Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

October 23rd, 2014

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
    - Select a clustering model, Characterize Model Fit
    - Choose the number of components for our mixture
2) Objective function:
    - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

    - Substantively Θ:
        - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
        - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
        - The mathematical "groupings" align with meaningful groupings
3) Optimization
    - Select the best model.
        - Run several candidate models$\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.
     - Run several candidate models$\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
     - Stats + Substance to select model + K

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} = f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.
     - Run several candidate models$\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
     - Stats + Substance to select model + K
4) Validation

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} = f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.
     - Run several candidate models$\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
     - Stats + Substance to select model + K
4) Validation
   - Is our statistic capturing what we want from the clustering?

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} = f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.
     - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
     - Stats + Substance to select model + K
4) Validation
   - Is our statistic capturing what we want from the clustering?
   - Are there features we're missing

Interpreting Clusterings + Computer Assisted Clusterings

1) Task:
   - Select a clustering model, Characterize Model Fit
   - Choose the number of components for our mixture
2) Objective function:
   - Mathematical objective function

$$\text{Math Obj} \quad = \quad f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta})$$

   - Substantively Θ:
     - Cohesive: words that are prominent in $\boldsymbol{\theta}_k$ actually occur together
     - Exclusive: words that are featured in $\boldsymbol{\theta}_k$ only occur in k
     - The mathematical "groupings" align with meaningful groupings
3) Optimization
   - Select the best model.
     - Run several candidate models $\rightsquigarrow$ optimize $\boldsymbol{\Theta}$ and $\boldsymbol{T}$
     - Stats + Substance to select model + K
4) Validation
   - Is our statistic capturing what we want from the clustering?
   - Are there features we're missing
   - Very Open Research Question

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

A Motivating Clustering Model$\leadsto$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

A Motivating Clustering Model$\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^{*} = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\boldsymbol{\tau}_i \sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}}$$

A Motivating Clustering Model$\leadsto$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}\boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\ \boldsymbol{x}_i^*|\tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}\end{aligned}$$

A Motivating Clustering Model$\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^{*} = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}
\boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\
\boldsymbol{x}_i^{*}|\tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}
\end{aligned}$$

Provides:

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned} \boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\ \boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}} \end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment

A Motivating Clustering Model$\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}
\boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\
\boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}
\end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment
- $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \rightsquigarrow$ Proportion of documents in each component

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* \quad = \quad \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned}
\boldsymbol{\tau}_i \quad &\sim \quad \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\
\boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k \quad &\sim \quad \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}}
\end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment
- $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \rightsquigarrow$ Proportion of documents in each component
- $\boldsymbol{\mu}_k \rightsquigarrow$ Exemplar document for cluster k

A Motivating Clustering Model $\rightsquigarrow$ Mixture of von Mises Fisher Distributions

J element long unit-length vector

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i}{\sqrt{\boldsymbol{x}_i^{'}\boldsymbol{x}_i}}$$

Mixture of von Mises-Fisher (vMF) distributions:

$$\begin{aligned} \boldsymbol{\tau}_i &\sim \overbrace{\text{Multinomial}(1, \boldsymbol{\pi})}^{\text{Mixture component}} \\ \boldsymbol{x}_i^* | \tau_{ik} = 1, \boldsymbol{\mu}_k &\sim \underbrace{\text{vMF}(\kappa, \boldsymbol{\mu}_k)}_{\text{Language model}} \end{aligned}$$

Provides:

- $\boldsymbol{\tau}_i \rightsquigarrow$ Each document's cluster assignment
- $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \rightsquigarrow$ Proportion of documents in each component
- $\boldsymbol{\mu}_k \rightsquigarrow$ Exemplar document for cluster k

EM algorithm in slides appendix

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?$\rightsquigarrow$ predict new documents?

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?$\rightsquigarrow$ predict new documents?
Problem

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?$\rightsquigarrow$ predict new documents?
Problem$\rightsquigarrow$ in sample evaluation leads to overfit.

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?$\rightsquigarrow$ predict new documents?
Problem$\rightsquigarrow$ in sample evaluation leads to overfit.
Solution$\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^*_{\text{out}}$

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^{*}_{\text{out}}$

$$\log p(\boldsymbol{x}^{*}_{\text{out}}|\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{X}) \quad = \quad \log \sum_{k=1}^{K} p(\boldsymbol{x}^{*}_{\text{out}}, \tau_{ik}|\boldsymbol{\mu}_k, \boldsymbol{\pi}, \boldsymbol{X})$$

Measuring Cluster Performance: Out of Sample Prediction

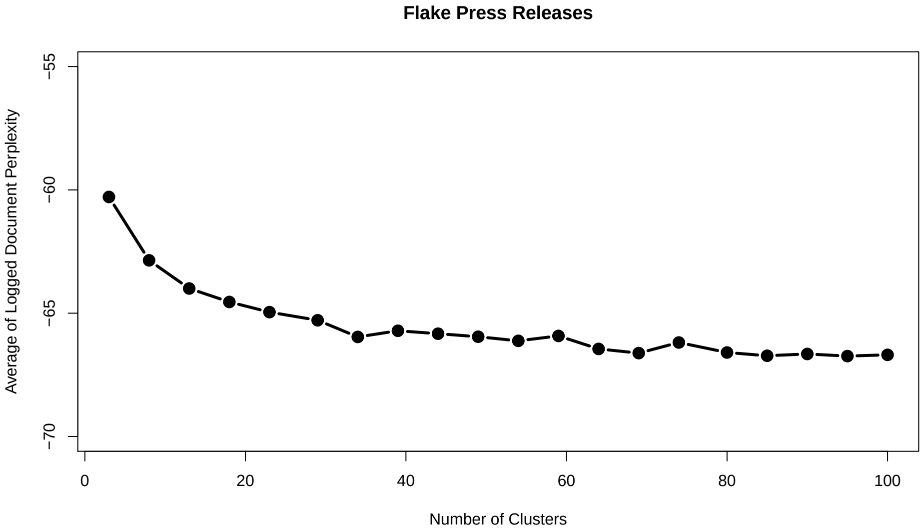
How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^{*}_{\text{out}}$

$$
\begin{aligned}
\log p(\boldsymbol{x}^{*}_{\text{out}}|\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{X}) &= \log \sum_{k=1}^{K} p(\boldsymbol{x}^{*}_{\text{out}}, \tau_{ik}|\boldsymbol{\mu}_k, \boldsymbol{\pi}, \boldsymbol{X}) \\
&= \log \sum_{k=1}^{K} \left[\pi_k \exp(\kappa \boldsymbol{\mu}_k^{'} \boldsymbol{x}^{*}_{\text{out}})\right]
\end{aligned}
$$

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?$\leadsto$ predict new documents?
Problem$\leadsto$ in sample evaluation leads to overfit.
Solution$\leadsto$ evaluate performance on held out data
For held out document $\boldsymbol{x}_{\text{out}}^{*}$

$$
\begin{aligned}
\log p(\boldsymbol{x}_{\text{out}}^{*}|\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{X}) &= \log \sum_{k=1}^{K} p(\boldsymbol{x}_{\text{out}}^{*}, \tau_{ik}|\boldsymbol{\mu}_k, \boldsymbol{\pi}, \boldsymbol{X}) \\
&= \log \sum_{k=1}^{K} \left[\pi_k \exp(\kappa \boldsymbol{\mu}_k^{'} \boldsymbol{x}_{\text{out}}^{*})\right] \\
\text{Perplexity}_{\text{word}} &= \exp\left(- \log p(\boldsymbol{x}_{\text{out}}^{*}|\boldsymbol{\mu}, \boldsymbol{\pi})\right)
\end{aligned}
$$


Flake Press Releases
Average of Logged Document Perplexity
−55
−60
−65
−70
0
20
40
60
80
100
Number of Clusters

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction$\rightsquigarrow$ One Task
- Do we care about it?

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction$\rightsquigarrow$ One Task
- Do we care about it?$\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy $\rightsquigarrow$ measure quality in topics and clusters

(Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task
- Do we care about it? $\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy $\rightsquigarrow$ measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al AJPS Forthcoming)

What's Prediction Got to Do With It?

- Prediction$\rightsquigarrow$ One Task
- Do we care about it?$\rightsquigarrow$ Social science application where we're predicting new texts?
- Does it correspond to how we might use the model?

Chang et al 2009 ("Reading the Tea Leaves") :

- Compare perplexity with human based evaluations
- NEGATIVE relationship between perplexity and human based evaluations

Different strategy$\rightsquigarrow$ measure quality in topics and clusters

- Statistics: measure cohesiveness and exclusivity (Roberts, et al AJPS Forthcoming)
- Experiments: measure topic and cluster quality

Measuring Cohesiveness and Exclusivity

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic? $\rightsquigarrow$ will see these words co-occur in documents

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic? $\rightsquigarrow$ will see these words co-occur in documents
- Define $\boldsymbol{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ be the top words for a topic

Measuring Cohesiveness and Exclusivity

- Consider the output of a 3-component mixture of model (say, Multinomials or von Mises-Fisher models)
- We might select 5 top words for each topic

Topic 1	bill	congressman	earmarks	following	house
Topic 2	immigration	reform	security	border	worker
Topic 3	earmark	egregious	pork	fiscal	today

- An ideal topic?$\rightsquigarrow$ will see these words co-occur in documents
- Define $\boldsymbol{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ be the top words for a topic
- For example $\boldsymbol{v}_3 =$ (earmark , egregious , pork , fiscal , today)

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$D(\text{earmark}, \text{egregious}) \quad = \quad \text{No. times earmark and egregious co-occur}$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$
\begin{aligned}
D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\
D(\text{egregious}) &= \text{Number of times Egregious occurs}
\end{aligned}
$$

Define cohesiveness for topic k as

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$
\begin{aligned}
D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\
D(\text{egregious}) &= \text{Number of times Egregious occurs}
\end{aligned}
$$

Define cohesiveness for topic k as

$$
\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log\left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})}\right)
$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic k as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)$$

Define overall cohesiveness as:

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic k as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log\left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})}\right)$$

Define overall cohesiveness as:

$$\text{Cohesive} = (\sum_{k=1}^{K} \text{Cohesive}_k)/K$$

Measuring Cohesiveness and Exclusivity

Define the function D as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic k as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log\left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})}\right)$$

Define overall cohesiveness as:

$$\begin{aligned} \text{Cohesive} &= \left(\sum_{k=1}^{K} \text{Cohesive}_k\right)/K \\ &= \left(\sum_{k=1}^{K} \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log\left(\frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})}\right)\right)/K \end{aligned}$$

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ few replicates of each topic

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\leadsto$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

$$\text{Exclusivity}_k = \sum_{j: v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}}$$

Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive $\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

$$\begin{aligned}
\text{Exclusivity}_k &= \sum_{j: v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \\
\text{Exclusivity} &= \left(\sum_{k=1}^{K} \text{Exclusivity}_k \right) / K
\end{aligned}$$

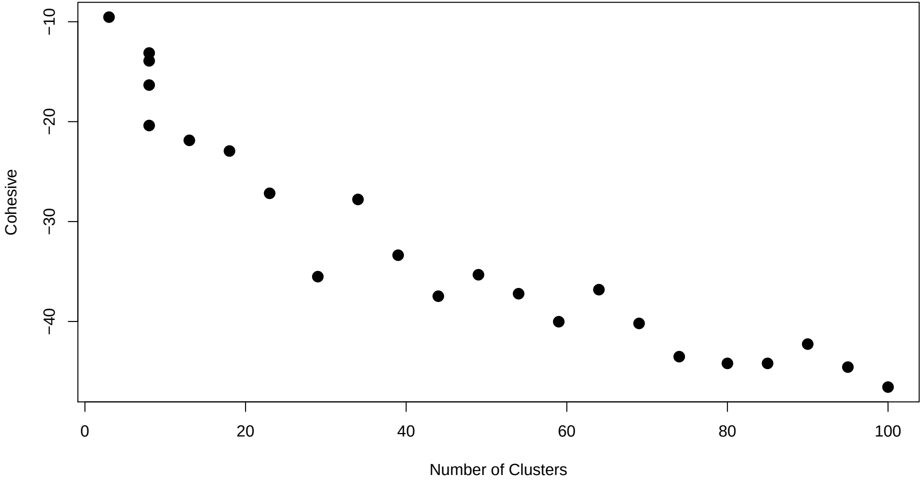
Measuring Cohesiveness and Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

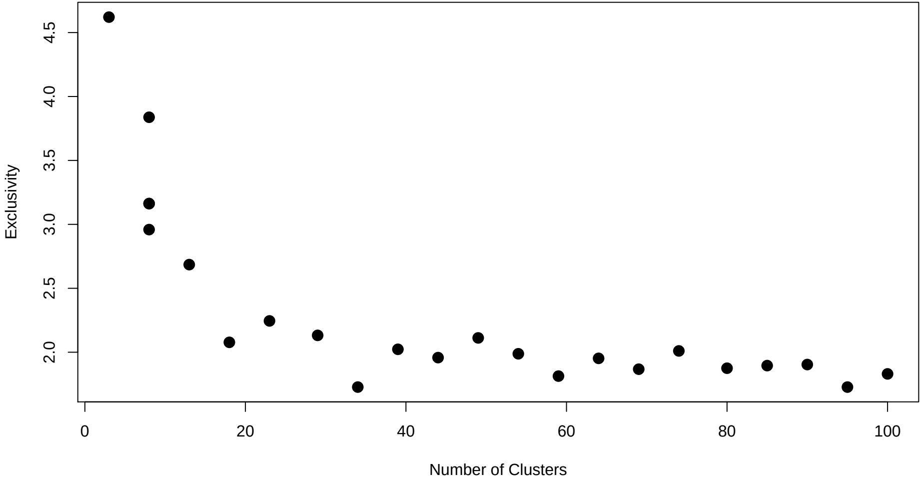
$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

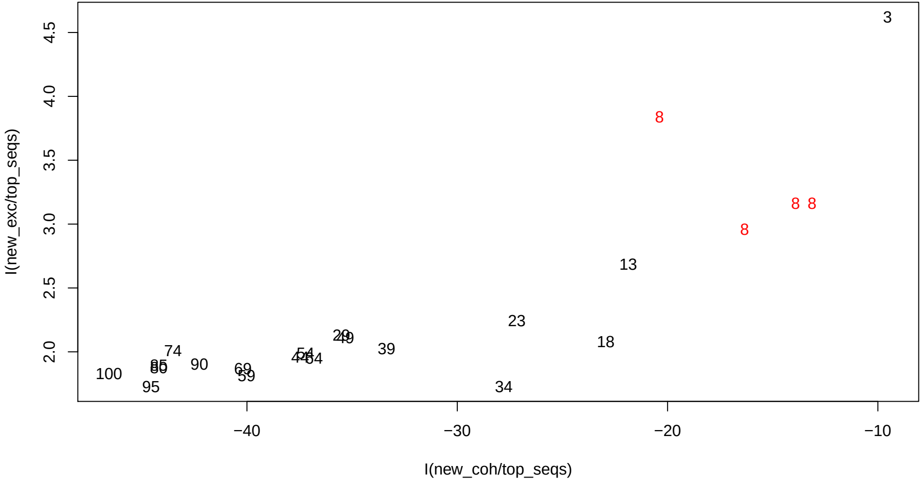
Suppose again we pick L top words. Measure Exclusivity for a topic as for a topic as:

$$\begin{aligned}
\text{Exclusivity}_k &= \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \\
\text{Exclusivity} &= \left(\sum_{k=1}^{K} \text{Exclusivity}_k \right) / K \\
&= \left(\sum_{k=1}^{K} \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \right) / K
\end{aligned}$$



Cohesive
−10
−20
−30
−40
0
20
40
60
80
100
Number of Clusters


Exclusivity
4.5
4.0
3.5
3.0
2.5
2.0
0
20
40
60
80
100
Number of Clusters


I(new_exc/top_seqs)
I(new_coh/top_seqs)
4.5
4.0
3.5
3.0
2.5
2.0
−40
−30
−20
−10
3
8
8
8
8
13
23
18
39
34
74
90
100
95
64

Experimental Approaches

Mathematical approaches

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model $\rightsquigarrow$ including text representation

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers

assumes we're in the model $\rightsquigarrow$ including text representation

Humans $\rightsquigarrow$ read texts

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model $\rightsquigarrow$ including text representation
Humans $\rightsquigarrow$ read texts
Humans $\rightsquigarrow$ use cluster output

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model $\rightsquigarrow$ including text representation
Humans $\rightsquigarrow$ read texts
Humans $\rightsquigarrow$ use cluster output
Do humans think the model is performing well?

Experimental Approaches

Mathematical approaches $\rightsquigarrow$ suppose we can capture quality with numbers
assumes we're in the model $\rightsquigarrow$ including text representation
Humans $\rightsquigarrow$ read texts
Humans $\rightsquigarrow$ use cluster output
Do humans think the model is performing well?

1) Topic Quality

Experimental Approaches

Mathematical approaches$\leadsto$ suppose we can capture quality with numbers
assumes we're in the model$\leadsto$ including text representation
Humans$\leadsto$ read texts
Humans$\leadsto$ use cluster output
Do humans think the model is performing well?

1) Topic Quality
2) Cluster Quality

Experimental Approaches

1) Take M top words for a topic
2) Randomly select a top word from another topic
   2a) Sample the topic number from l from $K-1$ (uniform probability)
   2b) Sample word j from the M top words in topic l
   2c) Permute the words and randomly insert the intruder:
      - List:

$$\text{test} = (v_{k,3}, v_{k,1}, v_{l,j}, v_{k,2}, v_{k,4}, v_{k,5})$$

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`bowl, flooding, olympic, olympics, nfl, coach`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`bowl, flooding, olympic, olympics, nfl, coach`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`stocks, investors, fed, guns, trading, earning`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

`stocks, investors, fed, guns, trading, earning`

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

Higher rate of intruder identification $\rightsquigarrow$ more exclusive/cohesive topics

Example Experiment: Word Intrusion (Weiss and Grimmer, In Progress)

Higher rate of intruder identification $\rightsquigarrow$ more exclusive/cohesive topics

Deploy on Mechanical Turk

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)
- Select clustering with highest cluster quality

Cluster Quality (Grimmer and King 2011)

Assessing Cluster Quality with experiments

- Goal: group together similar documents
- Who knows if similarity measure corresponds with semantic similarity

$\rightsquigarrow$ Inject human judgement on pairs of documents

Design to assess cluster quality

- Estimate clusterings
- Sample pairs of documents (hint: you only need to compare discrepant pairs)
- Scale: (1) unrelated, (2) loosely related, (3) closely related (richer instructions, based on thing you want to cluster on)
- Cluster Quality = mean(within cluster) - mean(between clusters)
- Select clustering with highest cluster quality
- Can be used to compare any clusterings, regardless of source

How do we Choose K?

Generate many candidate models

1) Assess Cohesiveness/Exclusivity, select models on frontier
2) Use experiments
3) Read
4) Final decision $\rightsquigarrow$ combination

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation , agglomerative Hierarchical

Computer Assisted Clustering Methods

There are a lot of different clustering models (and many variations within each):
k-means , Mixture of multinomials , k-medoids , affinity propagation , agglomerative Hierarchical fuzzy k-means, trimmed k-means, k-Harmonic means, fuzzy k-medoids, fuzzy k modes, maximum entropy clustering, model based hierarchical (agglomerative), proximus, ROCK, divisive hierarchical, DISMEA, Fuzzy, QTClust, self-organizing map, self-organizing tree, unnormalized spectral, MS spectral, NJW Spectral, SM Spectral, Dirichlet Process Multinomial, Dirichlet Process Normal, Dirichlet Process von-mises Fisher, Mixture of von mises-Fisher (EM), Mixture of von Mises Fisher (VA), Mixture of normals, co-clustering mutual information, co-clustering SVD, LLAhclust, CLUES, bclust, c-shell, qtClustering, LDA, Express Agenda Model, Hierarchical Dirichlet process prior, multinomial, uniform process mulitinomial, Chinese Restaurant Distance Dirichlet process multinomial, Pitmann-Yor Process multinomial, LSA, ...

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method —

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,. . .
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,. . .
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
    - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
    - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy *k*-modes, affinity propagation, self-organizing maps,...
    - Well-defined statistical, data analytic, or machine learning foundations
    - How to add substantive knowledge: With few exceptions, unclear
    - The literature: little guidance on when methods apply
    - Deriving such guidance: difficult or impossible

The Problem with Fully Automated Clustering (Grimmer and King 2011)

- Large quantitative literature on cluster analysis
- The Goal — an optimal application-independent cluster analysis method — is mathematically impossible:
  - No free lunch theorem: every possible clustering method performs equally well on average over all possible substantive applications
- Existing methods:
  - Many choices: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps,...
  - Well-defined statistical, data analytic, or machine learning foundations
  - How to add substantive knowledge: With few exceptions, unclear
  - The literature: little guidance on when methods apply
  - Deriving such guidance: difficult or impossible

Deep problem in cluster analysis literature: full automation requires more information

Fully Automated $\rightarrow$ Computer Assisted (Grimmer and King 2011)

Fully Automated → Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models

Fully Automated $\rightarrow$ Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering

Fully Automated → Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
  - Easy (if you don't think about it): list all clustering, choose best

Fully Automated → Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
  - Easy (if you don't think about it): list all clustering, choose best
  - Impossible in Practice

Fully Automated $\rightarrow$ Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
    - Easy (if you don't think about it): list all clustering, choose best
    - Impossible in Practice
    - Solution: Organized list

Fully Automated $\rightarrow$ Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
  - Easy (if you don't think about it): list all clustering, choose best
  - Impossible in Practice
  - Solution: Organized list
  - Insight: Many clusterings are perceptually identical

Fully Automated $\rightarrow$ Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
  - Easy (if you don't think about it): list all clustering, choose best
  - Impossible in Practice
  - Solution: Organized list
  - Insight: Many clusterings are perceptually identical
  - Consider two clusterings of 10,000 documents, we move one document from 5 to 6.

Fully Automated $\rightarrow$ Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
  - Easy (if you don't think about it): list all clustering, choose best
  - Impossible in Practice
  - Solution: Organized list
  - Insight: Many clusterings are perceptually identical
  - Consider two clusterings of 10,000 documents, we move one document from 5 to 6.
- How to organize clusterings so humans can undestand?

Fully Automated → Computer Assisted (Grimmer and King 2011)

- Fully Automated Clustering may succeed, fails in general. Too hard to know when to apply models
- An alternative: Computer Assisted Clustering
  - Easy (if you don't think about it): list all clustering, choose best
  - Impossible in Practice
  - Solution: Organized list
  - Insight: Many clusterings are perceptually identical
  - Consider two clusterings of 10,000 documents, we move one document from 5 to 6.
- How to organize clusterings so humans can undestand?
- Our answer: a geography of clusterings

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
    - Introduce sampling methods to extend search beyond existing methods

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
   - New Clustering: weighted average of clusterings from methods

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
    - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
    - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
   - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)
7) $\rightsquigarrow$ Millions of clusterings easily comprehended

A New Strategy (Grimmer and King 2011)

1) Code text as numbers (in one *or more* of several ways)
2) Apply many different clustering methods to the data — each representing different (unstated) substantive assumptions
   - Introduce sampling methods to extend search beyond existing methods
3) Develop a metric between clusterings
4) Create a metric space of clusterings, and a 2-D projection
5) Introduce the local cluster ensemble to summarize any point, including points with no existing clustering
   - New Clustering: weighted average of clusterings from methods
6) Use animated visualization: use the local cluster ensemble to explore the space of clusterings (smoothly morphing from one into others)
7) $\rightsquigarrow$ Millions of clusterings easily comprehended
8) (Or, our new strategy: represent entire Bell space directly; no need to examine document contents)

Crosas, Grimmer, King, and Stewart $\rightsquigarrow$ Consilience

A brief live demonstration

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)

- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming

Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking

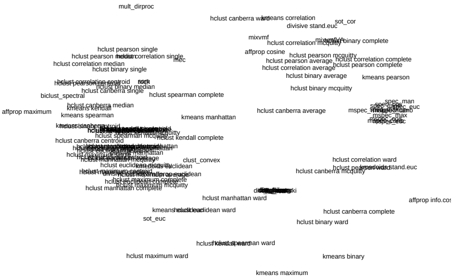
Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
  - Advertising
  - Credit Claiming
  - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)

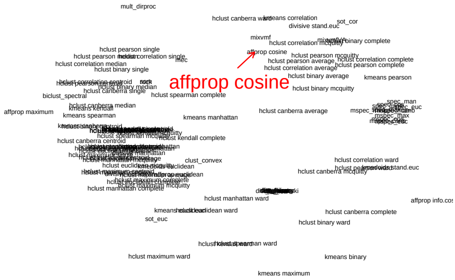
Example Discovery: What Do Members of Congress Do?

- Paper (Grimmer and King 2011): introduce new evaluation methods (like Cluster Quality)
- David Mayhew's (1974) famous typology
    - Advertising
    - Credit Claiming
    - Position Taking
- Data: 200 press releases from Frank Lautenberg's office (D-NJ)
- Apply our method (relying on many clustering algorithms)

Example Discovery



Example Discovery

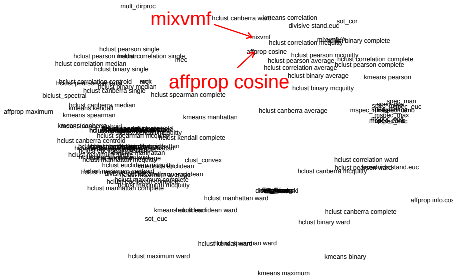


Each point is a clustering

Affinity Propagation-Cosine

(Dueck and Frey 2007)

Example Discovery



Each point is a clustering

Affinity Propagation-Cosine (Dueck and Frey 2007)

Close to:

Mixture of von Mises-Fisher distributions (Banerjee et. al. 2005)

$\Rightarrow$ Similar clustering of documents

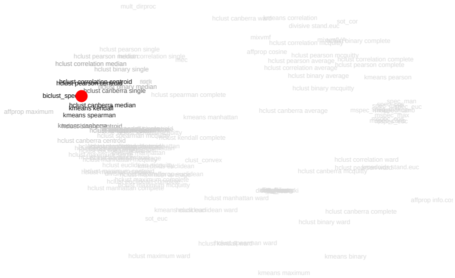
Example Discovery

Space between methods:

Example Discovery

Space between methods:

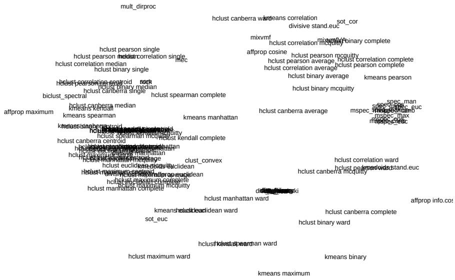
Example Discovery



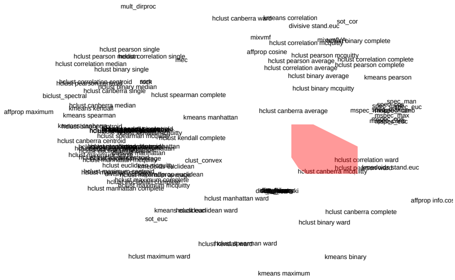
Space between methods:

local cluster ensemble

Example Discovery

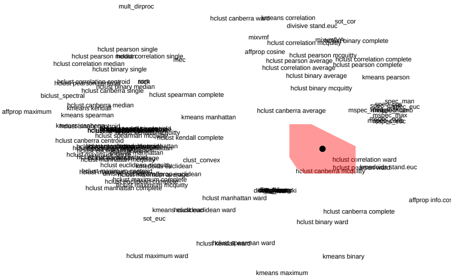


Example Discovery



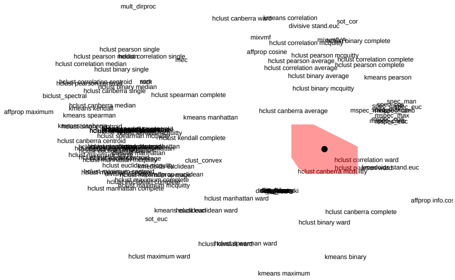
Found a region with clusterings that all reveal the same important insight

Example Discovery



Mixture:

Example Discovery



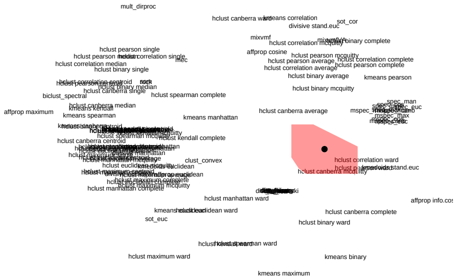
Mixture:

0.39 Hclust-Canberra-McQuitty

0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

Example Discovery



Mixture:

0.39 Hclust-Canberra-McQuitty

0.30 Spectral clustering Random Walk (Metrics 1-6)

0.13 Hclust-Correlation-Ward

0.09 Hclust-Pearson-Ward

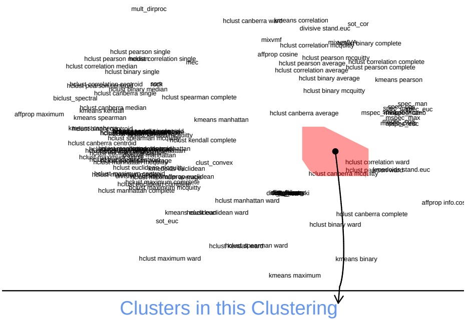
0.04 Spectral clustering Symmetric (Metrics 1-6)

Example Discovery

Mixture:

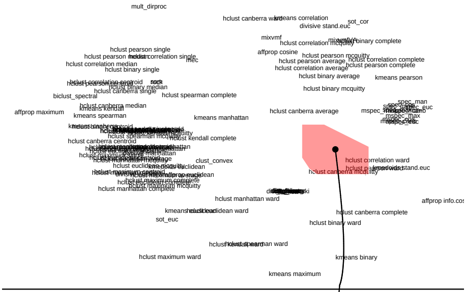
- 0.39 Hclust-Canberra-McQuitty
- 0.30 Spectral clustering Random Walk (Metrics 1-6)
- 0.13 Hclust-Correlation-Ward
- 0.09 Hclust-Pearson-Ward
- 0.05 Kmediods-Cosine
- 0.04 Spectral clustering Symmetric (Metrics 1-6)

Example Discovery

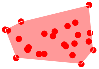


Mayhew

Example Discovery



Clusters in this Clustering

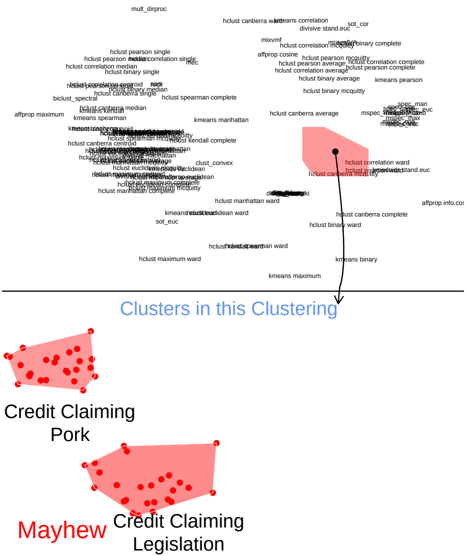


Credit Claiming
Pork

Credit Claiming, Pork:
"Sens. Frank R. Lautenberg (D-NJ) and Robert Menendez (D-NJ) announced that the U.S. Department of Commerce has awarded a $100,000 grant to the South Jersey Economic Development District"

Mayhew

Example Discovery

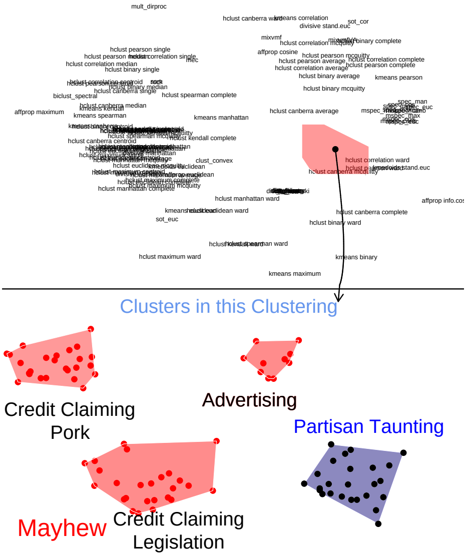


Credit Claiming, Legislation:
"As the Senate begins its recess, Senator Frank Lautenberg today pointed to a string of victories in Congress on his legislative agenda during this work period"

Example Discovery

Advertising:
"Senate Adopts Lautenberg/Menendez Resolution Honoring Spelling Bee Champion from New Jersey"

Example Discovery: Partisan Taunting



Partisan Taunting:
"Republicans Selling Out Nation on Chemical Plant Security"

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Sen. Lautenberg on Senate Floor 4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Sen. Lautenberg
on Senate Floor
4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Sen. Lautenberg on Senate Floor 4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities less safe." [Homeland Security]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Definition: Explicit, public, and negative attacks on another political party or its members

Sen. Lautenberg on Senate Floor 4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities less safe." [Homeland Security]

In Sample Illustration of Partisan Taunting

Important Concept Overlooked in Mayhew's (1974) typology

Definition: Explicit, public, and negative attacks on another political party or its members

Consequences for representation: Deliberative, Polarization, Policy

Sen. Lautenberg on Senate Floor 4/29/04

- "Senator Lautenberg Blasts Republicans as 'Chicken Hawks' " [Government Oversight]
- "The scopes trial took place in 1925. Sadly, President Bush's veto today shows that we haven't progressed much since then" [Healthcare]
- "Every day the House Republicans dragged this out was a day that made our communities less safe." [Homeland Security]

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.

Out of Sample Confirmation of Partisan Taunting

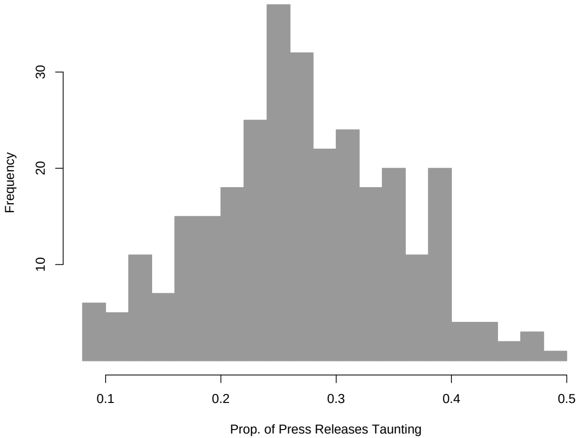
- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.

Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party

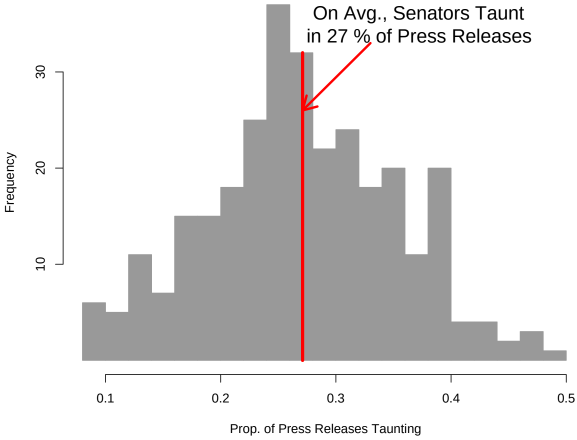
Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party

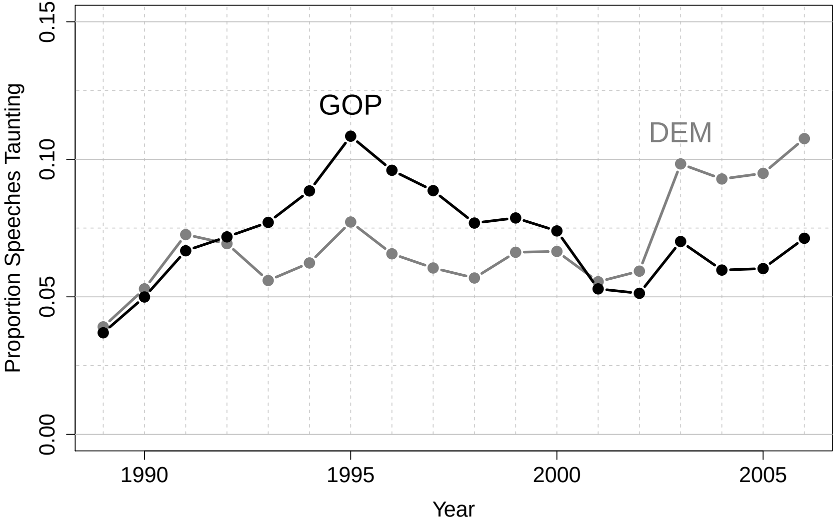


Out of Sample Confirmation of Partisan Taunting

- Discovered using 200 press releases; 1 senator.
- Demonstrate prevalence using senators' press releases.
- Apply supervised learning method: measure proportion of press releases a senator taunts other party



Over Time Tauting Rates in Speeches



How do we formulate conceptualizations?

How do we formulate conceptualizations?
Tension in potential methods

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
   - Provides single answer, uncertainty estimates

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
   - Provides single answer, uncertainty estimates
   - Imposes many unstated assumptions, narrow set of conceptualizations considered

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
    - Provides single answer, uncertainty estimates
    - Imposes many unstated assumptions, narrow set of conceptualizations considered
    - Difficult for political scientist to tune to their problem

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
   - Provides single answer, uncertainty estimates
   - Imposes many unstated assumptions, narrow set of conceptualizations considered
   - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
   - Provides single answer, uncertainty estimates
   - Imposes many unstated assumptions, narrow set of conceptualizations considered
   - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
   - Varies assumptions, ensures many different conceptualizations considered

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
   - Provides single answer, uncertainty estimates
   - Imposes many unstated assumptions, narrow set of conceptualizations considered
   - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
   - Varies assumptions, ensures many different conceptualizations considered
   - Burden on user to discover conceptualization

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
    - Provides single answer, uncertainty estimates
    - Imposes many unstated assumptions, narrow set of conceptualizations considered
    - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
    - Varies assumptions, ensures many different conceptualizations considered
    - Burden on user to discover conceptualization

- Best evaluation: An improbable experiment

How do we formulate conceptualizations?
Tension in potential methods

1) FAC methods tuned to problem
    - Provides single answer, uncertainty estimates
    - Imposes many unstated assumptions, narrow set of conceptualizations considered
    - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
    - Varies assumptions, ensures many different conceptualizations considered
    - Burden on user to discover conceptualization

- Best evaluation: An improbable experiment
    - Randomly assign incoming grad students to three conditions

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
    - Provides single answer, uncertainty estimates
    - Imposes many unstated assumptions, narrow set of conceptualizations considered
    - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
    - Varies assumptions, ensures many different conceptualizations considered
    - Burden on user to discover conceptualization

- Best evaluation: An improbable experiment
    - Randomly assign incoming grad students to three conditions
        - Topic Models (FAC)

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
    - Provides single answer, uncertainty estimates
    - Imposes many unstated assumptions, narrow set of conceptualizations considered
    - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
    - Varies assumptions, ensures many different conceptualizations considered
    - Burden on user to discover conceptualization

- Best evaluation: An improbable experiment
    - Randomly assign incoming grad students to three conditions
        - Topic Models (FAC)
        - Semi-supervised methods (CAC)

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
   - Provides single answer, uncertainty estimates
   - Imposes many unstated assumptions, narrow set of conceptualizations considered
   - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
   - Varies assumptions, ensures many different conceptualizations considered
   - Burden on user to discover conceptualization

- Best evaluation: An improbable experiment
  - Randomly assign incoming grad students to three conditions
    - Topic Models (FAC)
    - Semi-supervised methods (CAC)
    - Manual methods

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
    - Provides single answer, uncertainty estimates
    - Imposes many unstated assumptions, narrow set of conceptualizations considered
    - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
    - Varies assumptions, ensures many different conceptualizations considered
    - Burden on user to discover conceptualization

- Best evaluation: An improbable experiment
    - Randomly assign incoming grad students to three conditions
        - Topic Models (FAC)
        - Semi-supervised methods (CAC)
        - Manual methods
    - Observe group with most productivity 20-30 years later

How do we formulate conceptualizations?

Tension in potential methods

1) FAC methods tuned to problem
   - Provides single answer, uncertainty estimates
   - Imposes many unstated assumptions, narrow set of conceptualizations considered
   - Difficult for political scientist to tune to their problem
2) CAC methods to explore a space of partitions
   - Varies assumptions, ensures many different conceptualizations considered
   - Burden on user to discover conceptualization

- Best evaluation: An improbable experiment
  - Randomly assign incoming grad students to three conditions
    - Topic Models (FAC)
    - Semi-supervised methods (CAC)
    - Manual methods
  - Observe group with most productivity 20-30 years later
- To identify limits of methods, when to use which approach, need evaluations for the usefulness of conceptualizations

Clustering, FAC and CAC

This week

- Introduction to clustering
- Fully automated clustering algorithms
- Introduction to computer assisted clustering

Next week:

- Vanilla Topic models
- Structural Topic Models

EM Algorithm for Mixture of vMF Distributions

1) Initialize $\boldsymbol{\mu}$
2) Set r_{ik} to

$$r_{ik} = \frac{\pi_k \exp(\kappa \boldsymbol{\mu}_k^{'} \boldsymbol{x}_i^{*})}{\sum_{l=1}^{K} \pi_k \exp(\kappa \boldsymbol{\mu}_l^{'} \boldsymbol{x}_i^{*})}$$

3) Set $\boldsymbol{\mu}_k$ to

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} r_{ik} \boldsymbol{x}_i}{|| \sum_{i=1}^{N} r_{ik} \boldsymbol{x}_i ||}$$

   Set $\pi_k = \sum_{i=1}^{N} \frac{r_{ik}}{N}$

4) Assess change in objective function