



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Word Statistics

1/26/2026

Today (and next class)

- Word-level metrics, statistics, Bayesian Inference
- Exploratory text analysis
 - First approaches when working with a new data set – what can we do with minimal supervision? Minimal information about the data?

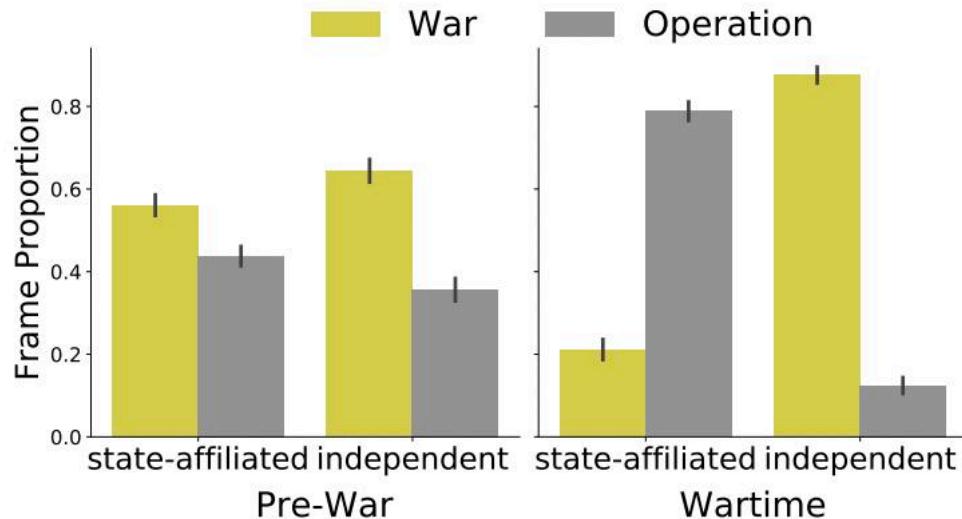
Background

- One of the most fundamental analyses we may want to conduct is, how does word usage differ in different corpora?
 - How do AI policy discussions differ in the U.S. and Europe?
 - Maybe U.S. politicians use words like “innovation” while European politicians use words like “privacy” [fictional example]
 - How do Wikipedia articles about men and women differ?
 - Articles about women focus on family and relationships more than articles about men (Wagner et al. 2015) [fictional words: “family”, “children”, “married”, “divorce”]
- “Entries in the burgeoning “text-as-data” movement are often accompanied by lists or visualizations of how word (or other lexical feature) usage differs across some pair or set of documents”

Example: Russia-Ukraine War



Example: State-affiliated outlets use “operation” over “war”



- We know to look for these terms because of laws passed in Russia, but what if we want to discover these differences?

Running Example: Congressional Record

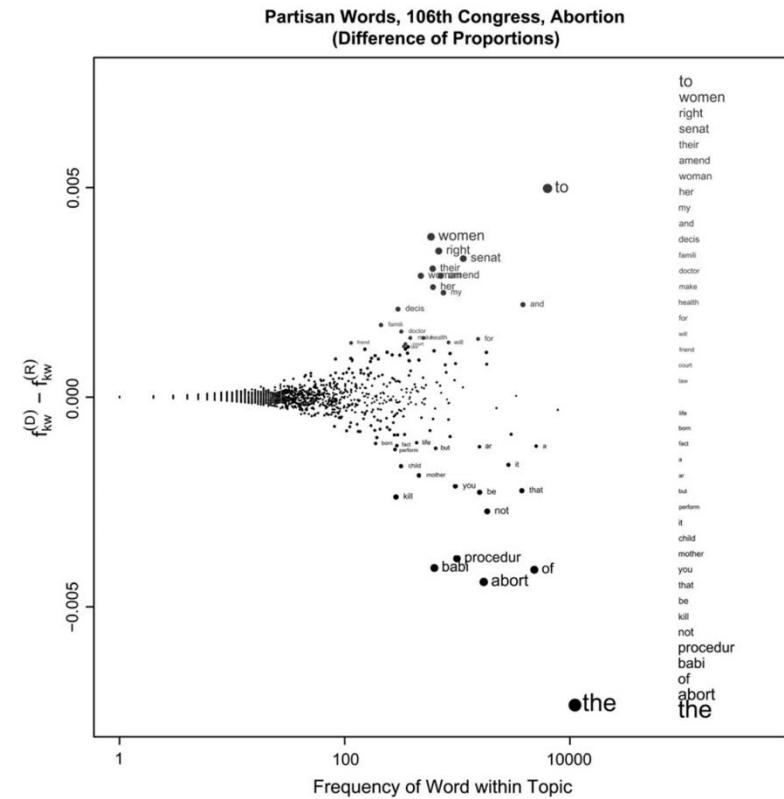
- How does word usage differ in speeches made by Republican and Democratic members of congress?
 - “The question is not which of these terms are partisan and which are not, but which are the most partisan, on which side, and by how much.” [Monroe et al. 2008]

Data credits:

- The corpus was originally constructed in plaintext format by Gentzkow, Shapiro, and Taddy (2018) ([repository for full download](#), [license](#)).
- Preprocessed by Rodriguez and Spirling (2021) ([code](#), [R data file](#)): remove non-alphabetic characters, lowercase, and remove words of length 2 or less, then filter to Congressional sessions 111-114 (Jan 2009 - Jan 2017) and to speakers with party labels D and R.
- Converted plaintext and csv files and subsampled by [Sandeep Soni](#) and [Connor Gilroy](#) ([code](#))

Some initial ideas: proportion of words

- Which words have the highest frequency in statements by Democrats?
 - "the", "and", "that", "this", "for", "have", "are", "not"
- Which words have the highest frequency in statements by Republicans?
 - "the", "and", "that", "for", "this", "have", "are", "our"



Some initial ideas: Odds ratio

- Odds ratio: $O_w^{(i)} = \frac{f_w}{1 - f_w}$, where f_w is the proportion of word w in corpus subset i
- Odds ratio between two groups: $\theta_w^{(i-j)} = \frac{o_w^i}{o_j^w}$
- Log-odds ratio: $\log(O_w^i) - \log(O_j^w)$  is symmetrical

Some initial ideas: Odds ratio

- Odds ratio: $\theta_w^{(i)} = \frac{f_w}{1 - f_w}$
 - Odds ratio between two groups: $\theta_w^{(i-j)} = \frac{\theta_w^i}{\theta_j^w}$
 - Log-odds ratio: $\log(\theta_w^i) - \log(\theta_j^w)$
- Becomes infinite/undefined if words only exist in one corpus

Odds ratio in Congressional data

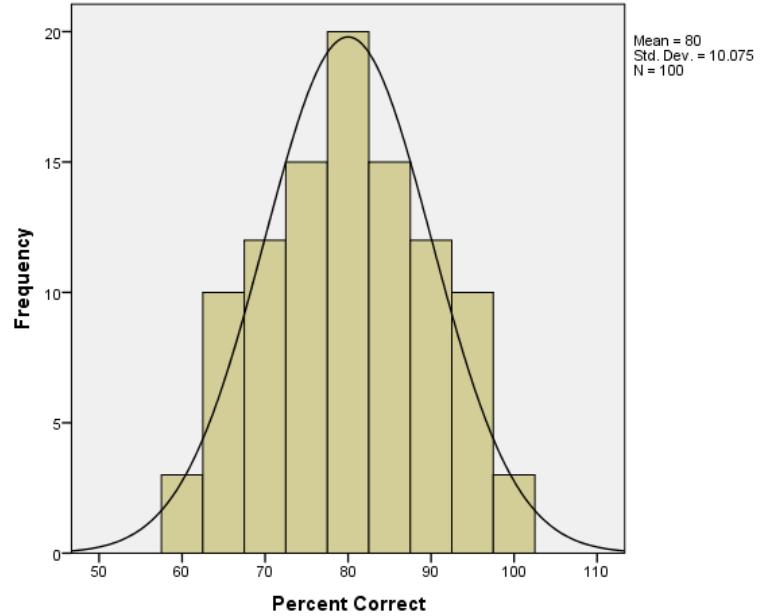
Word	Odds-Ratio	Frequency in Republican Speech	Frequency in Democratic Speech
governmentapproved	-4.90	239	2
capandtax	-4.98	477	4
partialbirth	-4.69	97	1
kansan	-4.4	73	1
corinthian	5.26	1	217
antihunger	5.08	1	180
trayvon	4.87	1	146
vermonters	4.81	4	554

Some initial ideas: Odds ratio

- Odds ratio: $\theta_w^{(i)} = \frac{f_w}{1 - f_w}$
 - Odds ratio between two groups: $\theta_w^{(i-j)} = \frac{\theta_w^i}{\theta_j^w}$
 - Log-odds ratio: $\log(\theta_w^i) - \log(\theta_j^w)$
-
- The diagram consists of three blue arrows originating from the right side of the equations in the list. The top arrow points from the first equation to the text "Becomes infinite if words only exist in one corpus". The middle arrow points from the second equation to the text "Becomes dominated by obscure words". The bottom arrow points from the third equation to the text "Becomes dominated by obscure words".
- Becomes infinite if words only exist in one corpus
- Becomes dominated by obscure words

Model-driven approach

- Clear that simple methods aren't going to work
- General statistical modeling approach:
 - Given a collection of data
 - Assume you generated this data from some model
 - Estimate model parameters
- Example:
 - Assume you gathered data by sampling from a normal distribution
 - Estimate mean and stdev



Model-driven approach

- High-level idea:
 - First model the word usage in the full collection of documents
 - Then investigate how subgroup-specific word usage diverges from that in the full collection of documents
- Incorporate a *prior*
 - Background estimate of how often a word is used in this type of document

Bag-of-words (BOW) assumption

- “the state of healthcare in this country is...”
- We ignore ordering of words and assume that we can represent the document collection as a “bag of words”
- [We’ve already been doing this implicitly]

country state the
in this
healthcare
is of

Terminology

- **y** = vector of term counts in the corpus

101	60	10	...	11	231
country	state	healthcare	...	employment	the

- n = number of terms in the corpus
- $n = 101 + 60 + 10 \dots + 11 + 231$

Terminology

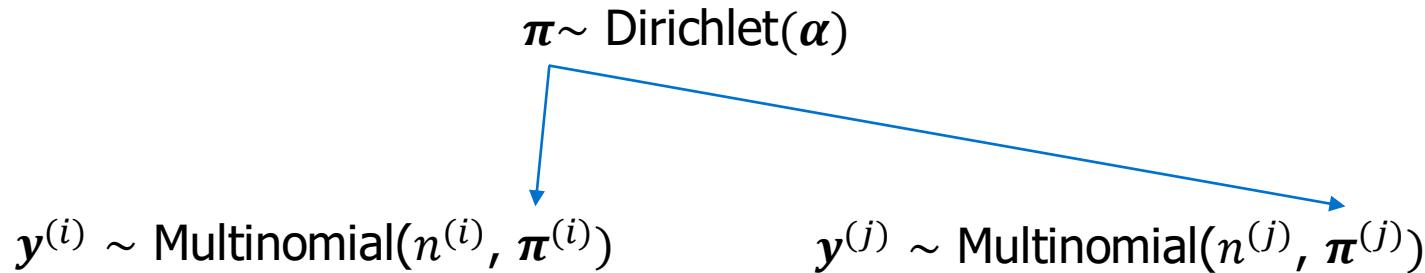
Define:

- \mathbf{y} = vector of term counts in the corpus
- n = number of terms in the corpus
- π = unknown distribution the vocabulary

▪ Assume:

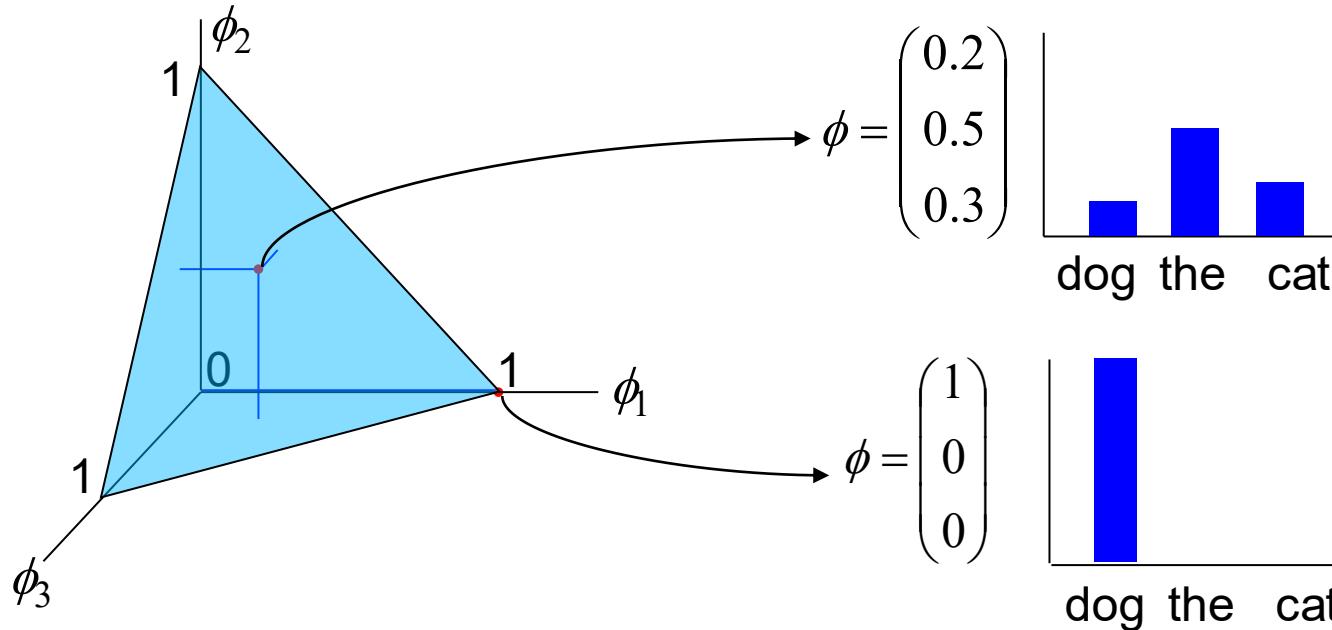
- $\mathbf{y} \sim \text{Multinomial}(n, \pi)$
- Intuition: we got \mathbf{y} by repeatedly sampling from a bag. π describes how many of each word is in the bag

Impose *Dirichlet Prior* on π



What is a Dirichlet distribution?

- We can plot multinomial probability distributions
- Shape we get is a *simplex*

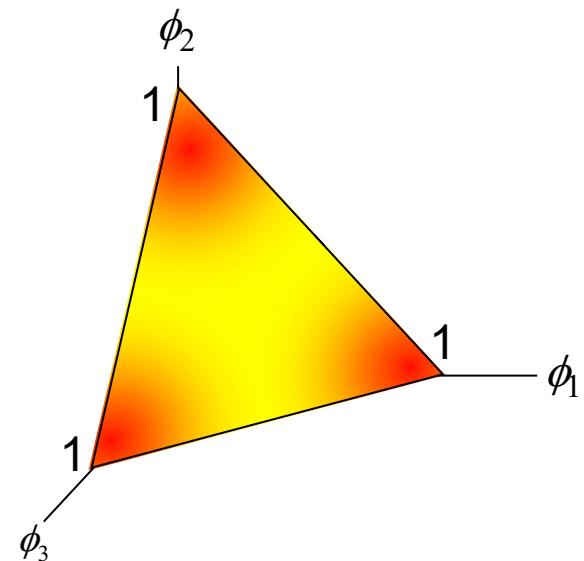
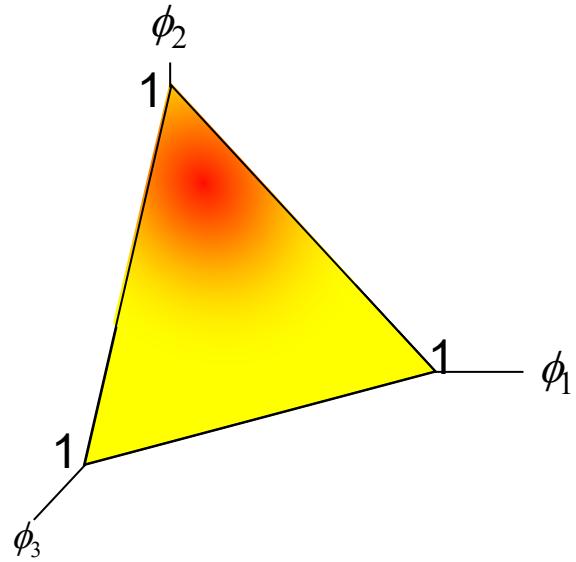
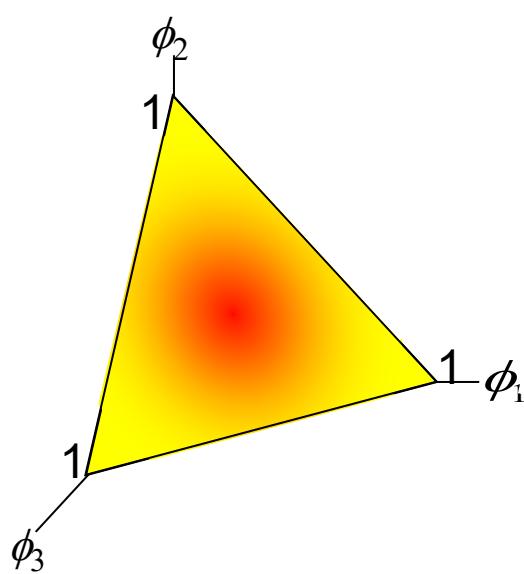


$$\sum_i \phi_i = 1$$

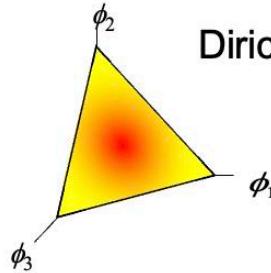
$$\sum_i \phi_i = 1$$

What is a Dirichlet distribution?

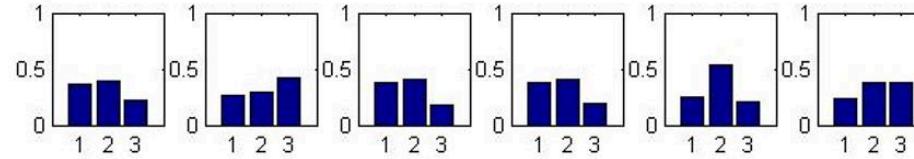
- A Dirichlet distribution is a distribution over multinomial distributions ϕ in the simplex



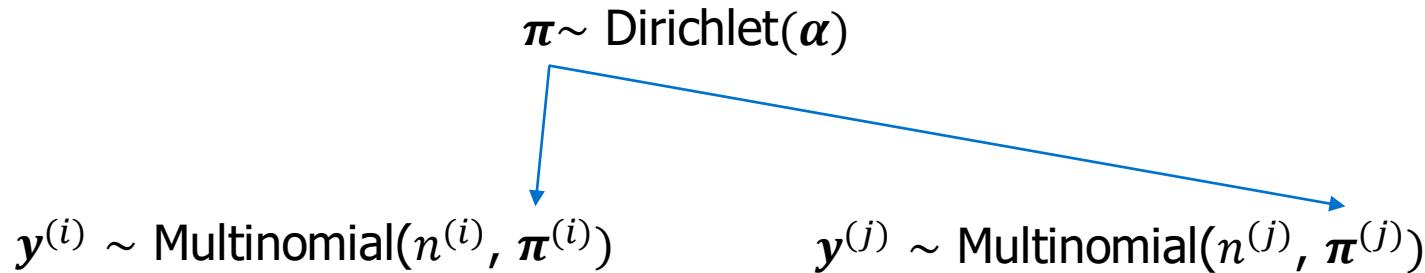
Example draws from a Dirichlet distribution over the 3-simplex



Dirichlet(5,5,5) [higher alpha – more dense]



Impose *Dirichlet Prior* on π



Impose *Dirichlet Prior* on π

$$\pi \sim \text{Dirichlet}(\alpha)$$



Frequency of a term in
the *entire* corpus

321	176	53	...	54	543
country	state	healthcare	...	employment	the

Impose *Dirichlet Prior* on π

$$\begin{array}{ccc} \pi \sim \text{Dirichlet}(\alpha) & & \\ \downarrow & & \searrow \\ y^{(i)} \sim \text{Multinomial}(n^{(i)}, \pi^{(i)}) & & y^{(j)} \sim \text{Multinomial}(n^{(j)}, \pi^{(j)}) \end{array}$$

$y^{(i)}$ can be word frequencies for Democrat Speech
 $y^{(j)}$ can be word frequencies for Republican Speech

Both are assumed to have the same prior – frequency in general congressional speech

Generative Story

1. Draw $\pi^{(i)} \sim \text{Dirichlet}(\alpha)$
2. For $n^{(i)}$ steps:
 1. Draw $w \sim \text{Multinomial}(\pi^{(i)})$

For each subset of our corpus,

- $y^{(i)}$, $n^{(i)}$ and α are observed in the data (where $y^{(i)}$ contains counts of w)
- $\pi^{(i)}$ is what we need to estimate

Another aside about distributions

- Prior distribution: $P(\pi)$
- Posterior distribution: $P(\pi | w)$
- When the posterior distribution is in the same family as the prior distribution, they are called **conjugate distributions**
- The Dirichlet distribution is a **conjugate prior** of the multinomial distribution
- [For our purposes, we often chose a Dirichlet prior for a multinomial distribution because it makes inference easier]

Point estimate of π

$$\hat{\pi}_w^{(i)} = \frac{y_w^{(i)} + \alpha_w}{n^{(i)} + \alpha_0} \quad \longrightarrow \text{Point estimate of } \pi, \text{ where } \alpha_0 = \sum \alpha_w$$

Intuitive interpretation: imagine we saw α_0 additional words and α_w were w

Point estimate of π

$$\hat{\pi}_w^{(i)} = \frac{y_w^{(i)} + \alpha_w}{n^{(i)} + \alpha_0} \longrightarrow \text{Point estimate of } \pi, \text{ where } \alpha_0 = \sum \alpha_w$$

$$\hat{\delta}_w^{(i-j)} = \log\left(\frac{\pi_w^{(i)}}{1-\pi_w^{(i)}}\right) - \log\left(\frac{\pi_w^{(j)}}{1-\pi_w^{(j)}}\right) \longrightarrow \text{Log-odds ratio with } \pi \text{ instead of frequencies}$$

$$\hat{\delta}_w^{(i-j)} = \log\left(\frac{y_w^{(i)} + \alpha_w}{n^{(i)} + \alpha_0 - y_w^{(i)} - \alpha_w}\right) - \log\left(\frac{y_w^{(j)} + \alpha_w}{n^{(j)} + \alpha_0 - y_w^{(j)} - \alpha_w}\right)$$

Congressional data with Dirichlet prior

Word	Frequency in Republican Speech	Frequency in Democratic Speech
idahoans	210	0
fairtax	130	0
cdh	102	0
isna	98	0
zinser	0	160
gaspee	0	127
vania	0	105
fiveminute	0	95

We don't have to drop zero counts anymore, but this isn't that much better than before!

We could impose a stronger prior?

Variance

- Report *z-score*: point estimate divided by variance
 - Lower-frequency words have higher variance

With some assumptions, we can estimate:

$$\sigma^2(\hat{\delta}_w^{(i-j)}) \approx \frac{1}{y_w^{(i)} + \alpha_w^{(i)}} + \frac{1}{y_w^{(j)} + \alpha_w^{(j)}}$$

And use as our final score:

$$\frac{\delta_w^{(i-j)}}{\sqrt{\sigma^2(\delta_w^{(i-j)})}}$$

Odds ratio in Congressional data

Top Republican Words	Score	Top Democrat Words	Score
spending	-66.26	republican	56.63
obamacare	-59.90	wealthiest	40.78
government	-47.92	rhode	39.43
going	-45.33	women	38.16
that	-44.58	pollution	33.66
trillion	-43.43	republicans	32.86
taxes	-42.39	gun	32.45
you	-40.85	investments	32.22
administration	-39.07	families	31.93
debt	-38.92	violence	30.88

New Example: Narrative framing in restaurant reviews

- As online reviews have become commonplace, they offer an opportunity to study consumer behavior
- How do consumers frame positive and negative sentiment online?
- Data:
 - 900,000 Yelp restaurant reviews from 9 cities: Boston, Chicago, Los Angeles, New York, Philadelphia, San Francisco, and Washington D.C
 - Corpus subsets:
 - “i” = one star reviews
 - “j” = 5 star reviews
 - Prior: entire review corpus

New Example: Narrative framing in restaurant reviews

Table 2: Top 50 words associated with one-star reviews by the Monroe, *et al.* (2008) method.

Linguistic class	Words in class
Negative sentiment	worst, rude, terrible, horrible, bad, awful, disgusting, bland, tasteless, gross, mediocre, overpriced, worse, poor
Linguistic negation	no, not
First person plural pronouns	we, us, our
Third person pronouns	she, he, her, him
Past tense verbs	was, were, asked, told, said, did, charged, waited, left, took
Narrative sequencers	after, then
Common nouns	manager, waitress, waiter, customer, customers, attitude, waste, poisoning, money, bill, minutes
Irrealis modals	would, should
Infinitives and complementizers	to, that

"In summary, one-star reviews were overwhelmingly focused on narrating experiences of trauma rather than discussing food, both portraying the author as a victim and using first person plural to express solace in community."

More serious example: Racial differences CPS services

- Words used in caseworker notes about families referred to child protective services
- Compare words used in notes about Black families vs. white families

Black-assoc.	Score	White-assoc.	Score
Referrals			
she	52.19	he	54.64
belt	47.37	heroin	41.87
her	45.39	PGF	36.08
BM	37.90	treatment	36.16
bus	30.95	anxiety	34.25
shelter	25.11	using	27.45
whooped	23.96	therapist	26.05
Cases			
school	56.80	F	130.67
housing	42.01	parents	59.26
informed	37.76	drug	37.65
pass	35.75	methadone	36.55

Alternate Approach: Pointwise-mutual information

- Probability/Information theory measure of association
- Common formulation: measure how often two events, x and y occur, compare with what we would expect if they were independent

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

How often we observe x and y together

How often we expect x and y to co-occur, if they each occur independently

The diagram illustrates the components of the Pointwise Mutual Information (PMI) formula. It shows the formula $PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$ with two blue arrows pointing from the terms $p(x, y)$ and $p(x)p(y)$ to their respective definitions. The top arrow points to the text "How often we observe x and y together". The bottom arrow points to the text "How often we expect x and y to co-occur, if they each occur independently".

Alternate Approach: Pointwise-mutual information

- Compute the co-occurrence between a word w and a label i

$$PMI(w, i) = \log \frac{p(w, i)}{p(w)p(i)}$$

Probability of w and i co-occurring

Probability of w occurring

Probability of i occurring

```
graph TD; PMI["PMI(w, i) = log p(w, i) / p(w)p(i)"] --> CoOccur["Probability of w and i co-occurring"]; PMI --> WProb["Probability of w occurring"]; PMI --> IProb["Probability of i occurring"]
```

Computing PMI

$$PMI(w, i) = \log \frac{p(w, i)}{p(w)p(i)} = \log \frac{p(w|i)p(i)}{p(w)p(i)} = \log \frac{p(w|i)}{p(w)}$$

	country	state	healthcare	...	employment	the	Total
Republican	321	176	15	...	54	500	10233
Democratic	100	31	53	...	20	543	12231
Total	421	207	68	...	74	1043	22464

$$PMI(\text{Republican}, \text{employment}) = \log \frac{\left(\frac{54}{22464}\right)}{\left(\frac{74}{22464}\right)\left(\frac{10233}{22464}\right)}$$

Alternate Approach: Pointwise-mutual information

- Compute the co-occurrence between a word w and a label i

$$PMI(w, i) = \log \frac{p(w, i)}{p(w)p(i)}$$

Number of times w occurs in i -labeled documents / number of total words

Proportion of w

Proportion of i -labeled terms

Alternate Approach: Pointwise-mutual information

- Common to use *Positive Pointwise mutual information (PPMI)*
 - Set PMI to 0 wherever it is negative
- Still run into problems with over-emphasizing rare words:
 - There are some fixes for this, including smoothing
- PMI scores are used frequently

Example of PMI: Gender Bias on Wikipedia

- [Only include words that occur in at least 1% of biographies]
- Women: actress (15.9%), women's (8.8%), female (5.6%), **her husband** (4.1%), women (5.3%), first woman (1.9%), film actress (1.6%), her mother (1.8%), woman (4.4%), **nee** (3.6%), feminist (1%), miss (1.9%), model (3.3%), girls (1.5%) and singer (6.5%).
- Men: played (14.2%), footballer who (3.0%), football (4.5%), league (5.9%), john (7.9%), major league (1.8%), football league (1.6%), college football (1.5%), son (7%), football player (2.2%), footballer (2%), served (11.7%), william (4.6%), national football (2%) and professional footballer (1%).

Additional Applications

- PPMI and variants of odds ratio are commonly used as *features* in other NLP tasks (not just for word statistics on their own)
 - Represent a document using one of these metrics instead of using word counts
 - Document vectors can be used for similarity metrics, e.g. clustering or information retrieval

	country	state	healthcare	...	employment	the
Republican	321	176	15	...	54	500
Democratic	100	31	53	...	20	543

Today's takeaways

- Counting words can be surprisingly hard!
- Key ideas behind two popular methods for examining word statistics:
 - Log-odds with a Dirichlet prior ("Fightin' Words")
 - Pointwise mutual information scores
- Examples of applications and understanding of when these methods are useful

Reminders

- Course website: <http://nlp-css-601-672.cs.jhu.edu/sp2025/>
- Join class Piazza
- Fill out course goals survey (linked on slides from last class)

References

- Jurafsky and Martin, 2022, Sec 6.6
 - https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf
- Monroe BL, Colaresi MP, Quinn KM. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*. 2008;16(4):372-403. doi:10.1093/pan/mpn018

End

- "We note, however, the practice of stop word elimination has been found generally to create more problems than it solves, across natural language processing applications. Manning et al. (2008) observe: "The general trend ... over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever" (p. 27). They give particular emphasis to the problems of searching for phrases that might disappear or change meaning without stop words (e.g., "to be or not to be"). In our example, a stop list would eliminate a word like *her*, which almost definitely has political content in the context of abortion,⁹ and a word like *their*, which might (e.g., women and their doctors)."
 - Core problem is not that they are stopwords but that they are particularly frequent



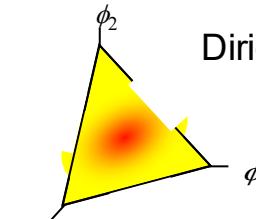
- “In general, our strategy is to first model word usage in the full collection of documents and to then investigate how subgroup-specific word usage diverges from that in the full collection of documents.”



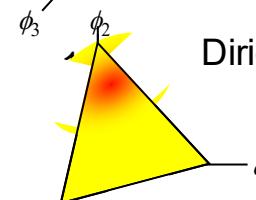
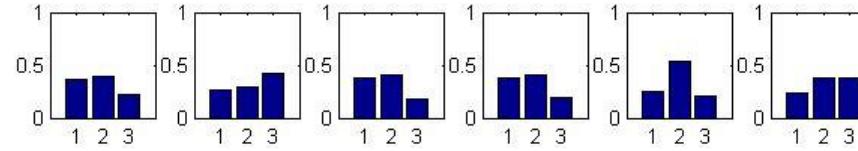
- One approach is to use more of what we know about the expected distribution of words. We can do this by specifying a prior proportional to the expected distribution of features in a random text. That is, we know the is used much more often than nuclear, and our prior can reflect that information.

Dirichlet Distribution

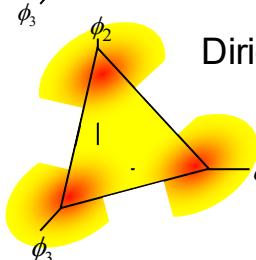
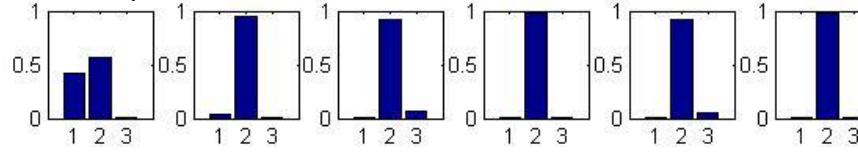
- Example draws from a Dirichlet Distribution over the 3-simplex:



Dirichlet(5,5,5)



Dirichlet(0.2, 5, 0.2)



Dirichlet(0.5,0.5,0.5)

