Political Science 452: Text as Data

Justin Grimmer

Assistant Professor
Department of Political Science
Stanford University

May 4th, 2011

Where We've Been, Where We're Going

- Class 1: Finding Text Data
- Class 2: Representing Texts Quantitatively
- Class 3: Dictionary Methods for Classification
- Class 4: Comparing Language Across Groups
- Class 5: Texts in Space
- Class 6: Clustering
- Class 7: Topic models
- Class 8: Supervised methods for classification
- Class 9: Ensemble methods for classification
- Class 10: Scaling Speech

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
DocN	0	0	...	4

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
⋮	⋮	⋮	⋱	⋮
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare entire documents

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare entire documents
  - Place documents in space

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
⋮	⋮	⋮	⋱	⋮
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare entire documents
  - Place documents in space
  - Measure similarity of documents

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
⋮	⋮	⋮	⋱	⋮
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare entire documents
  - Place documents in space
  - Measure similarity of documents
  - Interpret word weighting geometrically

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
⋮	⋮	⋮	⋱	⋮
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare entire documents
  - Place documents in space
  - Measure similarity of documents
  - Interpret word weighting geometrically
  - Facilitate visualization of documents, based on similarity

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
⋮	⋮	⋮	⋱	⋮
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare entire documents
  - Place documents in space
  - Measure similarity of documents
  - Interpret word weighting geometrically
  - Facilitate visualization of documents, based on similarity
  - Kernel Trick: richer comparisons of documents (Spirling Paper)

Texts and Geometry

Term Document Matrix

Docs	Word1	Word2	...	Word M
Doc1	1	0	...	0
Doc2	0	3	...	1
⋮	⋮	⋮	⋱	⋮
DocN	0	0	...	4

Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare entire documents
  - Place documents in space
  - Measure similarity of documents
  - Interpret word weighting geometrically
  - Facilitate visualization of documents, based on similarity
  - Kernel Trick: richer comparisons of documents (Spirling Paper)
  - Basis for clustering, supervised learning

Texts in Space

Texts in Space

$$\text{Doc1} = (1, 1, 3, \ldots, 5)$$

Texts in Space

$$\begin{aligned}
\text{Doc1} &= (1, 1, 3, \ldots, 5) \\
\text{Doc2} &= (2, 0, 0, \ldots, 1)
\end{aligned}$$

Texts in Space

$$
\begin{aligned}
\text{Doc1} &= (1, 1, 3, \ldots, 5) \\
\text{Doc2} &= (2, 0, 0, \ldots, 1) \\
\textbf{Doc1}, \textbf{Doc2} &\in \Re^{M}
\end{aligned}
$$

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \\ \textbf{Doc1}, \textbf{Doc2} &\in \Re^M \end{aligned}$$

Provides many operations that will be useful

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \\ \textbf{Doc1}, \textbf{Doc2} &\in \Re^M \end{aligned}$$

Provides many operations that will be useful

Inner Product between documents:

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \\ \textbf{Doc1}, \textbf{Doc2} &\in \Re^M \end{aligned}$$

Provides many operations that will be useful

Inner Product between documents:

$$\textbf{Doc1} \cdot \textbf{Doc2} = (1, 1, 3, \ldots, 5)^{\prime}(2, 0, 0, \ldots, 1)$$

Texts in Space

$$\begin{aligned}
\text{Doc1} &= (1, 1, 3, \ldots, 5) \\
\text{Doc2} &= (2, 0, 0, \ldots, 1) \\
\textbf{Doc1}, \textbf{Doc2} &\in \Re^{M}
\end{aligned}$$

Provides many operations that will be useful

Inner Product between documents:

$$\begin{aligned}
\textbf{Doc1} \cdot \textbf{Doc2} &= (1, 1, 3, \ldots, 5)^{'} (2, 0, 0, \ldots, 1) \\
&= 1 \times 2 + 1 \times 0 + 3 \times 0 + \ldots + 5 \times 1
\end{aligned}$$

Texts in Space

$$
\begin{aligned}
\text{Doc1} &= (1, 1, 3, \ldots, 5) \\
\text{Doc2} &= (2, 0, 0, \ldots, 1) \\
\mathbf{Doc1}, \mathbf{Doc2} &\in \Re^M
\end{aligned}
$$

Provides many operations that will be useful
Inner Product between documents:

$$
\begin{aligned}
\mathbf{Doc1} \cdot \mathbf{Doc2} &= (1, 1, 3, \ldots, 5)^{'}(2, 0, 0, \ldots, 1) \\
&= 1 \times 2 + 1 \times 0 + 3 \times 0 + \ldots + 5 \times 1 \\
&= 7
\end{aligned}
$$

Length of document:

Length of document:

$$
\begin{aligned}
||\mathbf{Doc1}|| &\equiv \sqrt{\mathbf{Doc1} \cdot \mathbf{Doc1}} \\
&= \sqrt{(1,1,3,\ldots,5)'(1,1,3,\ldots,5)} \\
&= \sqrt{1^2 + 1^2 + 3^2 + 5^2} \\
&= 6
\end{aligned}
$$

Length of document:

$$
\begin{aligned}
||\mathbf{Doc1}|| &\equiv \sqrt{\mathbf{Doc1}\cdot\mathbf{Doc1}} \\
&= \sqrt{(1,1,3,\ldots,5)'(1,1,3,\ldots,5)} \\
&= \sqrt{1^2+1^2+3^2+5^2} \\
&= 6
\end{aligned}
$$

Cosine of the angle between documents:

Length of document:

$$
\begin{aligned}
||\textbf{Doc1}|| &\equiv \sqrt{\textbf{Doc1} \cdot \textbf{Doc1}} \\
&= \sqrt{(1,1,3,\ldots,5)'(1,1,3,\ldots,5)} \\
&= \sqrt{1^2 + 1^2 + 3^2 + 5^2} \\
&= 6
\end{aligned}
$$

Cosine of the angle between documents:

$$
\begin{aligned}
\cos\theta &\equiv \left(\frac{\textbf{Doc1}}{||\textbf{Doc1}||}\right) \cdot \left(\frac{\textbf{Doc2}}{||\textbf{Doc2}||}\right) \\
&= \frac{7}{6 \times 2.24} \\
&= 0.52
\end{aligned}
$$

Measuring Similarity

Documents in space $\rightarrow$ measure similarity/dissimilarity

Measuring Similarity

Documents in space $\rightarrow$ measure similarity/dissimilarity
What properties should similarity measure have?

Measuring Similarity

Documents in space $\rightarrow$ measure similarity/dissimilarity
What properties should similarity measure have?

- Maximum: document with itself

Measuring Similarity

Documents in space $\rightarrow$ measure similarity/dissimilarity
What properties should similarity measure have?

- Maximum: document with itself
- Minimum: documents have no words in common (orthogonal)

Measuring Similarity

Documents in space $\rightarrow$ measure similarity/dissimilarity
What properties should similarity measure have?

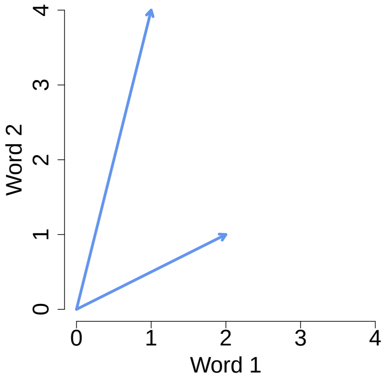
- Maximum: document with itself
- Minimum: documents have no words in common (orthogonal)
- Increasing when more of same words used

Measuring Similarity

Documents in space $\rightarrow$ measure similarity/dissimilarity
What properties should similarity measure have?

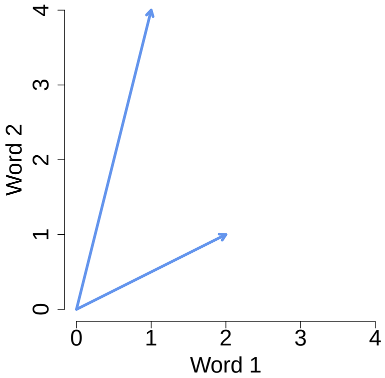
- Maximum: document with itself
- Minimum: documents have no words in common (orthogonal)
- Increasing when more of same words used
- ? $s(a, b) = s(b, a)$.

Measuring Similarity



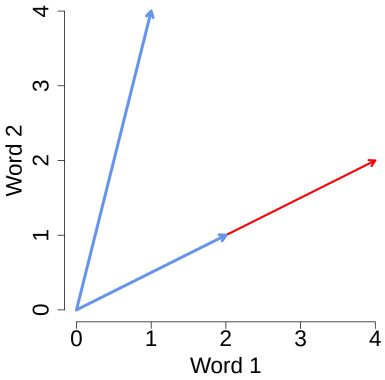
Measure 1: Inner product

Measuring Similarity

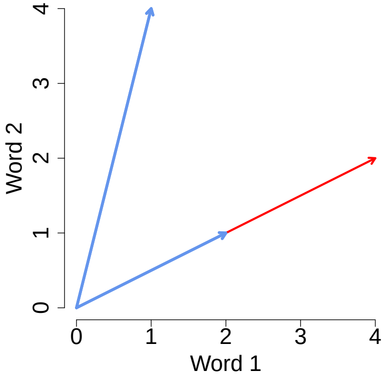


Measure 1: Inner product

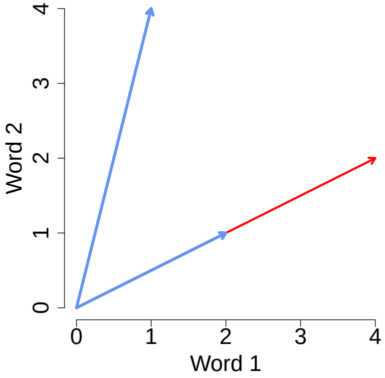
$$(2,1)' \cdot (1,4) \quad = \quad 6$$



Word 2
0
1
2
3
4
0
1
2
3
4
Word 1

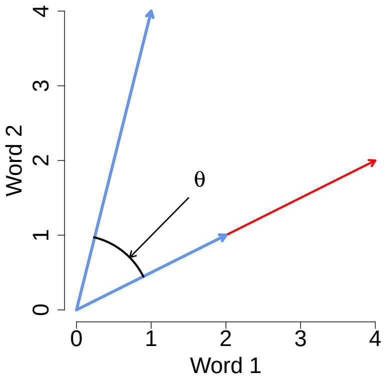


Problem(?): length dependent



Problem(?): length dependent

$$(4,2)^{'}(1,4) \quad = \quad 12$$



Problem(?): length dependent

$$(4,2)^{'}(1,4) = 12$$
$$a \cdot b = ||a|| \times ||b|| \times \cos\theta$$

Cosine Similarity

$\cos\theta$: removes document length from similarity measure

Cosine Similarity

$\cos\theta$: removes document length from similarity measure

$$\cos\theta = \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right)$$

Cosine Similarity

$\cos\theta$: removes document length from similarity measure

$$
\begin{aligned}
\cos\theta &= \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45)
\end{aligned}
$$

Cosine Similarity

$\cos\theta$: removes document length from similarity measure

$$
\begin{aligned}
\cos\theta &= \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45) \\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45)
\end{aligned}
$$

Cosine Similarity

$\cos\theta$: removes document length from similarity measure

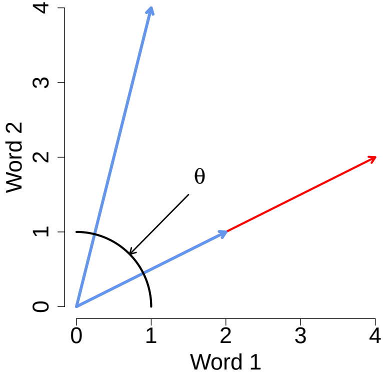
$$
\begin{aligned}
\cos\theta &= \left(\frac{a}{||a||}\right)\cdot\left(\frac{b}{||b||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45) \\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45) \\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97)
\end{aligned}
$$

Cosine Similarity

$\cos\theta$: removes document length from similarity measure

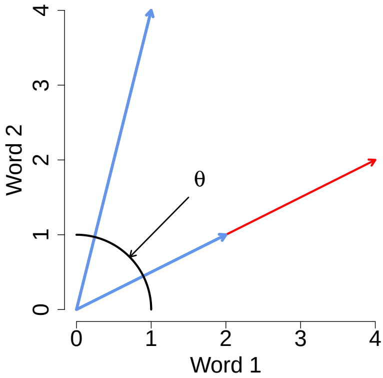
$$
\begin{aligned}
\cos\theta &= \left(\frac{a}{||a||}\right)\cdot\left(\frac{b}{||b||}\right)\\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45)\\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45)\\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97)\\
(0.89, 0.45)^{'}(0.24, 0.97) &= 0.65
\end{aligned}
$$

Cosine Similarity



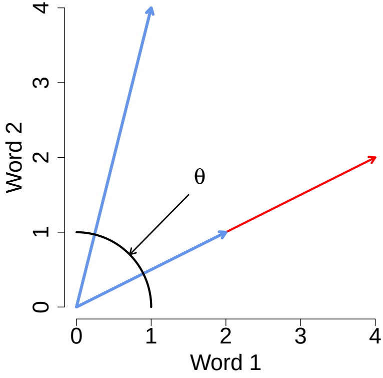
$\cos\theta$: removes document length from similarity measure

Cosine Similarity



$\cos\theta$: removes document length from similarity measure
Project onto Hypersphere

Cosine Similarity

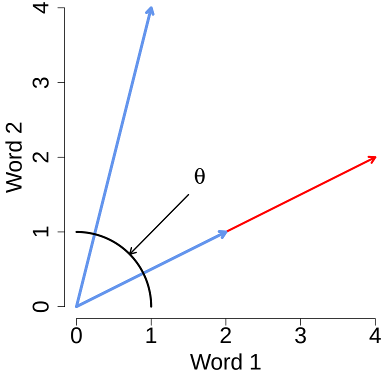


$\cos\theta$: removes document length from similarity measure

Project onto Hypersphere

$\cos\theta \rightarrow$ Inverse distance on Hypersphere

Cosine Similarity



$\cos\theta$: removes document length from similarity measure

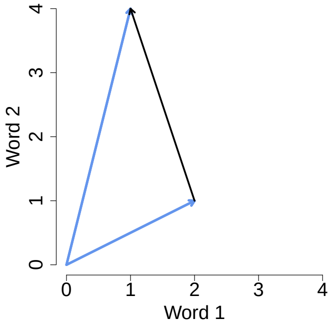
Project onto Hypersphere

$\cos\theta \rightarrow$ Inverse distance on Hypersphere

von Mises Fisher distribution : distribution on sphere surface

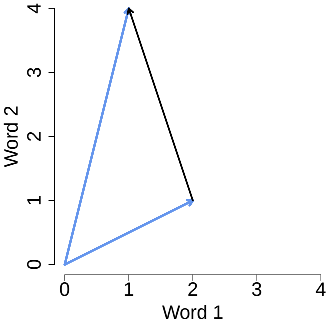
Measures of Dissimilarity

Measures of Dissimilarity



Measure distance or dissimilarity between documents

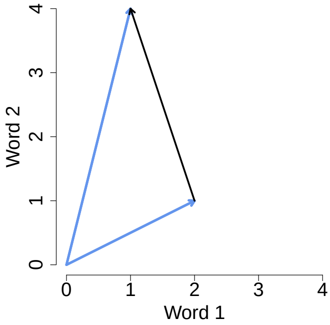
Measures of Dissimilarity



Measure distance or dissimilarity between documents

Euclidean distance:

Measures of Dissimilarity



Measure distance or dissimilarity between documents

Euclidean distance:

$$
\begin{aligned}
||\mathbf{a} - \mathbf{b}|| &= \sqrt{(a_1 - b_1)^2 + (a_2 + b_2)^2 + \ldots + (a_M - b_M)^2} \\
||(1,4) - (2,1)|| &= \sqrt{(1-2)^2 + (4-1)^2} \\
&= \sqrt{10}
\end{aligned}
$$

Measures of Dissimilarity

Many, Many Measures.

Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here

Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here
Manhattan metric

Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here
Manhattan metric

$$d_{\text{Man.}}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{M} |a_i - b_i|$$

Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here

Manhattan metric

$$
\begin{aligned}
d_{\text{Man.}}(\mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{M} |a_i - b_i| \\
d_{\text{Man.}}((1,4),(2,1)) &= |1| + |3| = 4
\end{aligned}
$$

Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here
Manhattan metric

$$
\begin{aligned}
d_{\text{Man.}}(\mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{M} |a_i - b_i| \\
d_{\text{Man.}}((1,4),(2,1)) &= |1| + |3| = 4
\end{aligned}
$$

Minkowski (p) metric

Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here
Manhattan metric

$$
\begin{aligned}
d_{\text{Man.}}(\mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{M} |a_i - b_i| \\
d_{\text{Man.}}((1,4),(2,1)) &= |1| + |3| = 4
\end{aligned}
$$

Minkowski (p) metric

$$
\begin{aligned}
d_p(\mathbf{a}, \mathbf{b}) &= \left(\sum_{i=1}^{M} (a_i - b_i)^p \right)^{1/p} \\
d_p((1,4),(2,1)) &= ((1-2)^p + (4-1)^p)^{1/p}
\end{aligned}
$$

What Does p Do?

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain maximum-metric

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \rightarrow \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) \quad = \quad \max_{i=1}^{M} |a_i - b_i|$$

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \rightarrow \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{M} |a_i - b_i|$$

Mapping Cosine similarity to dissimilarity

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{M} |a_i - b_i|$$

Mapping Cosine similarity to dissimilarity

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos \theta_{a,b}$$

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \rightarrow \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{M} |a_i - b_i|$$

Mapping Cosine similarity to dissimilarity

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos\theta_{a,b}$$

Quick proof that this makes sense

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{M} |a_i - b_i|$$

Mapping Cosine similarity to dissimilarity

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos \theta_{a,b}$$

Quick proof that this makes sense

- Restricted to nonnegative entries on documents

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{M} |a_i - b_i|$$

Mapping Cosine similarity to dissimilarity

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos\theta_{a,b}$$

Quick proof that this makes sense

- Restricted to nonnegative entries on documents
- Implies $\cos\theta \geq 0$

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \rightarrow \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{M} |a_i - b_i|$$

Mapping Cosine similarity to dissimilarity

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos\theta_{a,b}$$

Quick proof that this makes sense

- Restricted to nonnegative entries on documents
- Implies $\cos\theta \geq 0$
- $\cos\theta \leq 1$ (Cauchy-Schwartz)

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \rightarrow \infty$ Obtain maximum-metric

$$d_{\infty}(\mathbf{a}, \mathbf{b}) = \max_{i=1}^{M} |a_i - b_i|$$

Mapping Cosine similarity to dissimilarity

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos \theta_{a,b}$$

Quick proof that this makes sense

- Restricted to nonnegative entries on documents
- Implies $\cos \theta \geq 0$
- $\cos \theta \leq 1$ (Cauchy-Schwartz)
- $\cos \theta = 1 \iff \mathbf{a} = \mathbf{b}$

Weighting Words

Are all words created equal?

Weighting Words

Are all words created equal?

- Treat all words equally

Weighting Words

Are all words created equal?

- Treat all words equally
- Lots of noise

Weighting Words

Are all words created equal?

- Treat all words equally
- Lots of noise
- Reweight words

Weighting Words

Are all words created equal?

- Treat all words equally
- Lots of noise
- Reweight words
  - Accentuate words that are likely to be informative

Weighting Words

Are all words created equal?

- Treat all words equally
- Lots of noise
- Reweight words
  - Accentuate words that are likely to be informative
  - Make specific assumptions about characteristics of informative words

Weighting Words

Are all words created equal?

- Treat all words equally
- Lots of noise
- Reweight words
  - Accentuate words that are likely to be informative
  - Make specific assumptions about characteristics of informative words

How to generate weights?

Weighting Words

Are all words created equal?

- Treat all words equally
- Lots of noise
- Reweight words
  - Accentuate words that are likely to be informative
  - Make specific assumptions about characteristics of informative words

How to generate weights?

- Assumptions about separating words

Weighting Words

Are all words created equal?

- Treat all words equally
- Lots of noise
- Reweight words
  - Accentuate words that are likely to be informative
  - Make specific assumptions about characteristics of informative words

How to generate weights?

- Assumptions about separating words
- Use training set to identify separating words (Monroe, Ideology measurement)

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

$$n_j = \text{No. documents in which word } j \text{ occurs}$$

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

$$
\begin{aligned}
\mathrm{n}_j &= \text{No. documents in which word } j \text{ occurs} \\
\mathrm{idf}_j &= \log \frac{N}{n_j}
\end{aligned}
$$

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

$$\begin{aligned} \mathrm{n}_j &= \text{No. documents in which word } j \text{ occurs} \\ \mathrm{idf}_j &= \log \frac{N}{n_j} \\ \mathbf{idf} &= (\mathrm{idf}_1, \mathrm{idf}_2, \ldots, \mathrm{idf}_M) \end{aligned}$$

Weighting Words: TF-IDF Weighting

Why log ?

Weighting Words: TF-IDF Weighting

Why log ?

- Maximum at $n_j = 1$

Weighting Words: TF-IDF Weighting

Why log ?

- Maximum at $n_j = 1$
- Decreases at rate $\frac{1}{n_j} \Rightarrow$ diminishing "penalty" for more common use

Weighting Words: TF-IDF Weighting

Why log ?

- Maximum at $n_j = 1$
- Decreases at rate $\frac{1}{n_j} \Rightarrow$ diminishing "penalty" for more common use
- Other functional forms are fine, embed assumptions about penalization of common use

Weighting Words: TF-IDF

Weighting Words: TF-IDF

$$\mathbf{a}_{\text{idf}} \equiv \underbrace{\mathbf{a}}_{\text{tf}} \times \mathbf{idf} \quad = \quad (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M)$$

Weighting Words: TF-IDF

$$
\begin{aligned}
\mathbf{a}_{\text{idf}} \equiv \underbrace{\mathbf{a}}_{\text{tf}} \times \textbf{idf} &= (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M) \\
\mathbf{b}_{\text{idf}} \equiv \mathbf{b} \times \textbf{idf} &= (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M)
\end{aligned}
$$

Weighting Words: TF-IDF

$$\mathbf{a}_{\text{idf}} \equiv \underbrace{\mathbf{a}}_{\text{tf}} \times \textbf{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M)$$

$$\mathbf{b}_{\text{idf}} \equiv \mathbf{b} \times \textbf{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Weighting Words: TF-IDF

$$\mathbf{a}_{\text{idf}} \equiv \underbrace{\mathbf{a}}_{\text{tf}} \times \mathbf{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M)$$

$$\mathbf{b}_{\text{idf}} \equiv \mathbf{b} \times \mathbf{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M)$$

How Does This Matter For Measuring Similarity/Dissimilarity?
Inner Product

Weighting Words: TF-IDF

$$\mathbf{a}_{\text{idf}} \equiv \underbrace{\mathbf{a}}_{\text{tf}} \times \mathbf{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M)$$

$$\mathbf{b}_{\text{idf}} \equiv \mathbf{b} \times \mathbf{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M)$$

How Does This Matter For Measuring Similarity/Dissimilarity?
Inner Product

$$\mathbf{a}_{\text{idf}} \cdot \mathbf{b}_{\text{idf}} = (\mathbf{a} \times \mathbf{idf})'(\mathbf{b} \times \mathbf{idf})$$

Weighting Words: TF-IDF

$$\mathbf{a}_{\text{idf}} \equiv \underbrace{\mathbf{a}}_{\text{tf}} \times \mathbf{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M)$$

$$\mathbf{b}_{\text{idf}} \equiv \mathbf{b} \times \mathbf{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Inner Product

$$\begin{aligned}
\mathbf{a}_{\text{idf}} \cdot \mathbf{b}_{\text{idf}} &= (\mathbf{a} \times \mathbf{idf})^{'}(\mathbf{b} \times \mathbf{idf}) \\
&= (\text{idf}_1^2 \times a_1 \times b_1) + (\text{idf}_2^2 \times a_2 \times b_2) + \ldots + (\text{idf}_M^2 \times a_M \times b_M)
\end{aligned}$$

Weighting Words: Inner Product

Define:

Weighting Words: Inner Product

Define:

$$\mathbf{\Sigma} = \begin{pmatrix} \text{idf}_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \text{idf}_2^2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \text{idf}_M^2 \end{pmatrix}$$

Weighting Words: Inner Product

Define:

$$\mathbf{\Sigma} = \begin{pmatrix} \text{idf}_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \text{idf}_2^2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \text{idf}_M^2 \end{pmatrix}$$

We can then define the new inner product as

Weighting Words: Inner Product

Define:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{idf}_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \text{idf}_2^2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \text{idf}_M^2 \end{pmatrix}$$

We can then define the new inner product as

$$\mathbf{a}^{'}\boldsymbol{\Sigma}\mathbf{b} \quad = \quad \mathbf{a}_{\text{idf}} \cdot \mathbf{b}_{\text{idf}}$$

Weighting Words: Inner Product

Define:

$$\mathbf{\Sigma} = \begin{pmatrix} \text{idf}_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \text{idf}_2^2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \text{idf}_M^2 \end{pmatrix}$$

We can then define the new inner product as

$$\mathbf{a}^{'}\mathbf{\Sigma}\mathbf{b} = \mathbf{a}_{\text{idf}} \cdot \mathbf{b}_{\text{idf}}$$

Weighting Words: Inner Product

Why is this important?

Weighting Words: Inner Product

Why is this important?
Suggests general use of $\boldsymbol{\Sigma}$

Weighting Words: Inner Product

Why is this important?
Suggests general use of $\boldsymbol{\Sigma}$
If, for all $\mathbf{x}, \mathbf{y} \in \Re_{+}^{M}$

Weighting Words: Inner Product

Why is this important?
Suggests general use of $\boldsymbol{\Sigma}$
If, for all $\mathbf{x}, \mathbf{y} \in \Re_{+}^{M}$

$$\mathbf{x}^{'}\boldsymbol{\Sigma}\mathbf{y} \quad \geq \quad 0$$

Weighting Words: Inner Product

Why is this important?
Suggests general use of $\boldsymbol{\Sigma}$
If, for all $\mathbf{x}, \mathbf{y} \in \Re_{+}^{M}$

$$\mathbf{x}^{'}\boldsymbol{\Sigma}\mathbf{y} \geq 0$$

Then $\boldsymbol{\Sigma}$ defines a valid geometry

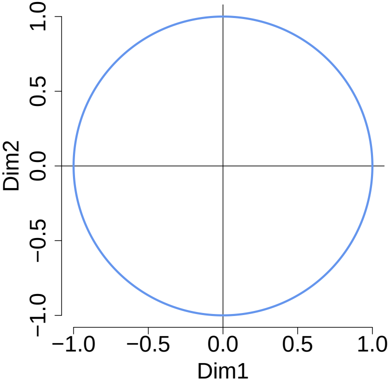
Weighting Words: Inner Product

Why is this important?
Suggests general use of $\boldsymbol{\Sigma}$
If, for all $\mathbf{x}, \mathbf{y} \in \Re_{+}^{M}$

$$\mathbf{x}^{'}\boldsymbol{\Sigma}\mathbf{y} \geq 0$$

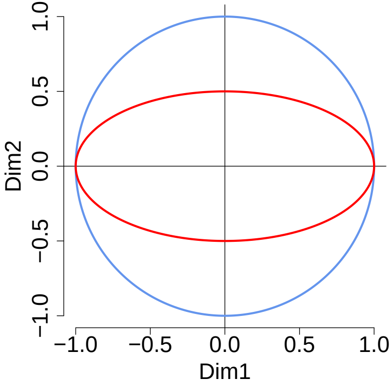
Then $\boldsymbol{\Sigma}$ defines a valid geometry
$\leadsto$ You can use $\boldsymbol{\Sigma}$ to modify similarity measures

Some Intuition: The Unit Circle



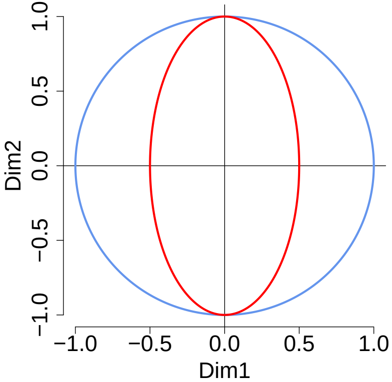
$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Some Intuition: The Unit Circle



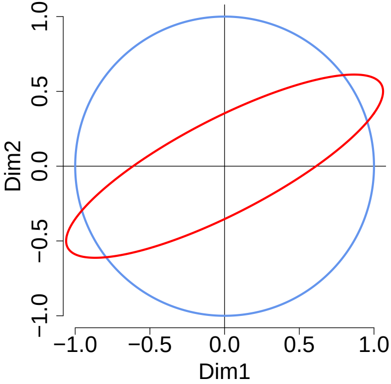
$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}$$

Some Intuition: The Unit Circle



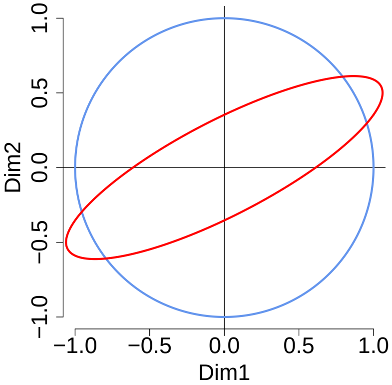
$$\mathbf{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$$

Some Intuition: The Unit Circle



$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

Some Intuition: The Unit Circle

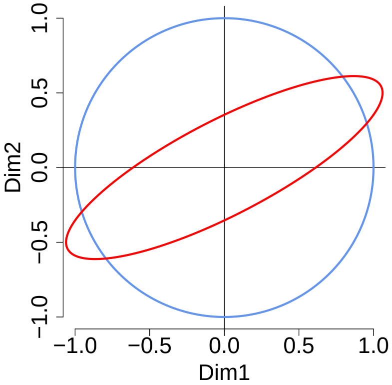


$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

Remember: Define inner product, define all other operations

$\boldsymbol{\Sigma}$ will be useful next week when clustering

Some Intuition: The Unit Circle



$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

Remember: Define inner product, define all other operations

$\boldsymbol{\Sigma}$ will be useful next week when clustering

Multidimensional Scaling and Projection

Multidimensional Scaling and Projection

Set of N documents, with M features.

Multidimensional Scaling and Projection

Set of N documents, with M features.
Use distance metric $d(\cdot, \cdot)$ to measure dissimilarities.

Multidimensional Scaling and Projection

Set of N documents, with M features.
Use distance metric $d(\cdot, \cdot)$ to measure dissimilarities.
Define $\mathbf{D}$ as $N \times N$ distance matrix

Multidimensional Scaling and Projection

Set of N documents, with M features.
Use distance metric $d(\cdot, \cdot)$ to measure dissimilarities.
Define $\mathbf{D}$ as $N \times N$ distance matrix

$$\mathbf{D} = \begin{pmatrix} 0 & d(1,2) & d(1,3) & \ldots & d(1,N) \\ d(2,1) & 0 & d(2,3) & \ldots & d(2,N) \\ d(3,1) & d(3,2) & 0 & \ldots & d(3,N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(N,1) & d(N,2) & d(N,3) & \ldots & 0 \end{pmatrix}$$

Multidimensional Scaling and Projection

Set of N documents, with M features.
Use distance metric $d(\cdot, \cdot)$ to measure dissimilarities.
Define $\mathbf{D}$ as $N \times N$ distance matrix

$$\mathbf{D} = \begin{pmatrix} 0 & d(1,2) & d(1,3) & \ldots & d(1,N) \\ d(2,1) & 0 & d(2,3) & \ldots & d(2,N) \\ d(3,1) & d(3,2) & 0 & \ldots & d(3,N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(N,1) & d(N,2) & d(N,3) & \ldots & 0 \end{pmatrix}$$

Lower Triangle contains unique information $N(N-1)/2$

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**.

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize
- Project into $\Re^J$, $J << M$

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize
- Project into $\Re^J$, $J << M$
- Key point: we will lose information

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize
- Project into $\Re^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\Re^J$ will not equal distances in $\Re^M$

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^{M}$
- Hard to visualize
- Project into $\Re^{J}$, $J << M$
- Key point: we will lose information
  - Distances between points in $\Re^{J}$ will not equal distances in $\Re^{M}$
- Why Project:

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize
- Project into $\Re^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\Re^J$ will not equal distances in $\Re^M$
- Why Project:
  - Identify systematic characteristics of data

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize
- Project into $\Re^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\Re^J$ will not equal distances in $\Re^M$
- Why Project:
  - Identify systematic characteristics of data
  - Visualize proximity

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize
- Project into $\Re^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\Re^J$ will not equal distances in $\Re^M$
- Why Project:
  - Identify systematic characteristics of data
  - Visualize proximity

Key question in Manifold learning (low-dimensional representation of high dimensional data):

Multidimensional Scaling and Projection

Learning low-dimensional structure of **D**. (Or: Machine Learning, 101)

- Assume: Documents reside in $\Re^M$
- Hard to visualize
- Project into $\Re^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\Re^J$ will not equal distances in $\Re^M$
- Why Project:
  - Identify systematic characteristics of data
  - Visualize proximity

Key question in Manifold learning (low-dimensional representation of high dimensional data):

What are the set of points in $\Re^J$ that "best" approximate points in $\Re^M$?

Classic Multidimensional Scaling Algorithms

Classic Multidimensional Scaling Algorithms

Begin: set of observations $\mathbf{Doc1}, \mathbf{Doc2}, \ldots, \mathbf{DocN} \in \Re^{M}$

Classic Multidimensional Scaling Algorithms

Begin: set of observations $\textbf{Doc1}, \textbf{Doc2}, \ldots, \textbf{DocN} \in \Re^{M}$
Goal: identify $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \Re^{J}$ that are "closest".

Classic Multidimensional Scaling Algorithms

Begin: set of observations $\mathbf{Doc1}, \mathbf{Doc2}, \ldots, \mathbf{DocN} \in \Re^{M}$
Goal: identify $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \Re^{J}$ that are "closest".
Classic MDS objective function

Classic Multidimensional Scaling Algorithms

Begin: set of observations $\textbf{Doc1}, \textbf{Doc2}, \ldots, \textbf{DocN} \in \Re^{M}$
Goal: identify $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \Re^{J}$ that are "closest".
Classic MDS objective function

$$\text{Stress}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} (d(\textbf{Doc}_j, \textbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2$$

Classic Multidimensional Scaling Algorithms

Begin: set of observations $\textbf{Doc1}, \textbf{Doc2}, \ldots, \textbf{DocN} \in \Re^M$
Goal: identify $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \Re^J$ that are "closest".
Classic MDS objective function

$$\text{Stress}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} (d(\textbf{Doc}_j, \textbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2$$

Identify $\mathbf{x}^*$ that minimizes the Stress

Classic Multidimensional Scaling Algorithms

Begin: set of observations $\textbf{Doc1}, \textbf{Doc2}, \ldots, \textbf{DocN} \in \Re^M$
Goal: identify $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \Re^J$ that are "closest".
Classic MDS objective function

$$\text{Stress}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} \left(d(\textbf{Doc}_j, \textbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i)\right)^2$$

Identify $\mathbf{x}^*$ that minimizes the Stress
`cmdscale` command in `R`

Classic MDS

Classic MDS

$\mathbf{x}^*$ is not unique.

Classic MDS

$\mathbf{x}^*$ is not unique.
If $\mathbf{x}^*$ minimize stress then all $\mathbf{x}^{**}$ that are rotations, translations, or shifts of $\mathbf{x}^*$ also minimize stress.

Classic MDS

$\mathbf{x}^*$ is not unique.
If $\mathbf{x}^*$ minimize stress then all $\mathbf{x}^{**}$ that are rotations, translations, or shifts of $\mathbf{x}^*$ also minimize stress.
Why?

Classic MDS

$\mathbf{x}^*$ is not unique.
If $\mathbf{x}^*$ minimize stress then all $\mathbf{x}^{**}$ that are rotations, translations, or shifts of $\mathbf{x}^*$ also minimize stress.
Why?

- Information only about relative positions

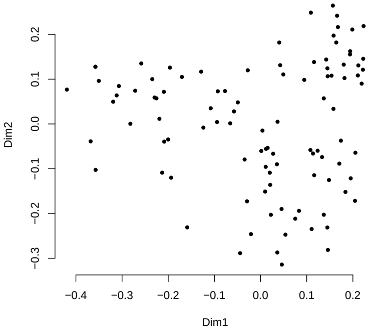
Classic MDS

$\mathbf{x}^*$ is not unique.
If $\mathbf{x}^*$ minimize stress then all $\mathbf{x}^{**}$ that are rotations, translations, or shifts of $\mathbf{x}^*$ also minimize stress.
Why?

- Information only about relative positions
- Many equivalent ways to place documents at same relative positions

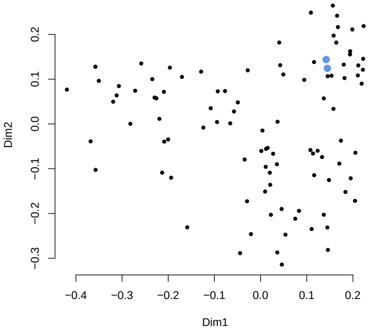
Visualizing Documents from Frank Lautenberg

Cosine dissimilarity, Classic MDS



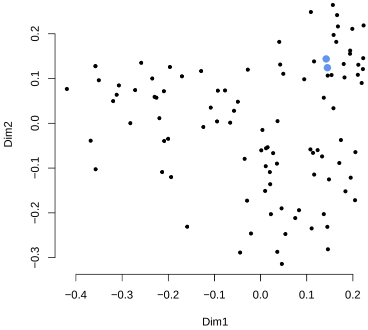
Visualizing Documents from Frank Lautenberg

Cosine dissimilarity, Classic MDS



Visualizing Documents from Frank Lautenberg

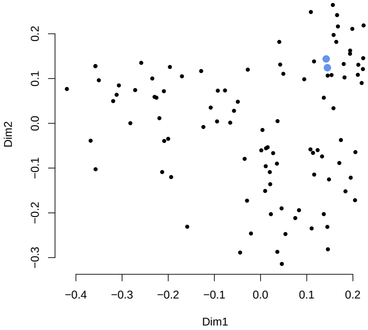
Cosine dissimilarity, Classic MDS



"The intolerance and discrimination we have seen from the Bush administration against gay and lesbian Americans is astounding, and anything but compassionate,"

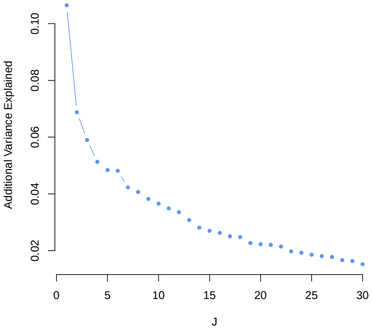
Visualizing Documents from Frank Lautenberg

Cosine dissimilarity, Classic MDS

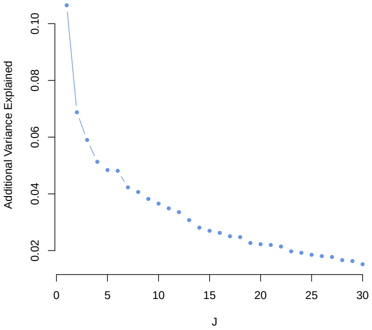


"Such a narrow-minded statement from the U.S. Secretary of Education is unacceptable...For Secretary Paige to say that the upbringing of one class of children offers superior morality compared to other children is offensive and hurtful to people of all other persuasions in America."

Classic Multidimensional Scaling Algorithms

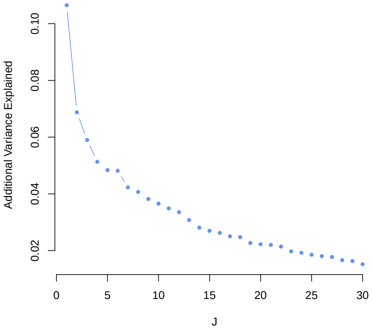


Classic Multidimensional Scaling Algorithms



What can we infer?

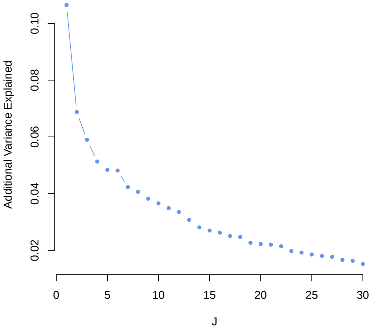
Classic Multidimensional Scaling Algorithms



What can we infer?

- Conditional on model, variance explained by factors

Classic Multidimensional Scaling Algorithms

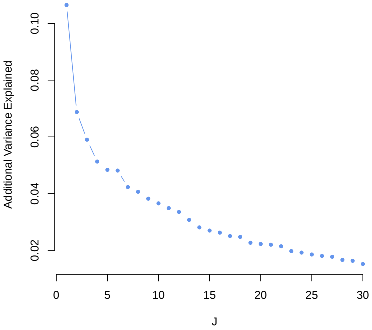


What can we infer?

- Conditional on model, variance explained by factors

What can't we infer?

Classic Multidimensional Scaling Algorithms



What can we infer?

- Conditional on model, variance explained by factors

What can't we infer?

- True Dimensionality

Sammon Multidimensional Scaling Algorithms

Many ways to infer low-dimensional structure from dissimilarities.

Sammon Multidimensional Scaling Algorithms

Many ways to infer low-dimensional structure from dissimilarities.
Consider one other method: Sammon Scaling

Sammon Multidimensional Scaling Algorithms

Many ways to infer low-dimensional structure from dissimilarities.

Consider one other method: Sammon Scaling

Classic MDS minimizes global stress

Sammon Multidimensional Scaling Algorithms

Many ways to infer low-dimensional structure from dissimilarities.

Consider one other method: Sammon Scaling

Classic MDS minimizes global stress

$$\text{Stress}(\mathbf{x}) \quad = \quad \sum_{j=2}^{N} \sum_{i<j} (d(\mathbf{Doc}_j, \mathbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2$$

Sammon Multidimensional Scaling Algorithms

Many ways to infer low-dimensional structure from dissimilarities.
Consider one other method: Sammon Scaling
Classic MDS minimizes global stress

$$\text{Stress}(\mathbf{x}) = \sum_{j=2}^{N}\sum_{i<j}(d(\mathbf{Doc}_j, \mathbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2$$

Often, we want a good approximation of neighborhoods (close to points), but don't care about far away distances

Sammon Multidimensional Scaling Algorithms

Many ways to infer low-dimensional structure from dissimilarities.
Consider one other method: Sammon Scaling
Classic MDS minimizes global stress

$$\text{Stress}(\mathbf{x}) = \sum_{j=2}^{N}\sum_{i<j}(d(\mathbf{Doc}_j, \mathbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2$$

Often, we want a good approximation of neighborhoods (close to points), but don't care about far away distances
Sammon Scaling

Sammon MDS

Sammon MDS

$$\text{Stress}_{\text{Sammon}}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} \frac{(d(\textbf{Doc}_j, \textbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2}{d(\textbf{Doc}_j, \textbf{Doc}_i)}$$

Sammon MDS

$$\text{Stress}_{\text{Sammon}}(\mathbf{x}) \quad = \quad \sum_{j=2}^{N} \sum_{i<j} \frac{\left(d(\mathbf{Doc}_j, \mathbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i)\right)^2}{d(\mathbf{Doc}_j, \mathbf{Doc}_i)}$$

Algorithm "cares" more about small distances $\rightsquigarrow$ prioritizes approximations for small distances

Sammon MDS

$$\text{Stress}_{\text{Sammon}}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} \frac{(d(\mathbf{Doc}_j, \mathbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2}{d(\mathbf{Doc}_j, \mathbf{Doc}_i)}$$

Algorithm "cares" more about small distances $\rightsquigarrow$ prioritizes approximations for small distances

`library(MASS)`

Sammon MDS

$$\text{Stress}_{\text{Sammon}}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} \frac{(d(\mathbf{Doc}_j, \mathbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2}{d(\mathbf{Doc}_j, \mathbf{Doc}_i)}$$

Algorithm "cares" more about small distances $\rightsquigarrow$ prioritizes approximations for small distances

```
library(MASS)
sammon
```

Sammon MDS

$$\text{Stress}_{\text{Sammon}}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} \frac{(d(\mathbf{Doc}_j, \mathbf{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2}{d(\mathbf{Doc}_j, \mathbf{Doc}_i)}$$

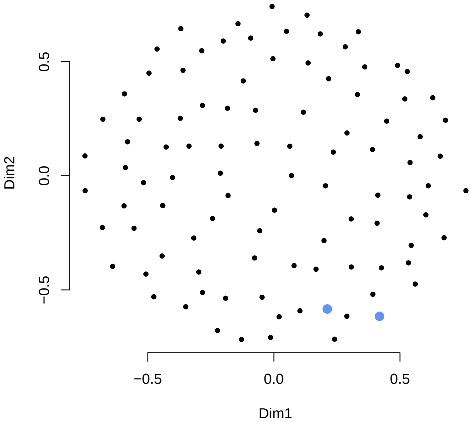
Algorithm "cares" more about small distances $\rightsquigarrow$ prioritizes approximations for small distances

`library(MASS)`

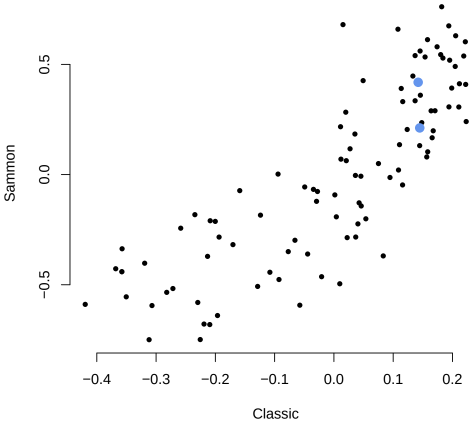
`sammon`

Pro tip: For all document $j \neq k$ $d(j,k) > 0$.

Comparing Sammon and Classic MDS



Comparing Sammon and Classic MDS



Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

- American political development

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Paper does a lot. We're going to focus on

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Paper does a lot. We're going to focus on

- Text representation and similarity calculation

Spirling and Indian Treaties

Spirling (2011): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Paper does a lot. We're going to focus on

- Text representation and similarity calculation
- Projecting to low dimensional space

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?

After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.

Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us
- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

```
- Peace Between Us

- No Peace Between Us
```

are identical.
Spirling uses complicated representation of texts to preserve word order $\rightsquigarrow$ quite useful

`Peace Between Us`

`Analyzes K-substrings`

Kernel Trick

Kernel Trick

- Kernel Methods: Represent texts, measure similarity

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$

$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$

$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

- If $a_n = 0$ or $b_n = 0$, then $a_n \times b_n = 0$.

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$

$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

- If $a_n = 0$ or $b_n = 0$, then $a_n \times b_n = 0$.
- Kernel Trick: Compare only substrings in both documents (without explicitly quantifying entire documents)

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$
$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

- If $a_n = 0$ or $b_n = 0$, then $a_n \times b_n = 0$.
- Kernel Trick: Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$

$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

- If $a_n = 0$ or $b_n = 0$, then $a_n \times b_n = 0$.
- Kernel Trick: Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$
$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

- If $a_n = 0$ or $b_n = 0$, then $a_n \times b_n = 0$.
- Kernel Trick: Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$

$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

- If $a_n = 0$ or $b_n = 0$, then $a_n \times b_n = 0$.
- Kernel Trick: Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Kernel Trick (Linear Algebra, 101) :

$$\mathbf{a} = (a_1, a_2, \ldots, a_K) \qquad \mathbf{b} = (b_1, b_2, \ldots, b_K)$$

$$\mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K$$

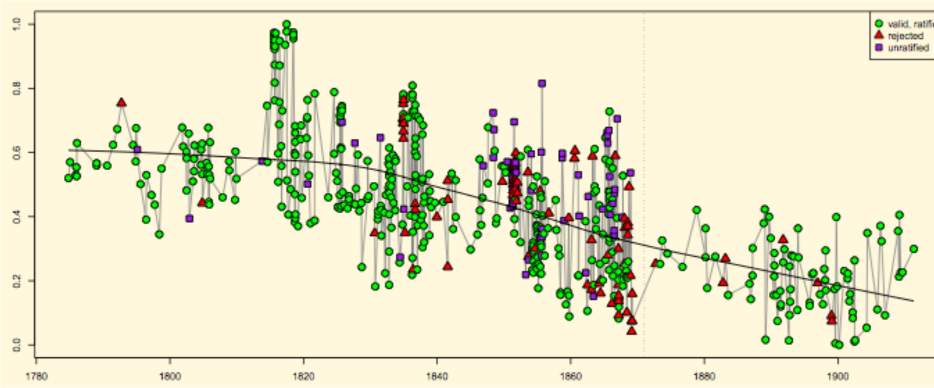
- If $a_n = 0$ or $b_n = 0$, then $a_n \times b_n = 0$.
- Kernel Trick: Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence Kernel : different sentences.

Kernel Trick

Apply kernel methods to simultaneously represent texts, measure similarity

- Creates dissimilarity matrix
- We can use projection methods to scale documents
- Spirling (2011): essentially uses classic MDS on dissimilarity measure

Harshness of Indian Treaties → Credible US Threats



Where We've Been Where We're Going

Today:

- Distance
- Projection

Next weeks:

- Clustering
- Topic Models
- Supervised learning

All require understanding material this week