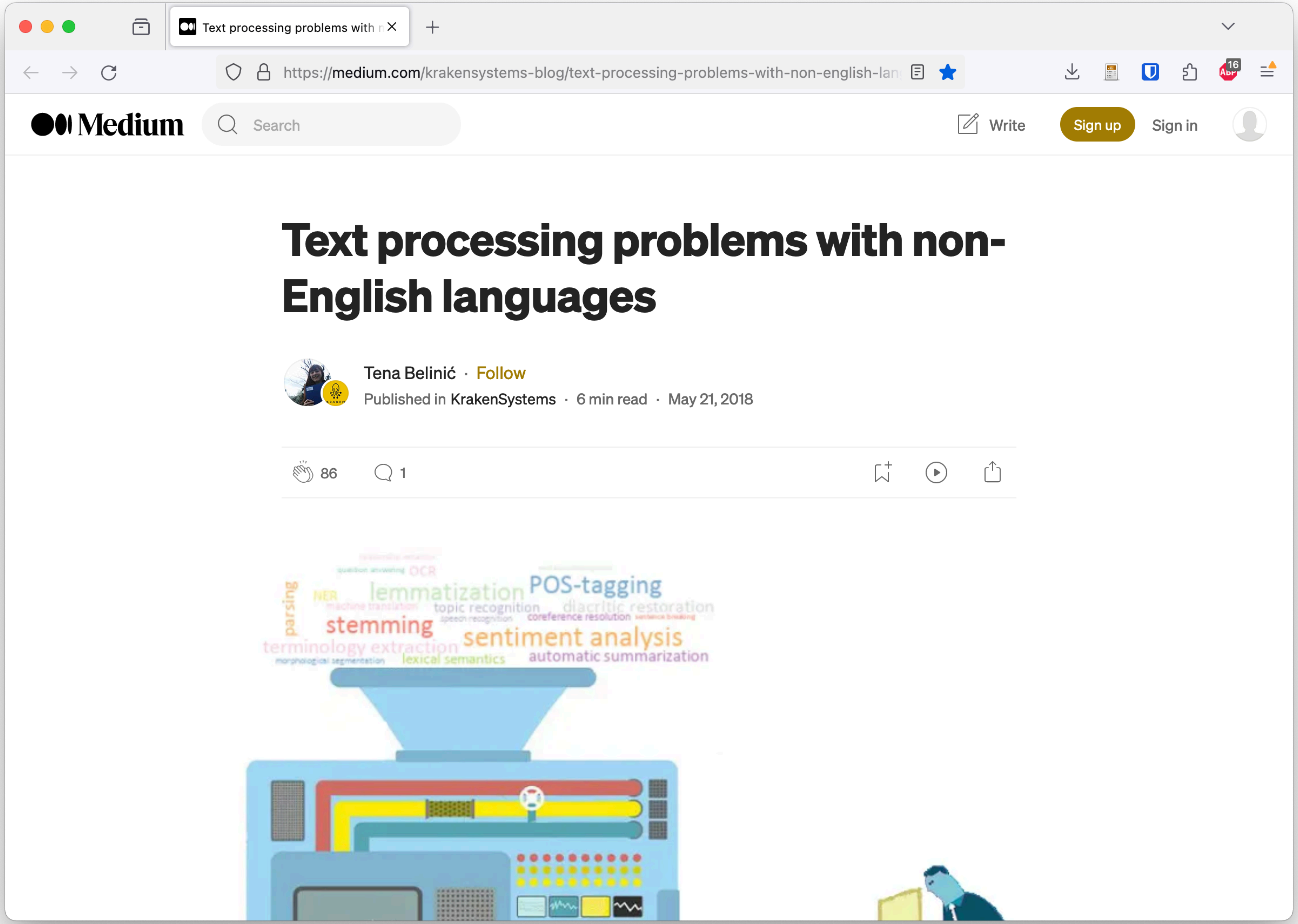
Indian; Sentiment Analysis; Indigenous Languages; Machine Learning; Deep learning; Data; Opinion Mining; Languages.

Subject

Computer Science and Mathematics, Artificial Intelligence and Machine Learning

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining. Sonali Rajesh Shah and Abhishek Kaushik https://doi.org/10.20944/preprints201911.0338.v1



Text processing problems with non-English languages. Tena Belinić
https://medium.com/krakensystems-blog/text-processing-problems-with-non-english-languages-82822d0945dd