Finding Distinctive terms

Sandeep Soni

09/13/2023

Class logistics

- Hw4 released

From Last Class

- Bag of words demo
- I've added code to create a counts term-document matrix

Questions for the day

“How to calculate the similarity between documents?”

“How to compare groups using words?”

Agenda

- How to measure the “similarity” or “differences” between documents?
- How to measure differences between groups?
- How to find words that differentiate between two groups?

Distance metrics

- Many ways to compare documents. A few examples:
  - Levenshtein distance
  - Jaccard index
  - Cosine similarity

Levenshtein Distance

- Objective: Transform one document to another by editing it
- Assign cost c for an edit operation
  - Insertion
  - Deletion
  - Replacement
- Dist (Doc1, Doc2) = min ($\sum_i c_i$)

Sherlock Holmes

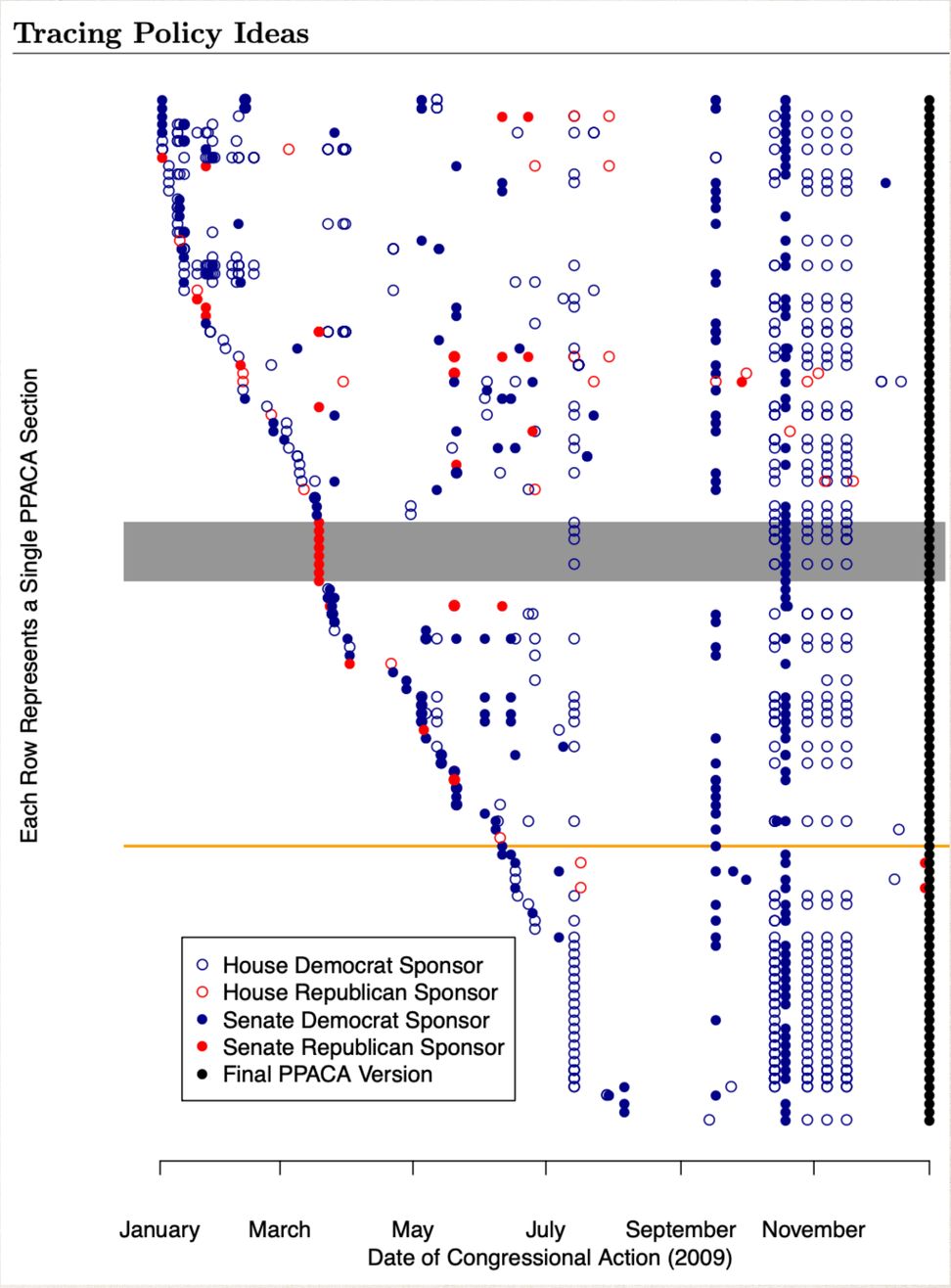
Doc 1

She locked Homes

Doc 2

Sherlock Holmes	Doc1
Sherlock Holmes	delete “l”
Sherlock Homes	replace r by “ ”
She locke Homes	insert “e”
She locked Homes	insert “d”
She locked Homes	Doc2

- Distance can be calculated using dynamic programming
- Many variations such as Needleman-Wunsch and Smith-Waterman
- Useful to quantify sequence alignment or text reuse



Wilkerson, John, David Smith, and Nicholas Stramp. "Tracing the flow of policy ideas in legislatures: A text reuse approach." *American Journal of Political Science* 59.4 (2015): 943-956.

Jaccard Index

- Objective: measure commonality between documents
- Treat documents as sets of tokens
- $J(Doc1, Doc2) = \frac{|Doc1 \cap Doc2|}{|Doc1 \cup Doc2|}$

To Sherlock Holmes she is always *the* woman.

Doc 1

Sherlock Holmes is always right.

Doc 2

- $J(Doc1, Doc2) = \frac{4}{9}$

- Many variations and generalizations (e.g., overlap coefficient, Tversky index)
- Useful to quantify lexical overlap and lexical variation

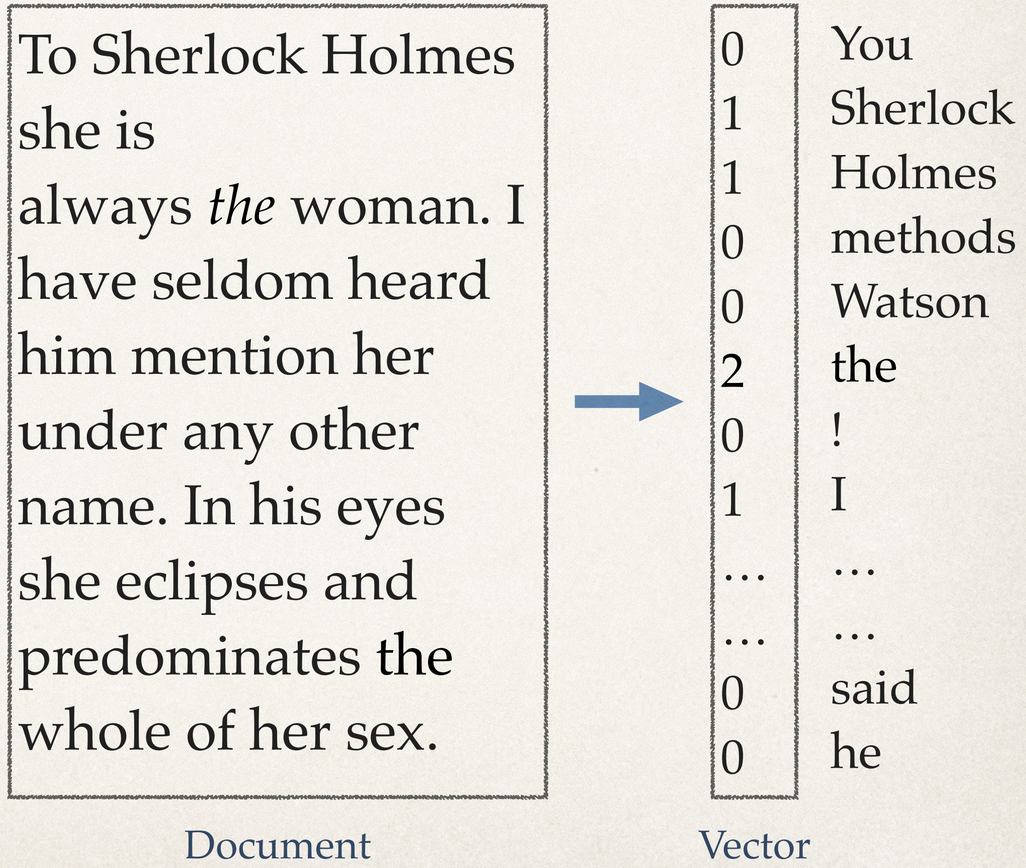
Num	Obsecro Te 1
1	Obsecro Te domina sancta maria mater dei pietate plenissima summi
2	regis filia mater gloriosissima mater orphanorum consolatio
3	desolatorum via errantium **salus in te** sperantium virgo ante
4	partum virgo in partu **et** virgo post partum **Fons misericordie**
5	fons salutis et gratie fons pietatis et leticie fons consolationis
6	et indulgencie Per illam sanctam ineffabilem leticiam
7	qua exultavit spiritus tuus in illa hora quando tibi per gabrielem
8	annunciatus filius dei fuit
9	Et per illud divinum mysterium quod tunc operatus est spiritus sanctus

Obsecro Te 2
Obsecro Te domina sancta maria mater dei pietate plenissima summi
regis filia mater gloriosissima mater orphanorum consolatio
desolatorum via errantium **salus et spes in te** sperantium virgo ante
partum virgo in partu virgo post partum
fons salutis et gratie fons pietatis et leticie fons consolationis
et indulgencie **Et** per illam sanctam inestimabilem leticiam
qua exultavit spiritus tuus in illa hora quando tibi per gabrielem
archangelum annunciatus **et conceptus** filius dei fuit
Et per illud divinum mysterium quod tunc operatus est spiritus sanctus **in te**

Hazem, Amir, et al. "Towards automatic variant analysis of ancient devotional texts." Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. 2019.

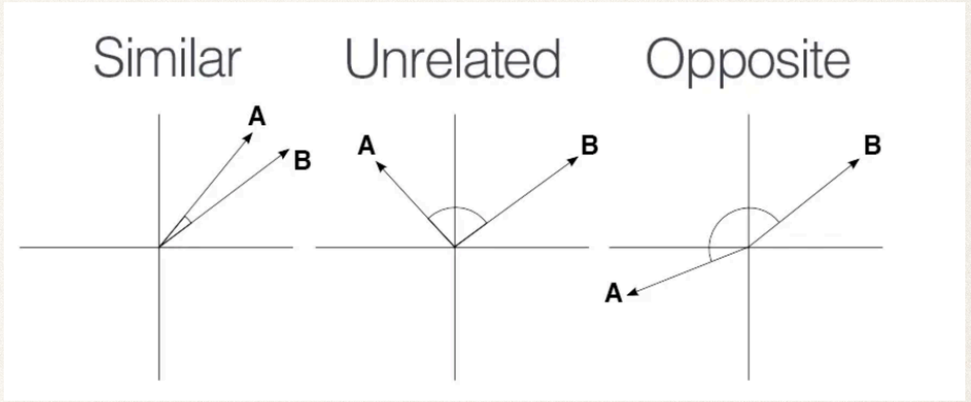
Documents as vectors

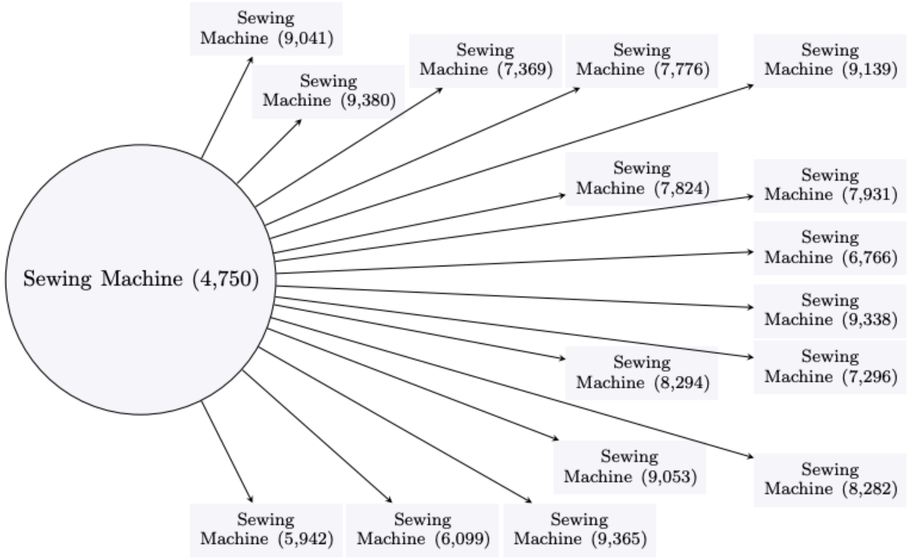
- Let d be a document, then $d \in \mathbb{R}^{|V|}$
- If D is a set of documents, then $D = [d^{(1)}; d^{(2)}; \cdots; d^{(|D|)}] \in \mathbb{R}^{|D| \times |V|}$

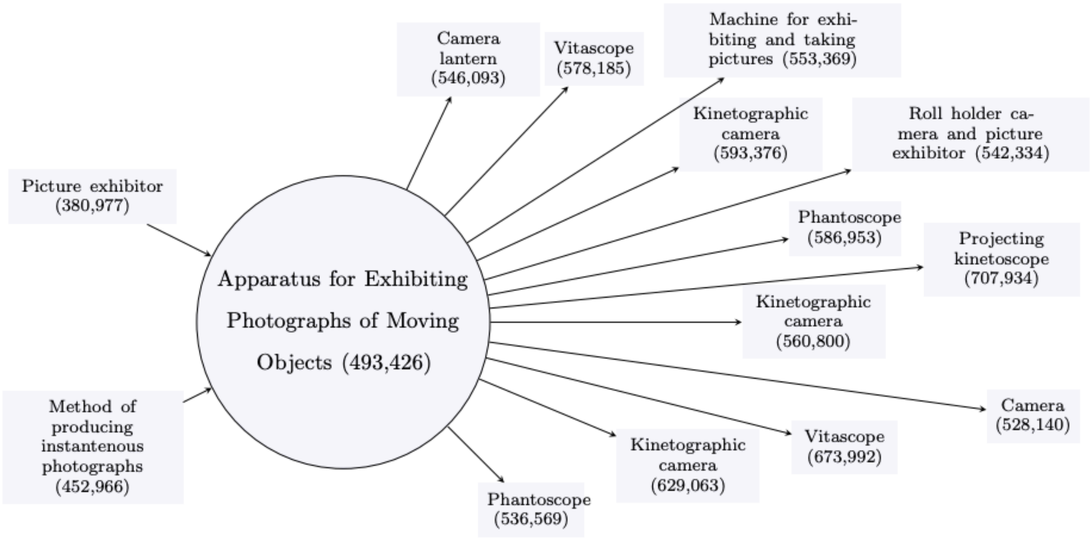


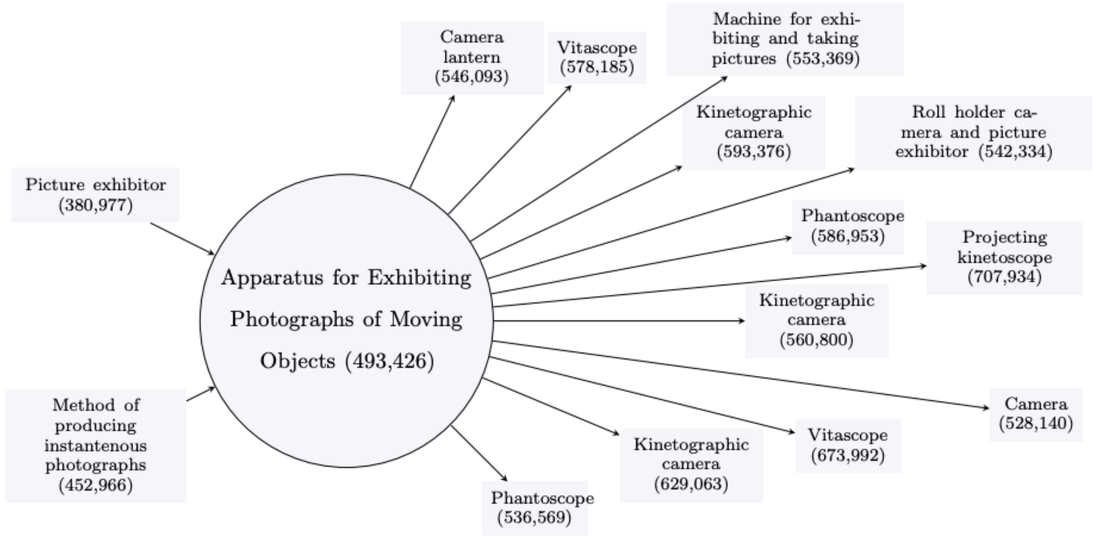
Cosine similarity

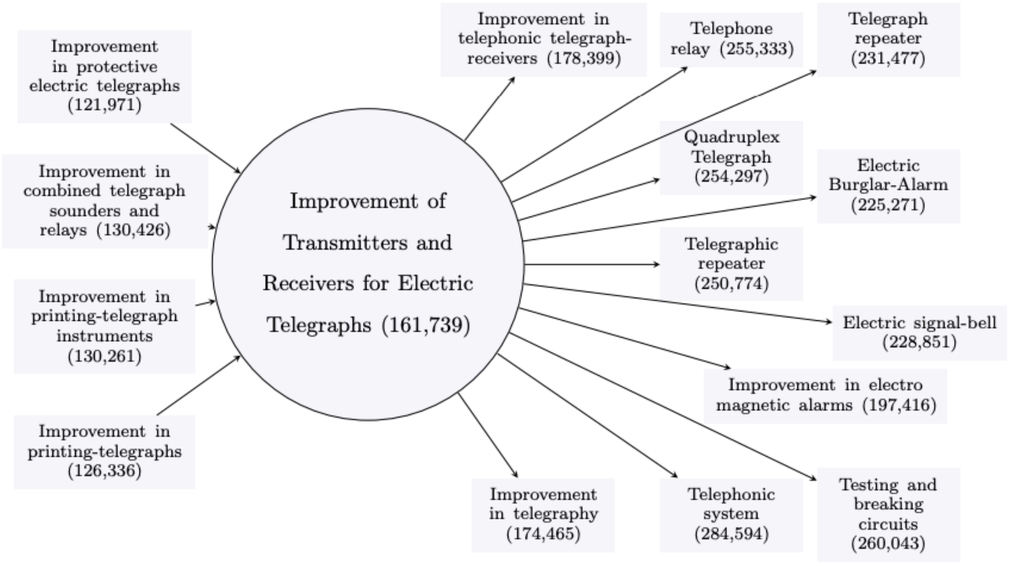
- Objective: Measure similarity by treating documents as vectors
- CosSim (Doc1, Doc2) = $\frac{d^{(1)} \cdot d^{(2)}}{\|d^{(1)}\| \|d^{(2)}\|}$




Sewing Machine (4,750)
Sewing Machine (9,041)
Sewing Machine (9,380)
Sewing Machine (7,369)
Sewing Machine (7,776)
Sewing Machine (9,139)
Sewing Machine (7,824)
Sewing Machine (7,931)
Sewing Machine (6,766)
Sewing Machine (9,338)
Sewing Machine (8,294)
Sewing Machine (7,296)
Sewing Machine (9,053)
Sewing Machine (8,282)
Sewing Machine (5,942)
Sewing Machine (6,099)
Sewing Machine (9,365)


Apparatus for Exhibiting Photographs of Moving Objects (493,426)
Picture exhibitor (380,977)
Method of producing instantenous photographs (452,966)
Camera lantern (546,093)
Vitascope (578,185)
Machine for exhibiting and taking pictures (553,369)
Kinetographic camera (593,376)
Roll holder camera and picture exhibitor (542,334)
Phantoscope (586,953)
Projecting kinetoscope (707,934)
Kinetographic camera (560,800)
Camera (528,140)
Vitascope (673,992)
Kinetographic camera (629,063)
Phantoscope (536,569)


Apparatus for Exhibiting Photographs of Moving Objects (493,426)
Picture exhibitor (380,977)
Method of producing instantenous photographs (452,966)
Camera lantern (546,093)
Vitascope (578,185)
Machine for exhibiting and taking pictures (553,369)
Kinetographic camera (593,376)
Roll holder camera and picture exhibitor (542,334)
Phantoscope (586,953)
Projecting kinetoscope (707,934)
Kinetographic camera (560,800)
Camera (528,140)
Vitascope (673,992)
Kinetographic camera (629,063)
Phantoscope (536,569)


Improvement of Transmitters and Receivers for Electric Telegraphs (161,739)
Improvement in protective electric telegraphs (121,971)
Improvement in combined telegraph sounders and relays (130,426)
Improvement in printing-telegraph instruments (130,261)
Improvement in printing-telegraphs (126,336)
Improvement in telephonic telegraph-receivers (178,399)
Telephone relay (255,333)
Telegraph repeater (231,477)
Quadruplex Telegraph (254,297)
Electric Burglar-Alarm (225,271)
Telegraphic repeater (250,774)
Electric signal-bell (228,851)
Improvement in electro magnetic alarms (197,416)
Testing and breaking circuits (260,043)
Telephonic system (284,594)
Improvement in telegraphy (174,465)

Distance metrics

- Many ways to compare documents. A few examples:
  - Edit based (e.g., Levenshtein distance)
  - Token based (e.g., Jaccard index)
  - Vector based (e.g., Cosine similarity)

Metadata

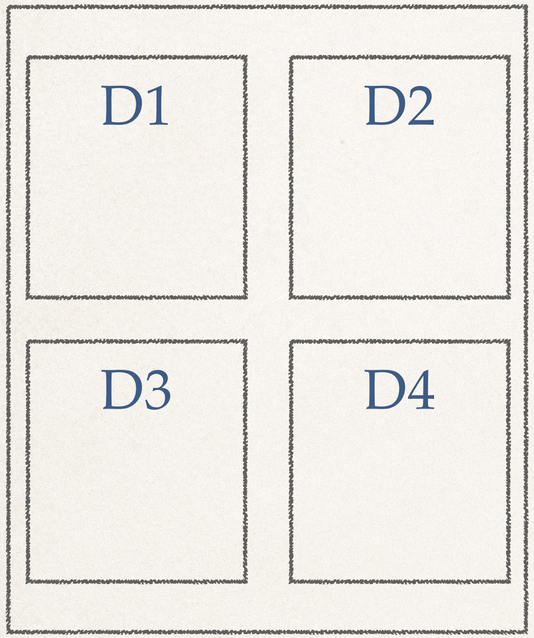
Metadata

- Documents have metadata (e.g., author, date, genre)

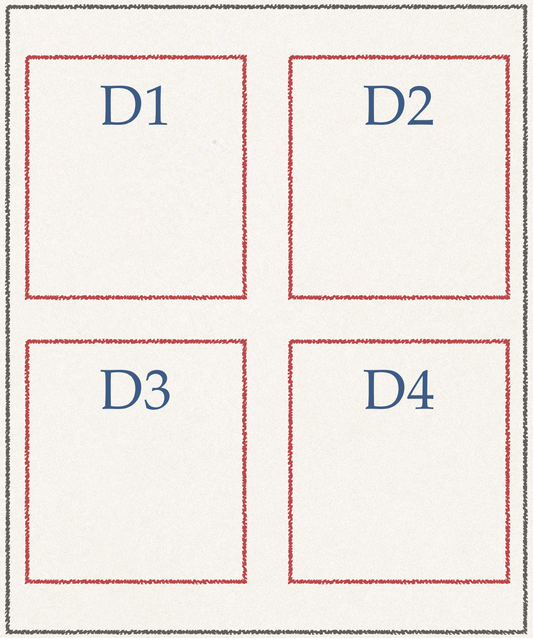
Metadata

- Documents have metadata (e.g., author, date, genre)
- How to quantify differences between groups?

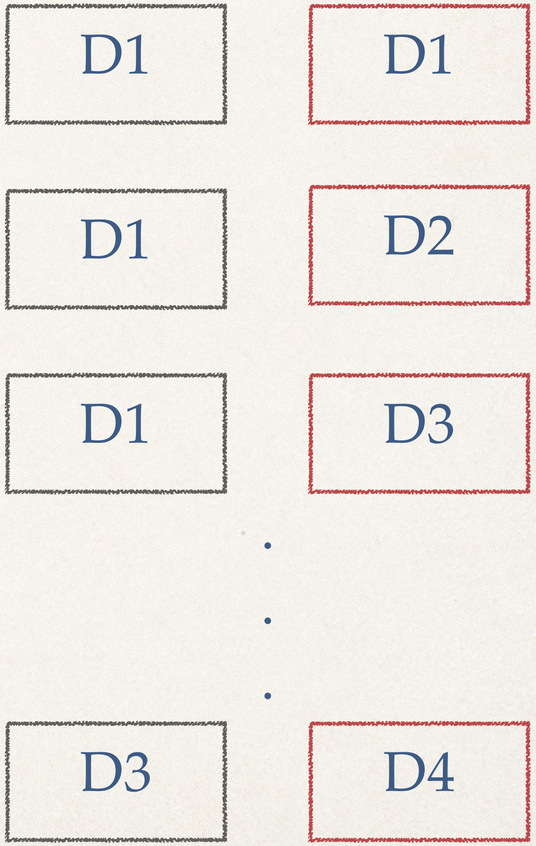
Pairwise Metrics



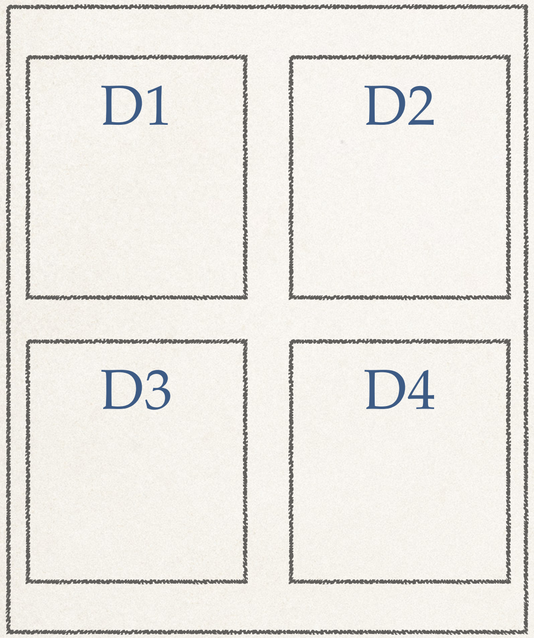
Corpus for group 1



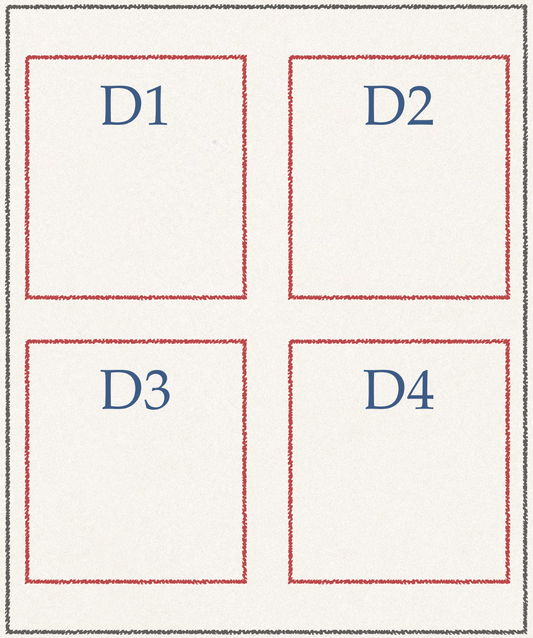
Corpus for group 2



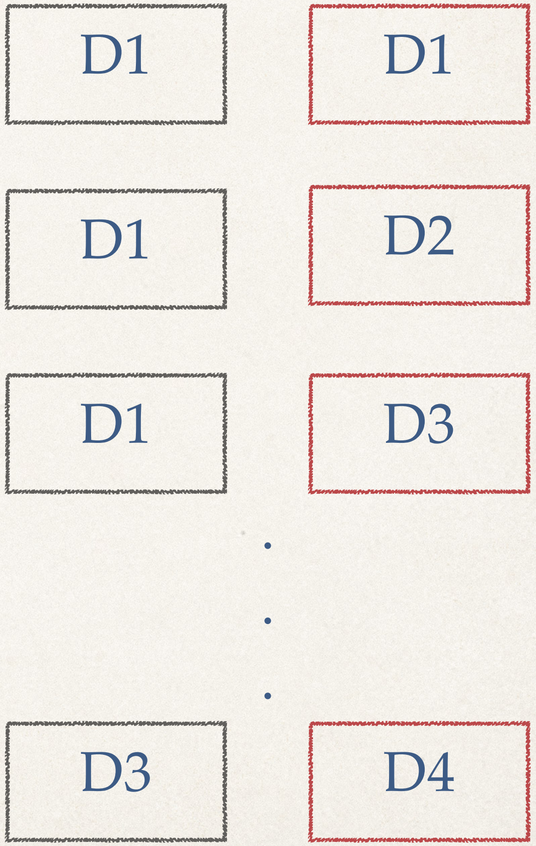
Pairwise Metrics



Corpus for group 1



Corpus for group 2



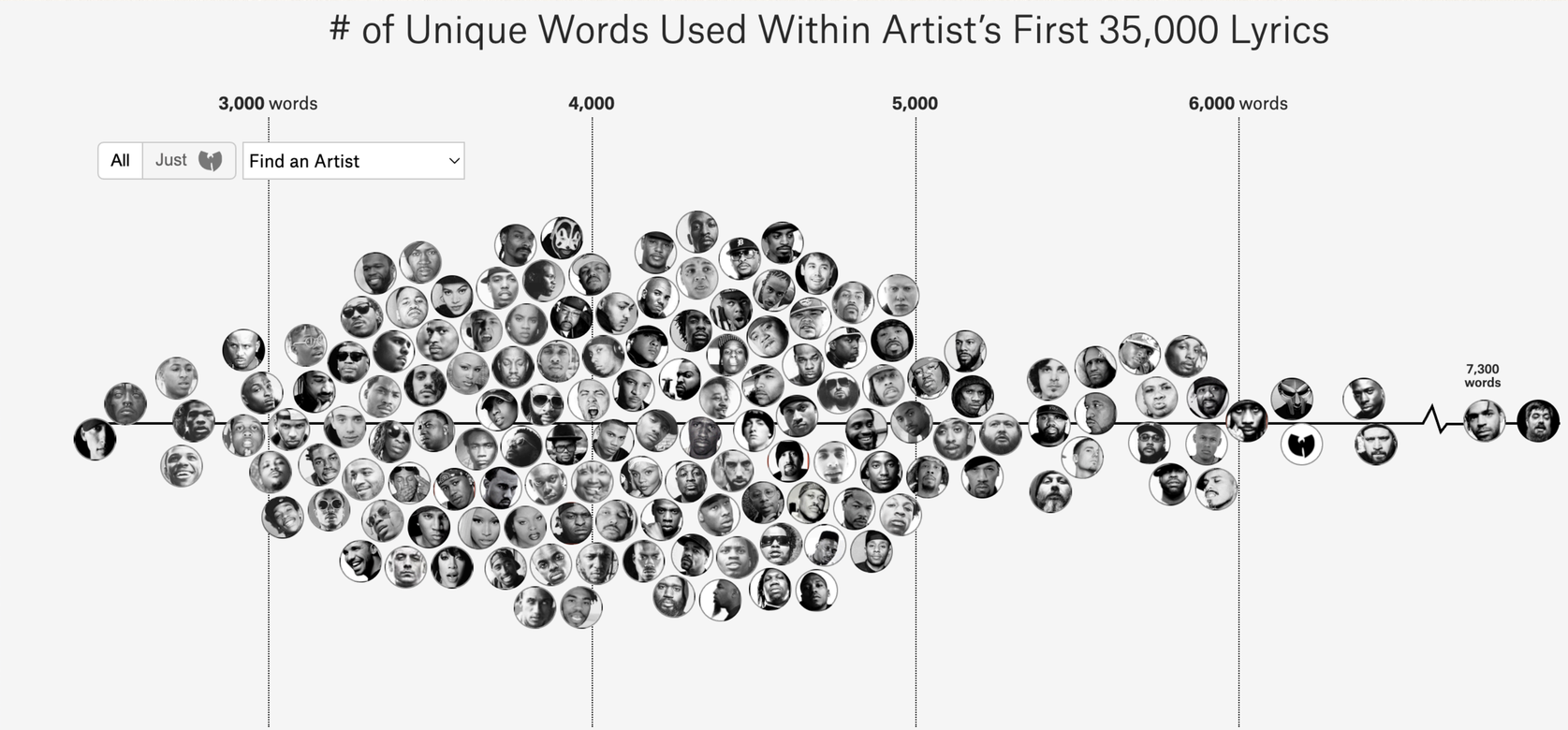
- Average similarity between document pairs as measure of similarity between groups

metrics.pairwise.additive_chi2_kernel(X[, Y])	Compute the additive chi-squared kernel between observations in X and Y.
metrics.pairwise.chi2_kernel(X[, Y, gamma])	Compute the exponential chi-squared kernel between X and Y.
metrics.pairwise.cosine_similarity(X[, Y, ...])	Compute cosine similarity between samples in X and Y.
metrics.pairwise.cosine_distances(X[, Y])	Compute cosine distance between samples in X and Y.
metrics.pairwise.distance_metrics()	Valid metrics for pairwise_distances.
metrics.pairwise.euclidean_distances(X[, Y, ...])	Compute the distance matrix between each pair from a vector array X and Y.
metrics.pairwise.haversine_distances(X[, Y])	Compute the Haversine distance between samples in X and Y.
metrics.pairwise.kernel_metrics()	Valid metrics for pairwise_kernels.
metrics.pairwise.laplacian_kernel(X[, Y, gamma])	Compute the laplacian kernel between X and Y.
metrics.pairwise.linear_kernel(X[, Y, ...])	Compute the linear kernel between X and Y.
metrics.pairwise.manhattan_distances(X[, Y, ...])	Compute the L1 distances between the vectors in X and Y.
metrics.pairwise.nan_euclidean_distances(X)	Calculate the euclidean distances in the presence of missing values.
metrics.pairwise.pairwise_kernels(X[, Y, ...])	Compute the kernel between arrays X and optional array Y.
metrics.pairwise.polynomial_kernel(X[, Y, ...])	Compute the polynomial kernel between X and Y.
metrics.pairwise.rbf_kernel(X[, Y, gamma])	Compute the rbf (gaussian) kernel between X and Y.
metrics.pairwise.sigmoid_kernel(X[, Y, ...])	Compute the sigmoid kernel between X and Y.
metrics.pairwise.paired_euclidean_distances(X, Y)	Compute the paired euclidean distances between X and Y.
metrics.pairwise.paired_manhattan_distances(X, Y)	Compute the paired L1 distances between X and Y.
metrics.pairwise.paired_cosine_distances(X, Y)	Compute the paired cosine distances between X and Y.
metrics.pairwise.paired_distances(X, Y, *[, ...])	Compute the paired distances between X and Y.
metrics.pairwise_distances(X[, Y, metric, ...])	Compute the distance matrix from a vector array X and optional Y.
metrics.pairwise_distances_argmin(X, Y, *[, ...])	Compute minimum distances between one point and a set of points.
metrics.pairwise_distances_argmin_min(X, Y, *)	Compute minimum distances between one point and a set of points.
metrics.pairwise_distances_chunked(X[, Y, ...])	Generate a distance matrix chunk by chunk with optional reduction.

sklearn.metrics.pairwise.chi2_kernel

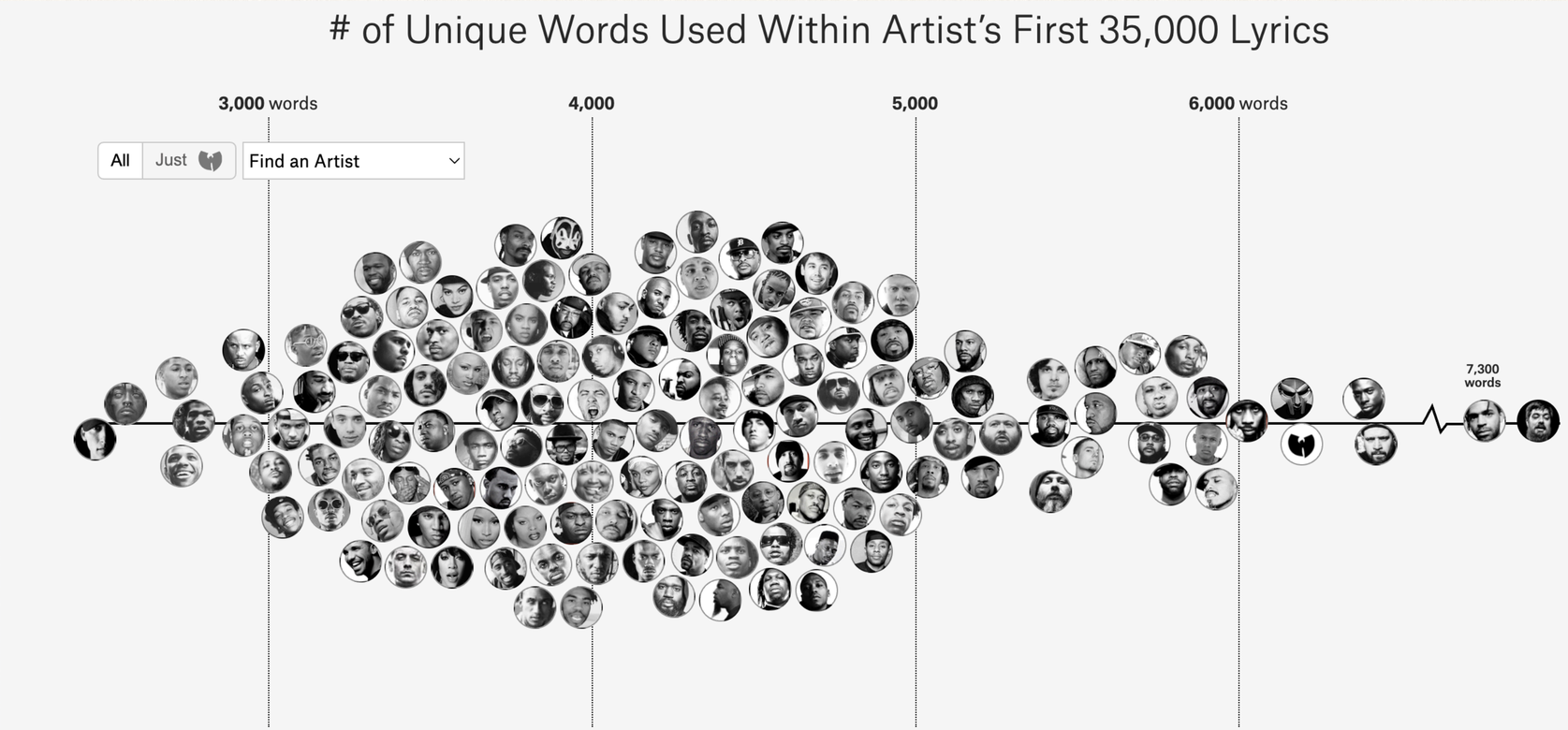
The different pairwise metrics in sklearn

Finding Differentiators



Source: https://pudding.cool/projects/vocabulary/index.html

Finding Differentiators



Source: https://pudding.cool/projects/vocabulary/index.html

We are often interested in finding the difference markers

Difference in proportions

- If w is the word, k is the category and c_w is the count
- $f_{w,k} = \frac{C_{w,k}}{\sum_{w'} C_{w',k}}$ is the normalized count or proportion of occurrence of the word w in category k
- $f_{w,k_1} - f_{w,k_2}$ gives the difference in proportions across two groups

Difference in proportions

- Simple and easy to measure and interpret
- Overemphasizes common words; for common words, there differences are also large
- No correction for chance or determination of statistical significance

$$x^2$$

χ^2

Does the word “robot” occur significantly more frequently in science fiction?

χ^2

Does the word “robot” occur significantly more frequently in science fiction?

	robot	¬robot	
sci-fi	104	1004	= 10.3%
¬sci-fi	2	13402	= 0.015%

Slide credit: David Bamman’s Info 256 class

χ^2

We can calculate the following statistic, which is the sum of squared difference between the observed value in each cell and the expected value assuming independence

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Slide adapted from David Bamman's class

χ^2

	robot	¬robot	sum	frequency
sci-fi	104	1004	1108	0.076
¬sci-fi	2	13402	13404	0.924
sum	106	14406		
frequency	0.007	0.993		

Assuming independence:

$$P(\text{robot}, \text{scifi}) = P(\text{robot}) \times P(\text{scifi})$$
$$= 0.007 \times 0.076 = 0.00053$$

Among 14512 words, we would expect to see 7.69 occurrences of *robot* in sci-fi texts.

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

$P(\text{scifi})$	0.076
$P(\neg\text{scifi})$	0.924

$P(\text{robot})$	$P(\neg\text{robot})$
0.007	0.993

Slide adapted from David Bamman's class

χ^2

	robot	¬robot
sci-fi	104	1004
¬sci-fi	2	13402

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

Left is observed counts; right is expected counts assuming complete independence

Slide adapted from David Bamman's class

χ^2

How different are these two tables?

	robot	¬robot
sci-fi	104	1004
¬sci-fi	2	13402

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

Left is observed counts; right is expected counts assuming complete independence

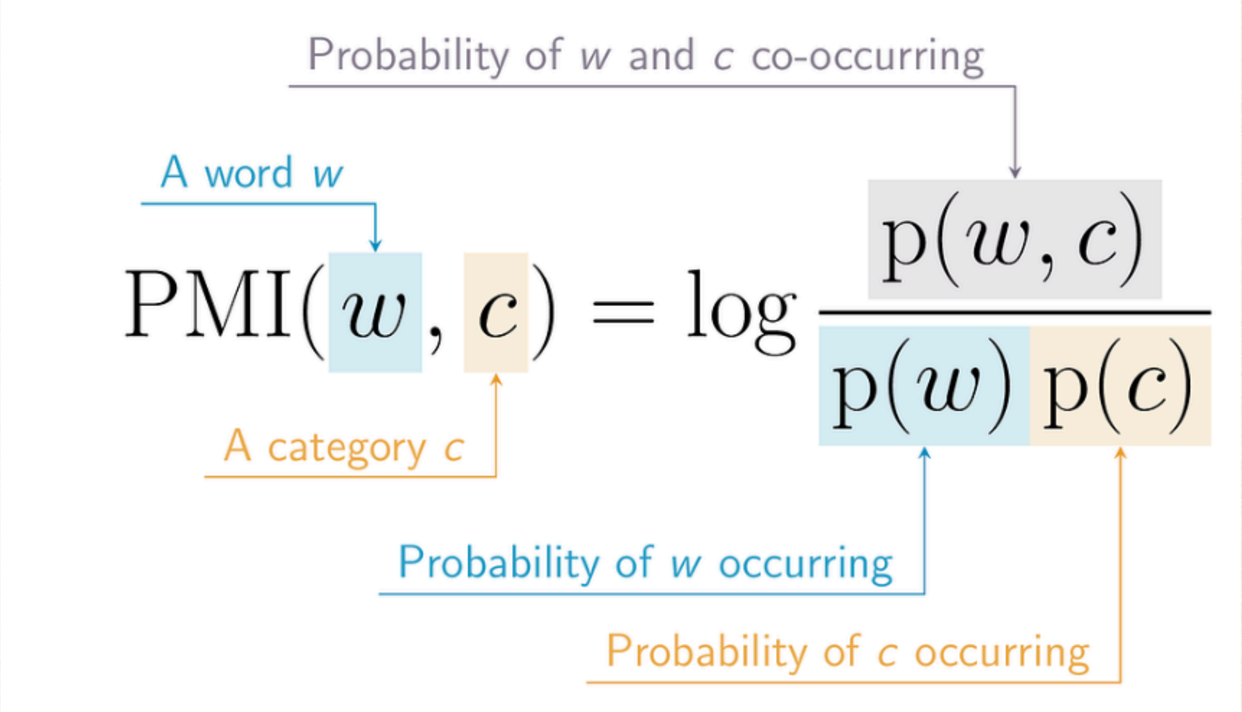
Slide adapted from David Bamman's class

χ^2

- Useful statistic to find differentiating markers
- We have a way to test for statistical significance
- Assumes each word is independent from the others

PMI

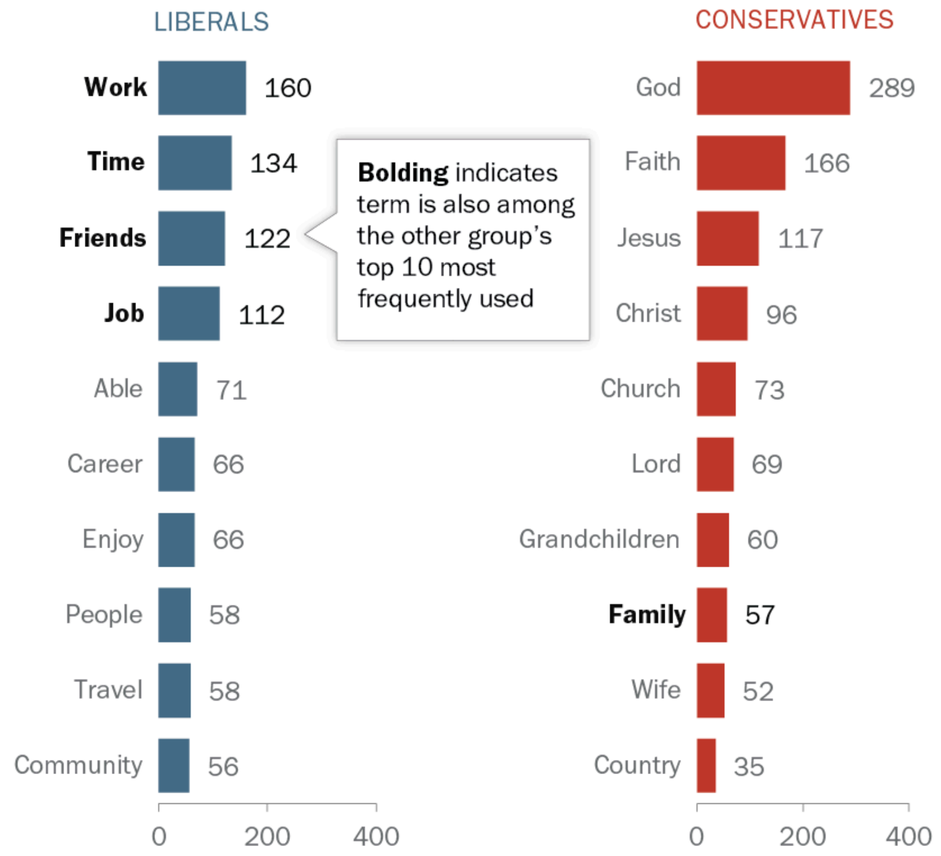
- A more information theoretic approach
- Generalization to more than 2 categories
- Statistical significance can be calculated using non-parametric methods



Source: Bestvater and Shah; Pew research article

Identifying terms used more often by one group than another doesn't always indicate distinctivenes

*Top 10 terms used **more** frequently by ___, by difference in word count*

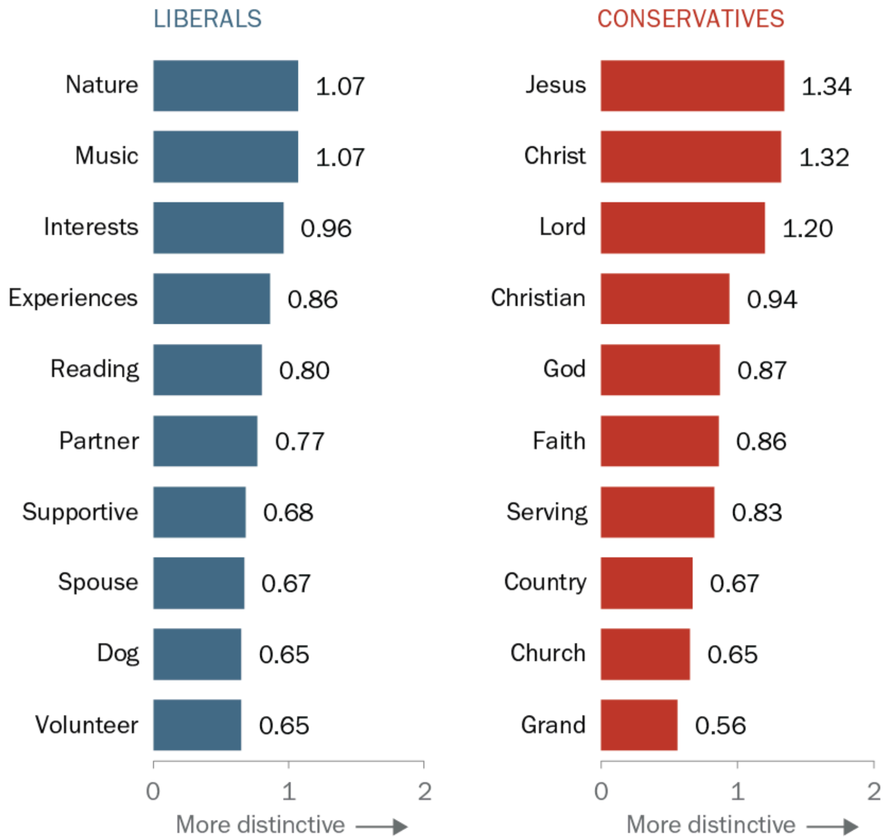


Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

Pointwise mutual information helps identify words that are used distinctively by one group, not just used more

Most dinstinctive terms used by each group, by PMI score



Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

Pointwise mutual information emphasizes different words than simple term frequencies or frequency differences

Top 10 terms used by conservatives, as measured by ...

Rank	Frequency	Difference in frequency	PMI
1	Family	God	Jesus
2	God	Faith	Christ
3	Friends	Jesus	Lord
4	Children	Christ	Christian
5	Health	Church	God
6	Love	Lord	Faith
7	Time	Grandchildren	Serving
8	Work	Family	Country
9	Faith	Wife	Church
10	Job	Country	Grand
...			
60			Family

Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

Pointwise mutual information emphasizes different words than simple term frequencies or frequency differences

Top 10 terms used by conservatives, as measured by ...

Rank	Frequency	Difference in frequency	PMI
1	Family	God	Jesus
2	God	Faith	Christ
3	Friends	Jesus	Lord
4	Children	Christ	Christian
5	Health	Church	God
6	Love	Lord	Faith
7	Time	Grandchildren	Serving
8	Work	Family	Country
9	Faith	Wife	Church
10	Job	Country	Grand
...			
60			Family

Notes: Terms used in open-ended responses to the question, "What about your life do you currently find meaningful, fulfilling, or satisfying? What keeps you going, and why?" Reported numbers are unweighted.
Source: Survey conducted Sept. 14-28, 2017, among U.S. adults.

PEW RESEARCH CENTER

Other methods

- Many other methods to characterize differences
  - Model based methods that assume parametric distributions and Bayesian priors (e.g., Monroe et. al. 2009)
  - Unsupervised and Bayesian (e.g., SAGE; Eisenstein et. al. 2011)
  - Supervised learning to learn precise features that are informative in separating categories (e.g., Underwood et. al. 2018)

In class

- Finding distinctive terms demo