Mining Clustering Dimensions

Sajib Dasgupta and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Clustering Dimensions

- dimensions along which a dataset can be naturally clustered

- Movie reviews can be clustered by
  - **genre** (action, romantic, documentary, …)
  - **sentiment** (positive, negative, …)
  - **…**

Clustering Dimensions

- dimensions along which a dataset can be naturally clustered

- Movie reviews can be clustered by
  - **genre** (action, romantic, documentary, …)
  - **sentiment** (positive, negative, …)
  - **…**

clustering dimensions

Task

- Given data *X*, discover in an unsupervised manner the dimensions along which *X* can be meaningfully clustered

Task

- Given data *X*, discover in an unsupervised manner the dimensions along which *X* can be meaningfully clustered

Task

- Given data *X*, discover in an unsupervised manner the dimensions along which *X* can be meaningfully clustered

- A meaningful clustering is a clustering that is
  - human interpretable
  - qualitatively strong

Why bother?

- Exploratory data analysis
  - useful for someone who doesn’t know how the data can be clustered

Goal

- Propose a text clustering algorithm that can
  - produce multiple clusterings of a text collection from which we induce its important clustering dimensions

Goal

- Propose a text clustering algorithm that can
  - produce multiple clusterings of a text collection from which we induce its important clustering dimensions
  - allow a user to **visualize** these dimensions

Goal

- Propose a text clustering algorithm that can
  - produce multiple clusterings of a text collection from which we induce its important clustering dimensions
  - allow a user to **visualize** these dimensions
    - by representing each dimension using a small number of unigrams

Goal

- Propose a text clustering algorithm that can
  - produce multiple clusterings of a text collection from which we induce the important clustering dimensions
  - allow a user to **visualize** these dimensions
    - by representing each dimension using a small number of unigrams

Example

- Given a set of book and DVD reviews …

Example

- Given a set of book and DVD reviews …

Dimension 1	Dimension 2	Dimension 3
reader information research important text	wonderful excellent music highly collection	bought workout recipes information disappointed
music script actors films comedy	boring waste novel worst pages	young men scene cast role

Example

- Given a set of book and DVD reviews …

Dimension 1	Dimension 2	Dimension 3
reader information research important text	wonderful excellent music highly collection	bought workout recipes information disappointed
music script actors films comedy	boring waste novel worst pages	young men scene cast role

Example

- Given a set of book and DVD reviews …

Dimension 1	Dimension 2	Dimension 3
Book		
reader	wonderful	bought
information	excellent	workout
research	music	recipes
important	highly	information
text	collection	disappointed
DVD		
music	boring	young
script	waste	men
actors	novel	scene
films	worst	cast
comedy	pages	role

Example

- Given a set of book and DVD reviews …

Topic	Dimension 2	Dimension 3
Book		
reader	wonderful	bought
information	excellent	workout
research	music	recipes
important	highly	information
text	collection	disappointed
DVD		
music	boring	young
script	waste	men
actors	novel	scene
films	worst	cast
comedy	pages	role

Example

- Given a set of book and DVD reviews …

Topic	Dimension 2	Dimension 3
Book		
reader	wonderful	bought
information	excellent	workout
research	music	recipes
important	highly	information
text	collection	disappointed
DVD		
music	boring	young
script	waste	men
actors	novel	scene
films	worst	cast
comedy	pages	role

Example

- Given a set of book and DVD reviews …

Topic	Dimension 2	Dimension 3
Book	**Positive**	
reader	wonderful	bought
information	excellent	workout
research	music	recipes
important	highly	information
text	collection	disappointed
DVD	**Negative**	
music	boring	young
script	waste	men
actors	novel	scene
films	worst	cast
comedy	pages	role

Example

- Given a set of book and DVD reviews …

Topic	**Sentiment**	**Dimension 3**
Book	**Positive**	
reader	wonderful	bought
information	excellent	workout
research	music	recipes
important	highly	information
text	collection	disappointed
DVD	**Negative**	
music	boring	young
script	waste	men
actors	novel	scene
films	worst	cast
comedy	pages	role

Our Text Clustering Algorithm

- Two steps:
- **Step 1**
  - Produce multiple clusterings
- **Step 2**
  - Represent each dimension with representative words

Our Text Clustering Algorithm

- Two steps:

- **Step 1**
  - Produce multiple clusterings

- **Step 2**
  - Represent each dimension with representative words

Producing Multiple Clusterings

- Can we use traditional clustering algorithms to discover clustering dimensions?

Producing Multiple Clusterings

- Can we use traditional clustering algorithms to discover clustering dimensions?
  - Perhaps no …
  - Typically only one clustering is produced

Only one clustering dimension can be recovered

Producing Multiple Clusterings

- What if we tweak these traditional clustering algorithms using human knowledge?

Producing Multiple Clusterings

- What if we tweak these traditional clustering algorithms using human knowledge?
  - design different similarity functions or objective functions so that multiple meaningful clusterings can be produced

Producing Multiple Clusterings

- What if we tweak these traditional clustering algorithms using human knowledge?
  - design different similarity functions or objective functions so that multiple meaningful clusterings can be produced

Defeats the purpose of exploratory data analysis

Producing Multiple Clusterings

- Other attempts
  - Gondek & Hofmann (2004), Davidson & Qi (2007), …
  - assume that one clustering is provided; the goal is to induce a distinctly different clustering

Producing Multiple Clusterings

- Other attempts
  - Gondek & Hofmann (2004), Davidson & Qi (2007), …
  - assume that one clustering is provided; the goal is to induce a distinctly different clustering

Semi-supervised: still require knowledge of the data

Producing Multiple Clusterings

- Meta clustering (Caruana et al., 2006)
  - unsupervised method
  - run k-means multiple times, each time with a random selection of seeds and a random weighting of features
  - treat each local minimum as a possible clustering

Producing Multiple Clusterings

- Meta clustering (Caruana et al., 2006)
  - unsupervised method
  - run k-means multiple times, each time with a random selection of seeds and a random weighting of features
  - treat each local minimum as a possible clustering

Many local minima are qualitatively poor

Producing Multiple Clusterings

- Jain et al. (2008)
  - unsupervised method
  - learns two clusterings in a “decorrelated” k-means framework
  - model aims to achieve typical k-means objectives and ensure the two induced clusterings are distinctly different

Producing Multiple Clusterings

- Jain et al. (2008)
  - learns two clusterings in a "decorrelated" k-means framework
  - model aims to achieve typical k-means objectives and ensure the two induced clusterings are distinctly different
  - objective function:

$$\sum_{i=1}^{k_1} \sum_{x \in C_i^1} ||x - \mu_i||^2 + \sum_{j=1}^{k_2} \sum_{x \in C_j^2} ||x - \nu_j||^2$$

$$+\lambda \sum_{i,j} (\beta_j^T \mu_i)^2 + \lambda \sum_{i,j} (\alpha_i^T \nu_j)^2$$

Producing Multiple Clusterings

- Jain et al. (2008)
  - learns two clusterings in a “decorrelated” k-means framework
  - model aims to achieve typical k-means objectives and ensure the two induced clusterings are distinctly different
  - objective function:

$$\sum_{i=1}^{k_1} \sum_{x \in C_i^1} ||x - \mu_i||^2 + \sum_{j=1}^{k_2} \sum_{x \in C_j^2} ||x - \nu_j||^2$$
$$+ \lambda \sum_{i,j} (\beta_j^T \mu_i)^2 + \lambda \sum_{i,j} (\alpha_i^T \nu_j)^2$$

Producing Multiple Clusterings

- Jain et al. (2008)
  - learns two clusterings in a “decorrelated” k-means framework
  - model aims to achieve typical k-means objectives and ensure the two induced clusterings are distinctly different
  - objective function:

$$\sum_{i=1}^{k_1} \sum_{x \in C_i^1} ||x - \mu_i||^2 + \sum_{j=1}^{k_2} \sum_{x \in C_j^2} ||x - \nu_j||^2$$
$$+ \lambda \sum_{i,j} (\beta_j^T \mu_i)^2 + \lambda \sum_{i,j} (\alpha_i^T \nu_j)^2$$

Objective can become very convoluted as # clusterings $\uparrow$

Producing Multiple Clusterings

- Can we have a method for producing multiple clusterings that
  - is simple
  - is unsupervised
  - employs a single similarity function and a single objective
  - can produce distinctly different and qualitatively strong clusterings?

Idea

- Go beyond producing the clustering that is optimal w.r.t. the objective function and produce suboptimal clusterings

Idea

- Go beyond producing the clustering that is optimal w.r.t. the objective function and produce suboptimal clusterings

but not overly suboptimal

How?

- Use spectral clustering
- Ng et al. (2001)

Spectral Clustering (Ng et al., 2001)

- Given data D and a pairwise similarity function $\varnothing$,
  1. form similarity matrix $S=\varnothing(D)$
  2. form diagonal matrix G, where $G(i,i)$=sum of the i-th row of S
  3. form Laplacian matrix $L=G^{-1/2}\, S\, G^{1/2}$
  4. find the eigenvectors of L
  5. apply k-means to cluster using these eigenvectors

Spectral Clustering (Ng et al., 2001)

- Given data D and a pairwise similarity function $\varnothing$,
  1. form similarity matrix $S=\varnothing(D)$
  2. form diagonal matrix G, where $G(i,i)$=sum of the i-th row of S
  3. form Laplacian matrix $L=G^{-1/2}\ S\ G^{1/2}$
  4. find the eigenvectors of L
  5. apply k-means to cluster using these eigenvectors

How to produce the optimal clustering and suboptimal clusterings using these eigenvectors?

Producing the Optimal Clustering

- Use $\mathbf{e}_2$, the second eigenvector
  - real-valued solution to the normalized min-cut objective

Producing Suboptimal Clusterings

- Each of $\mathbf{e}_3$, $\mathbf{e}_4$, $\mathbf{e}_5$, … are suboptimal solutions to the normalized cut objective
  - $\mathbf{e}_3$ is the optimal solution to objective orthogonal to $\mathbf{e}_2$
  - $\mathbf{e}_4$ is the optimal solution to objective orthogonal to $\mathbf{e}_2$ and $\mathbf{e}_3$
  - ...

Why does it make sense?

- e_3, e_4, e_5, … are suboptimal, but perhaps reasonably good, solutions to the normalized cut objective
  - may yield **qualitatively strong** clusterings

Why does it make sense?

- e_3, e_4, e_5, … are suboptimal, but perhaps reasonably good, solutions to the normalized cut objective
  - may yield **qualitatively strong** clusterings

- The eigenvectors are orthogonal to each other
  - may yield **distinctly different** clusterings

Why does it make sense?

- e_3, e_4, e_5, … are suboptimal, but perhaps reasonably good, solutions to the normalized cut objective
  - may yield **qualitatively strong** clusterings

- The eigenvectors are orthogonal to each other
  - may yield **distinctly different** clusterings

To produce multiple clusterings …

- Use each of the top eigenvectors to produce a clustering
  - $\mathbf{e}_2$ Clustering 1
  - $\mathbf{e}_3$ Clustering 2
  - $\mathbf{e}_4$ Clustering 3
  - $\mathbf{e}_5$ Clustering 4
  - **…**

- To produce *m* clusterings, we use the top (*m*+1) eigenvectors (excluding $\mathbf{e}_1$)

To produce multiple clusterings …

- Use a single similarity function: dot product
- Use a single objective function: normalized cut

Our Text Clustering Algorithm

- Two steps:

- **Step 1**
  - Produce multiple clusterings

- **Step 2**
  - Represent each dimension with representative words

Selecting the Representative Words

- Given a clustering, we rank its words using the weighted log-likelihood ratio (WLLR):

$$P(w_i \mid C_j) \cdot \log \frac{P(w_i \mid C_j)}{P(w_i \mid \neg C_j)}$$

where w_i: i-th feature, C_j: j-th cluster

Selecting the Representative Words

- Given a clustering, we rank its words using the weighted log-likelihood ratio (WLLR):

$$P(w_i \mid C_j) \cdot \log \frac{P(w_i \mid C_j)}{P(w_i \mid \neg C_j)}$$

where w_i: i-th feature, C_j: j-th cluster

- w_i has a high rank in C_j if it appears frequently in C_j and infrequently in $\neg C_j$

Selecting the Representative Words

- Given a clustering, we rank its words using the weighted log-likelihood ratio (WLLR):

$$P(w_i \mid C_j) \cdot \log \frac{P(w_i \mid C_j)}{P(w_i \mid \neg C_j)}$$

where w_i: i-th feature, C_j: j-th cluster

- w_i has a high rank in C_j if it appears frequently in C_j and infrequently in $\neg C_j$

- An induced clustering dimension is represented using the top-ranked features in each cluster.

Evaluation

Goal:

Determine whether our algorithm

- induces clustering dimensions that are human-interpretable
- produces clusterings that are qualitatively strong

given a text collection

Datasets

- Two Newsgroups (TNG)
  - `talks.politics` and `sci.crypt` (**politics vs. science**)
- Blitzer et al.'s datasets: book (BOO) and DVD reviews
  - Each contains 2000 customer reviews of books and DVDs
- The BOO-DVD dataset
  - Composed of the 2000 book reviews and 2000 DVD reviews
- The politics (POL) dataset
  - 2000 political articles written by columnists who identified themselves as Democrats or Republicans

Gold-Standard Creation

Step 1: Identify the clustering dimensions

- Five students
  - agreed on the 2-way clustering dimensions for each dataset

Gold-Standard Creation

Step 1: Identify the clustering dimensions

- Five students
  - agreed on the 2-way clustering dimensions for each dataset
  - proposed 13 clustering dimensions for the five datasets

Dataset	Clustering Dimensions
TNG	Topic
BOO	Sentiment, Subjectivity, Strength
DVD	Sentiment, Subjectivity, Strength
BOO-DVD	Sentiment, Subjectivity, Strength, Topic
POL	Political Affiliation, Policy

Gold-Standard Creation

Step 1: Identify the clustering dimensions

- Five students
  - agreed on the 2-way clustering dimensions for each dataset
  - proposed 13 clustering dimensions for the five datasets

Dataset	Clustering Dimensions
TNG	Topic
BOO	Sentiment, Subjectivity, Strength
DVD	Sentiment, Subjectivity, Strength
BOO-DVD	Sentiment, Subjectivity, Strength, Topic
POL	Political Affiliation, Policy

Gold-Standard Creation

Step 1: Identify the clustering dimensions

- Five students
  - agreed on the 2-way clustering dimensions for each dataset
  - proposed 13 clustering dimensions for the five datasets

Dataset	**Clustering Dimensions**
TNG	Topic
BOO	Sentiment, Subjectivity, Strength
DVD	Sentiment, Subjectivity, Strength
BOO-DVD	Sentiment, Subjectivity, Strength, Topic
POL	Political Affiliation, Policy

Gold-Standard Creation (Cont'd)

Step 2: Annotate documents along each dimension

Applying Our Clustering Algorithm

- For each dataset,
  - cluster using $\mathbf{e}_2$ through $\mathbf{e}_5$ (2nd through 5th eigenvectors), yielding four 2-way clustering
  - represent each clustering dimension with unigrams selected via WLLR

Experiment 1: Human Interpretability

- Goals: determine
  - whether an induced dimension is human-interpretable when represented as two ranked lists of features
  - how well our algorithm can recover the clustering dimensions manually identified for each dataset

Experimental Setup

- Perform experiments involving 10 students
  - None of them were involved in data annotation
- For each clustering produced by our algorithm
  - Show each human judge the top 100 features selected for each cluster of each of the 4 clusterings according to WLLR
  - Ask her to label the resulting dimension, if possible

Experimental Setup

- Perform experiments involving 10 CS graduate students
  - None of them were involved in data annotation
- For each clustering produced by our algorithm
  - Show each human judge the top 100 features selected for each cluster of each of the 4 clusterings according to WLLR
  - Ask her to label the resulting dimension, if possible
- They did **not** know the set of possible dimension labels

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Fraction of judges who thought the dimension is interpretable

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Label assigned by the majority of the judges if more than five judges think that the dimension is interpretable

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

How many clustering dimensions in the gold standard were being recovered?

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Dataset	Clustering Dimensions
TNG	Topic
BOO	Sentiment, Subjectivity, Strength
DVD	Sentiment, Subjectivity, Strength
BOO/DVD	Sentiment, Subjectivity, Strength, Topic
POL	Political Affiliation, Policy

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Dataset	Clustering Dimensions
TNG	Topic
BOO	Sentiment, Subjectivity, Strength
DVD	Sentiment, Subjectivity, Strength
BOO/DVD	Sentiment, Subjectivity, Strength, Topic
POL	Political Affiliation, Policy

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Dataset	Clustering Dimensions
TNG	Topic ✓
BOO	Sentiment ✓ Subjectivity ✓ Strength ✗
DVD	Sentiment, Subjectivity, Strength
BOO/DVD	Sentiment, Subjectivity, Strength, Topic
POL	Political Affiliation, Policy

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Dataset	Clustering Dimensions
TNG	Topic ✓
BOO	Sentiment ✓, Subjectivity ✓, Strength ✗
DVD	Sentiment ✓, Subjectivity ✓, Strength ✗
BOO/DVD	Sentiment ✓, Subjectivity ✓, Strength ✗, Topic ✓
POL	Political Affiliation ✓, Policy ✓

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Dataset	Clustering Dimensions
TNG	Topic ✓
BOO	Sentiment ✓, Subjectivity ✓, Strength ✗
DVD	Sentiment ✓, Subjectivity ✓, Strength ✗
BOO/DVD	Sentiment ✓, Subjectivity ✓, Strength ✗, Topic ✓
POL	Political Affiliation ✓, Policy ✓

Recall = 77%

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Did the judges agree on which dimension label should be assigned when a dimension was found to be human-interpretable?

Human Interpretability Results

Dataset	2nd eigenvector		3rd eigenvector		4th eigenvector		5th eigenvector	
TNG	1.0	Topic	1.0	Topic	1.0	Topic	0.0	---
BOO	0.0	---	0.8	Subjectivity	1.0	Sentiment	0.4	---
DVD	0.8	Subjectivity	1.0	Sentiment	0.0	---	0.2	---
BOO/DVD	1.0	Topic	0.7	Subjectivity	1.0	Sentiment	1.0	Sentiment
POL	0.7	Political Affil	1.0	War/Non-war	1.0	War/Non-war	0.0	---

Did the judges agree on which dimension label should be assigned when a dimension was found to be human-interpretable?

Agreement rate: ≥70%

Experiment 2: Clustering Quality

- Since many of the induced clustering dimensions are human-interpretable, the clusterings are presumably qualitatively strong, but …
  - how strong are they?

Experiment 2: Clustering Quality

- Since many of the induced clustering dimensions are human-interpretable, the clusterings are presumably qualitatively strong, but …
  - how strong are they?
    - evaluate them against gold-standard clusterings
      - Find the best bipartite matching between the clusterings proposed by our algorithm and the gold clusterings
      - Use accuracy as the evaluation measure

Baseline Systems

1. **Spectral clustering** (Ng et al., 2001)
   - 2-means clustering using the second eigenvector

Baseline Systems

1. **Spectral clustering** (Ng et al., 2001)
    - 2-means clustering using the second eigenvector

2. **Non-Negative Matrix Factorization** (Xu et al., 2003)

Baseline Systems

1. **Spectral clustering** (Ng et al., 2001)
   - 2-means clustering using the second eigenvector

2. **Non-Negative Matrix Factorization** (Xu et al., 2003)

3. **Meta clustering** (Caruana et al., 2006)
   - 2-means with random weighting of features and initializations

Baseline Systems

1. **Spectral clustering** (Ng et al., 2001)
   - 2-means clustering using the second eigenvector

2. **Non-Negative Matrix Factorization** (Xu et al., 2003)

3. **Meta clustering** (Caruana et al., 2006)
   - 2-means with random weighting of features and initializations

4. **Iterative feature removal**
   - use Ng et al.'s spectral algorithm to produce a 2-way clustering
   - remove the informative features from each cluster
   - repeat these two steps if more clusterings are needed

Baseline Systems: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6

Baseline Systems: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6

Baseline Systems: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6

Baseline Systems: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6

Baseline Systems: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6

Baseline Systems: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6

- Best baseline: Ng et al.'s spectral clustering algorithm
- Worst baseline: NMF

Our Clustering Algorithm: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6
Our system	83.8	69.5	63.8	56.7	70.7	60.5	55.4	69.7	70.2

Our Clustering Algorithm: Results

	TNG	BOO			DVD			POL	
System	Topic	Sent.	Subj.	Stren.	Topic	Subj.	Stren.	Affili.	Policy
Spectral	89.8	58.9	58.8	51.5	54.9	61.5	54.9	54.3	67.6
NMF	85.2	52.1	57.8	50.7	50.3	60.5	51.9	53.0	61.1
Meta clustering	76.2	50.8	51.2	51.5	53.9	71.0	52.9	59.4	61.6
IFR	83.8	58.9	63.2	50.2	51.2	60.5	50.1	57.8	61.6
Our system	83.8	69.5	63.8	56.7	70.7	60.5	55.4	69.7	70.2

- Our system
  - often outperforms the best baseline for each dimension
  - achieves more stable performance across the dimensions

Summary of Contributions

- The insight that multiple kinds of clusterings in a dataset may be overlaid and should be teased apart to achieve a clustering along the desired dimension
- A novel application of spectral clustering
  - the insight that the eigenvectors of the Laplacian enable us to tease apart different kinds of clusterings of a text collection
- An intelligent choice of evaluation datasets can provide valuable algorithmic insights