Clustering

Clustering

Clustering:

Clustering

Clustering: look for 'groups' in data explicitly.

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

Clustering

Clustering: look for ‘groups’ in data explicitly.

Partition methods are most common:

$\rightarrow$ Include K-means, for which one pre-specifies cluster number,

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

so pick s such that $\sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} ||\mathbf{x_j} - \boldsymbol{\mu_i}||^2$ is minimized where $\boldsymbol{\mu_i}$ is (vector) mean of points in cluster S_i.

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

$\rightarrow$ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

so pick s such that $\sum_{i=1}^{k}\sum_{\mathbf{x_j}\in S_i}||\mathbf{x_j}-\boldsymbol{\mu_i}||^2$ is minimized where $\boldsymbol{\mu_i}$ is (vector) mean of points in cluster S_i.

$\rightarrow$ observations (documents) within clusters should be as similar as possible,

Clustering

Clustering: look for 'groups' in data explicitly.

Partition methods are most common:

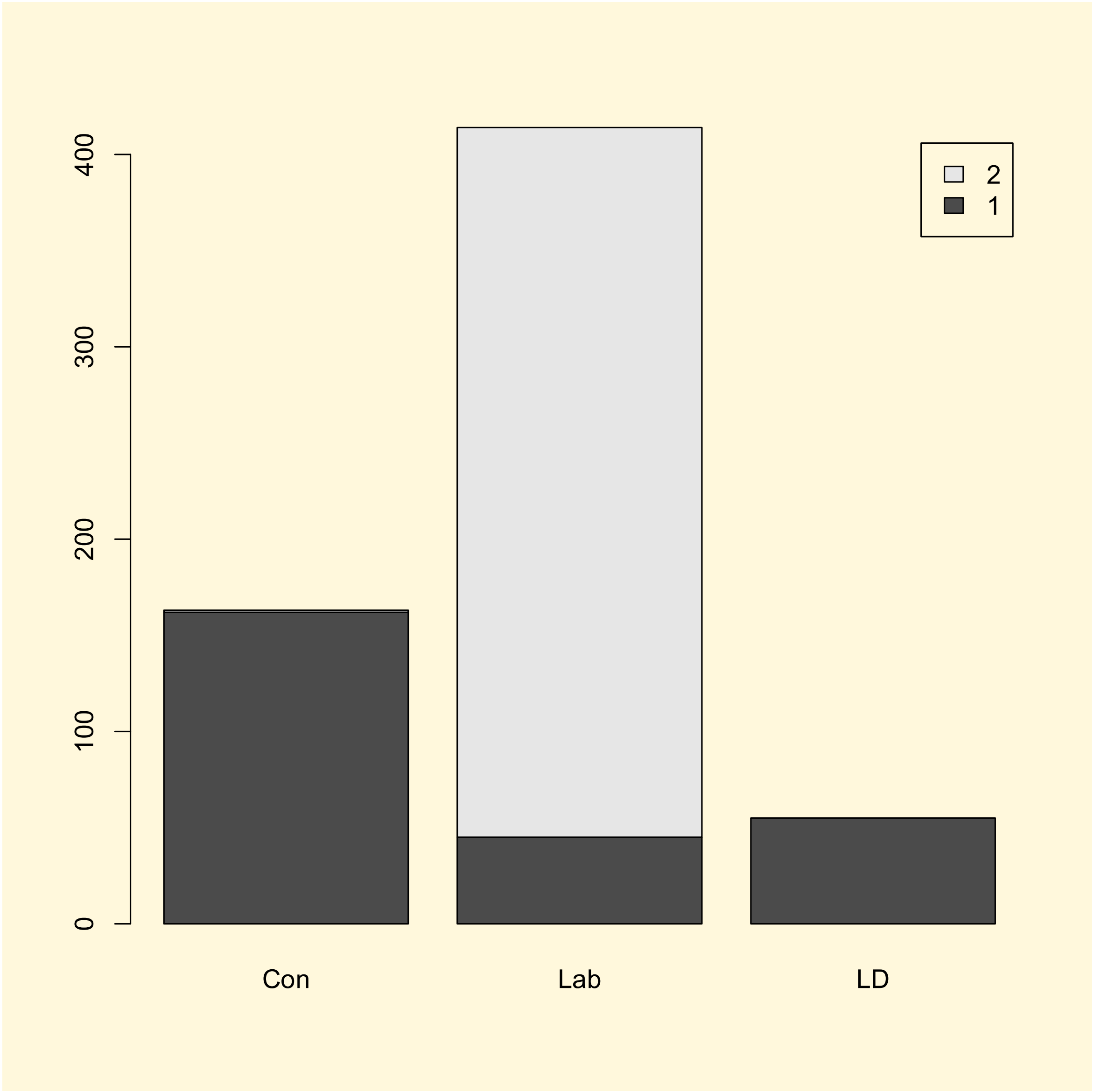
→ Include K-means, for which one pre-specifies cluster number, and algorithm puts observations into clusters which minimize within cluster sum of squares

so pick s such that $\sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} ||\mathbf{x_j} - \boldsymbol{\mu_i}||^2$ is minimized where $\boldsymbol{\mu_i}$ is (vector) mean of points in cluster S_i.

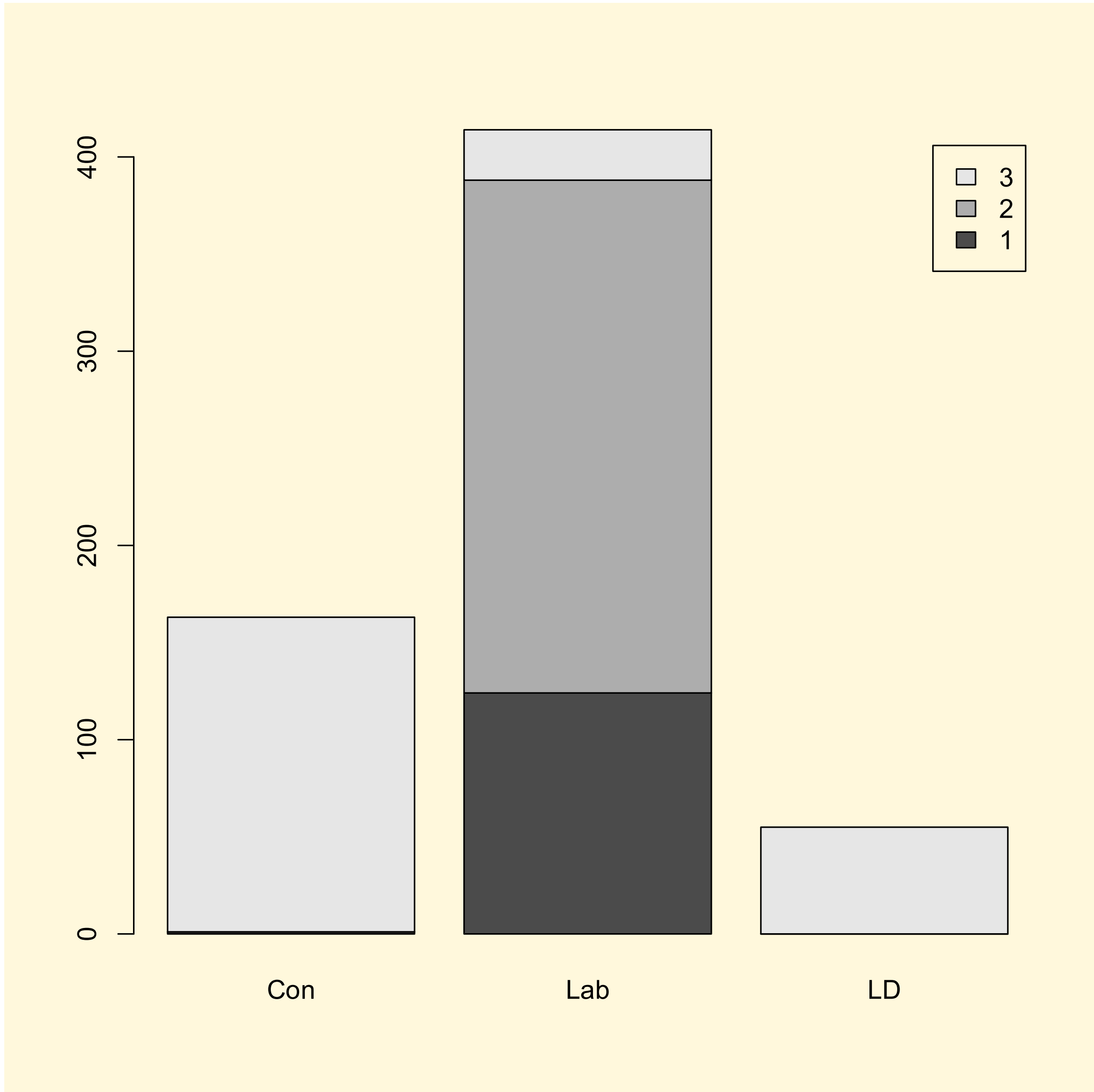
→ observations (documents) within clusters should be as similar as possible, observations (documents) in different clusters should be as different as possible.

k-means on Commons Roll Calls

k-means on Commons Roll Calls



k-means on Commons Roll Calls



Hierarchical Methods

Hierarchical Methods

Successively aggregate groups of observations.

Hierarchical Methods

Successively aggregate groups of observations.

Can be agglomerative/bottom-up in sense that everything starts in own cluster and then groups are formed by putting observations together

Hierarchical Methods

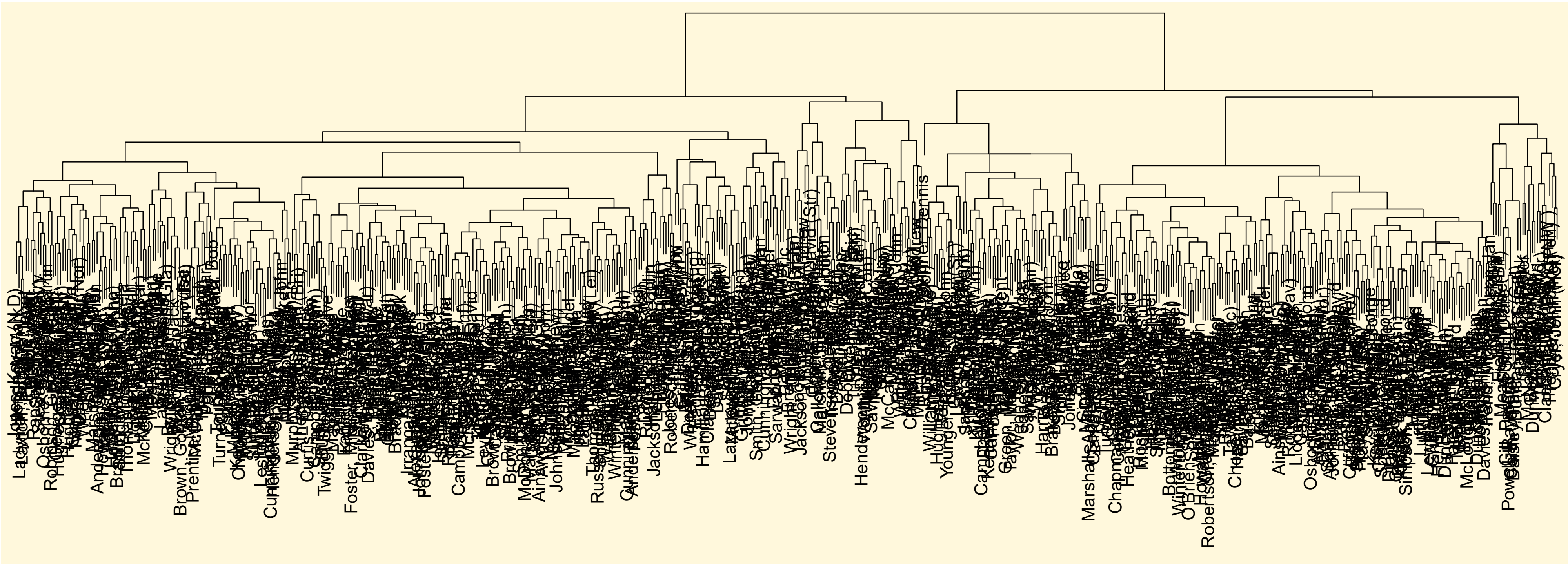
Successively aggregate groups of observations.

Can be agglomerative/bottom-up in sense that everything starts in own cluster and then groups are formed by putting observations together

Or Divisive/top down in sense that everything starts in same cluster and then splits are performed (typically on one feature) to form clusters.

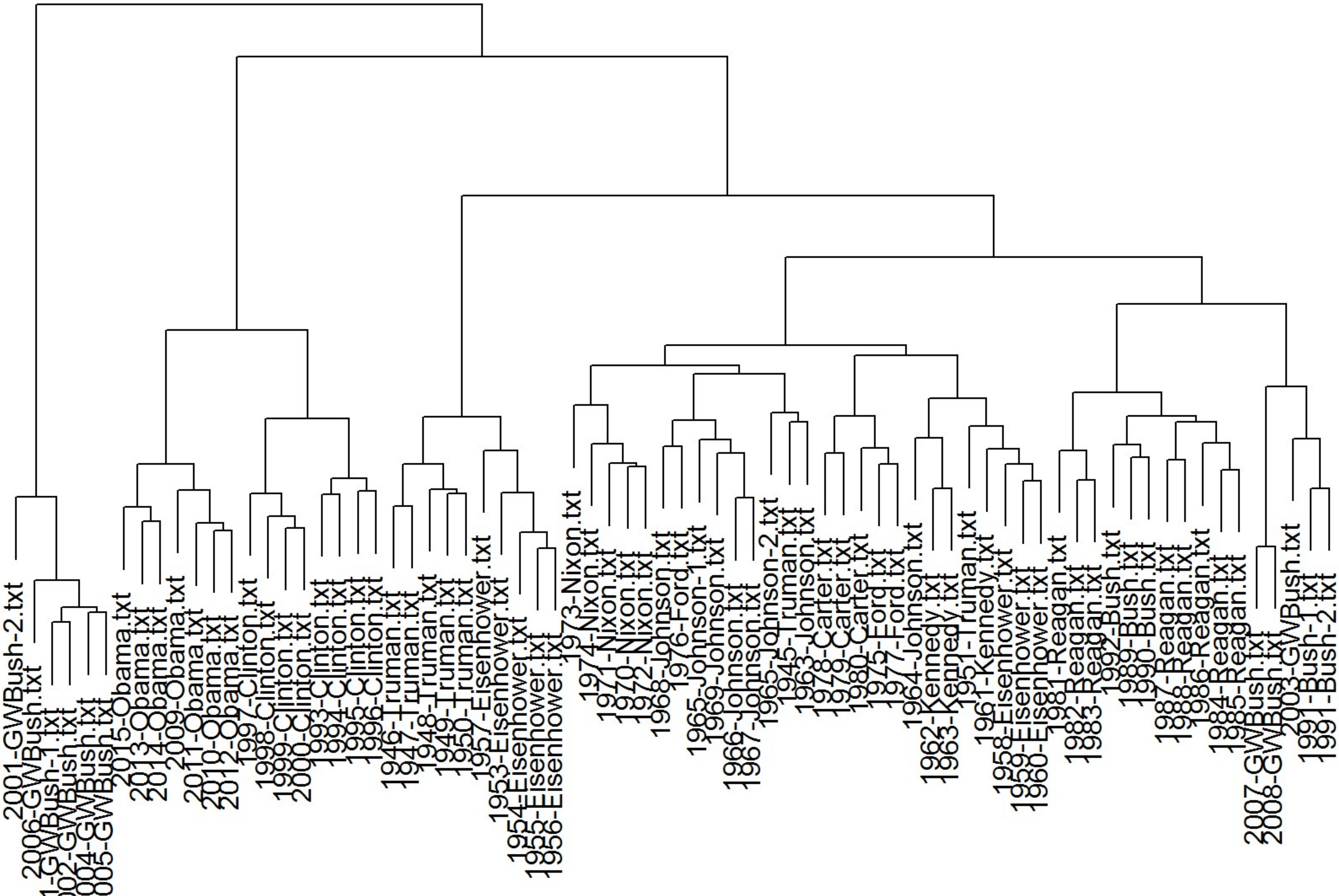
Hierarchical: Commons

Hierarchical: Commons



Hierarchical: SOTU (Frank Evans, `dzone.com`)

Hierarchical: SOTU (Frank Evans, `dzone.com`)



Notes

Notes

Gaussian assumptions probably off-base for many text problems

Notes

Gaussian assumptions probably off-base for many text problems

$n > p$ not met in many text examples so cannot always apply techniques ‘off the shelf’

Notes

Gaussian assumptions probably off-base for many text problems

$n > p$ not met in many text examples so cannot always apply techniques 'off the shelf' and missing data not trivial to handle

Notes

Gaussian assumptions probably off-base for many text problems

$n > p$ not met in many text examples so cannot always apply techniques 'off the shelf' and missing data not trivial to handle

Hard to compare across specifications:

Notes

Gaussian assumptions probably off-base for many text problems

$n > p$ not met in many text examples so cannot always apply techniques 'off the shelf' and missing data not trivial to handle

Hard to compare across specifications: e.g. $k = 2$, $k = 3$

Notes

Gaussian assumptions probably off-base for many text problems

$n > p$ not met in many text examples so cannot always apply techniques 'off the shelf' and missing data not trivial to handle

Hard to compare across specifications: e.g. $k = 2$, $k = 3$

No underlying model of human behavior/text generation gives rise to these techniques

Notes

Gaussian assumptions probably off-base for many text problems

$n > p$ not met in many text examples so cannot always apply techniques 'off the shelf' and missing data not trivial to handle

Hard to compare across specifications: e.g. $k = 2$, $k = 3$

No underlying model of human behavior/text generation gives rise to these techniques

"General purpose computer-assisted clustering and conceptualization", Grimmer and King (2010)

“General purpose computer-assisted clustering and conceptualization”, Grimmer and King (2010)

Motivation: there are an infinite number of possible algorithms,

"General purpose computer-assisted clustering and conceptualization", Grimmer and King (2010)

Motivation: there are an infinite number of possible algorithms, and hard to choose which one makes most 'sense' for given substantive problem.

“General purpose computer-assisted clustering and conceptualization”, Grimmer and King (2010)

Motivation: there are an infinite number of possible algorithms, and hard to choose which one makes most ‘sense’ for given substantive problem.

G&K Use ‘all’ of them,

"General purpose computer-assisted clustering and conceptualization", Grimmer and King (2010)

Motivation: there are an infinite number of possible algorithms, and hard to choose which one makes most 'sense' for given substantive problem.

G&K Use 'all' of them, and allow users to choose one (or more) that maximizes some 'insightful-ness' criteria.

"General purpose computer-assisted clustering and conceptualization", Grimmer and King (2010)

Motivation: there are an infinite number of possible algorithms, and hard to choose which one makes most 'sense' for given substantive problem.

G&K Use 'all' of them, and allow users to choose one (or more) that maximizes some 'insightful-ness' criteria.

This requires thoughtful visualization,

“General purpose computer-assisted clustering and conceptualization”, Grimmer and King (2010)

Motivation: there are an infinite number of possible algorithms, and hard to choose which one makes most ‘sense’ for given substantive problem.

G&K Use ‘all’ of them, and allow users to choose one (or more) that maximizes some ‘insightful-ness’ criteria.

This requires thoughtful visualization, to help humans select particular partition.

“General purpose computer-assisted clustering and conceptualization”, Grimmer and King (2010)

Motivation: there are an infinite number of possible algorithms, and hard to choose which one makes most ‘sense’ for given substantive problem.

G&K Use ‘all’ of them, and allow users to choose one (or more) that maximizes some ‘insightful-ness’ criteria.

This requires thoughtful visualization, to help humans select particular partition.

Plus simultaneously allow users to select combinations of clusterings that look ‘useful’.

Steps

Steps

1. standard pre-processing of texts,

Steps

1. standard pre-processing of texts, throwing out very rare and very common terms.

Steps

1. standard pre-processing of texts, throwing out very rare and very common terms.

2. apply very large number of clustering algorithms, with multiple specifications of each.

Steps

1. standard pre-processing of texts, throwing out very rare and very common terms.

2. apply very large number of clustering algorithms, with multiple specifications of each.
3. calculate distance between J clusterings as function of number of pairs of documents not placed together in same cluster.

Steps

1. standard pre-processing of texts, throwing out very rare and very common terms.
2. apply very large number of clustering algorithms, with multiple specifications of each.
3. calculate distance between J clusterings as function of number of pairs of documents not placed together in same cluster.
4. project this $J \times J$ clustering matrix down to 2D Euclidean space using Sammon MDS to preserve small distances better (than large ones)

Steps

1. standard pre-processing of texts, throwing out very rare and very common terms.

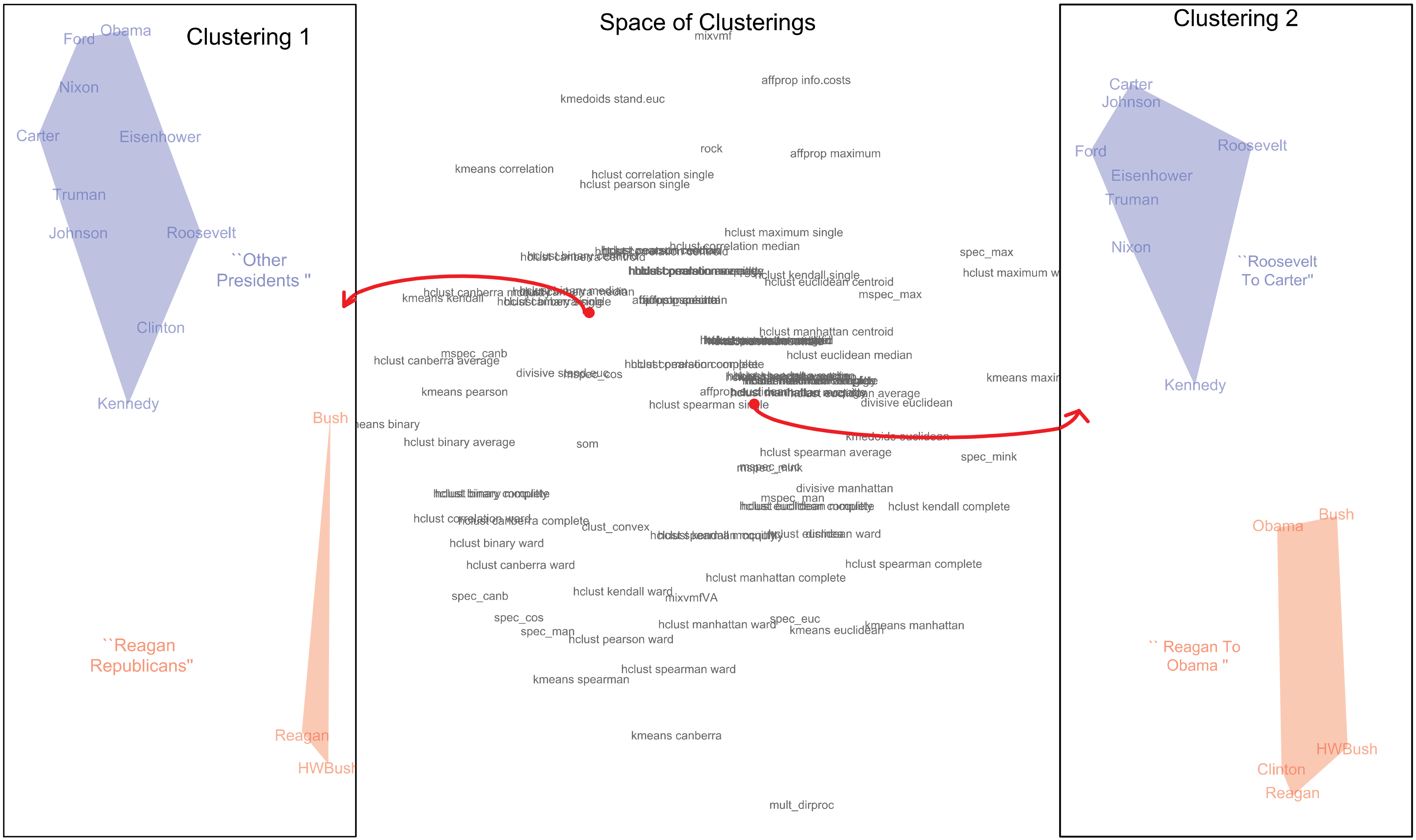
2. apply very large number of clustering algorithms, with multiple specifications of each.
3. calculate distance between J clusterings as function of number of pairs of documents not placed together in same cluster.
4. project this $J \times J$ clustering matrix down to 2D Euclidean space using Sammon MDS to preserve small distances better (than large ones)
5. allow for local cluster ensemble which is (a new) clustering composed of combination of clusterings that are nearby any given point in 2D space.

Steps

1. standard pre-processing of texts, throwing out very rare and very common terms.
2. apply very large number of clustering algorithms, with multiple specifications of each.
3. calculate distance between J clusterings as function of number of pairs of documents not placed together in same cluster.
4. project this $J \times J$ clustering matrix down to 2D Euclidean space using Sammon MDS to preserve small distances better (than large ones)
5. allow for local cluster ensemble which is (a new) clustering composed of combination of clusterings that are nearby any given point in 2D space.
6. visualize for users.

Example: Biographies of Presidents

Example: Biographies of Presidents



Evaluating Clusterings

Evaluating Clusterings

A clustering is good if “the user, or the user’s intended audience, finds the chosen clustering useful or insightful.”

Evaluating Clusterings

A clustering is good if "the user, or the user's intended audience, finds the chosen clustering useful or insightful."

But this is possibly unfalsifiable,

Evaluating Clusterings

A clustering is good if "the user, or the user's intended audience, finds the chosen clustering useful or insightful."

But this is possibly unfalsifiable, and not necessarily scientific. . .

Evaluating Clusterings

A clustering is good if "the user, or the user's intended audience, finds the chosen clustering useful or insightful."

But this is possibly unfalsifiable, and not necessarily scientific. . .

So suggest some more measurable/formal evaluation mechanisms:

Evaluating Clusterings

A clustering is good if "the user, or the user's intended audience, finds the chosen clustering useful or insightful."

But this is possibly unfalsifiable, and not necessarily scientific...

So suggest some more measurable/formal evaluation mechanisms:

1. Cluster Quality: randomly draw pairs of documents from same cluster and different clusters,

Evaluating Clusterings

A clustering is good if "the user, or the user's intended audience, finds the chosen clustering useful or insightful."

But this is possibly unfalsifiable, and not necessarily scientific...

So suggest some more measurable/formal evaluation mechanisms:

1. Cluster Quality: randomly draw pairs of documents from same cluster and different clusters, and ask human coders how closely related they are.

Evaluating Clusterings

A clustering is good if "the user, or the user's intended audience, finds the chosen clustering useful or insightful."

But this is possibly unfalsifiable, and not necessarily scientific. . .

So suggest some more measurable/formal evaluation mechanisms:

1. Cluster Quality: randomly draw pairs of documents from same cluster and different clusters, and ask human coders how closely related they are.
2. Discovery Quality: show scholars the cluster space and see if it improves their understanding of own data

Discovery of Partisan Taunting in Press Releases

Discovery of Partisan Taunting in Press Releases

Discovery of Partisan Taunting in Press Releases