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

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>
**Texts and Geometry**

**Term Document Matrix**

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>

Inference about documents:
- Word by word comparison
  - Dictionary methods
  - Class labelling methods
### Texts and Geometry

#### Term Document Matrix

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Inference about documents:

- Word by word comparison
  - Dictionary methods
  - Class labelling methods
- Compare **entire** documents
Texts and Geometry

Term Document Matrix

<table>
<thead>
<tr>
<th></th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td></td>
<td>:</td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
<td>...</td>
<td>:</td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>...</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
<td>...</td>
<td>:</td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Docs</th>
<th>Word1</th>
<th>Word2</th>
<th>\ldots</th>
<th>Word M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>0</td>
<td>\ldots</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>3</td>
<td>\ldots</td>
<td>1</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>DocN</td>
<td>0</td>
<td>0</td>
<td>\ldots</td>
<td>4</td>
</tr>
</tbody>
</table>

### Term Document Matrix

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

Inner Product between documents:

\[ \text{Doc}_1 \cdot \text{Doc}_2 = (1, 1, 3, \ldots, 5) \cdot (2, 0, 0, \ldots, 1) = 1 \times 2 + 1 \times 0 + 3 \times 0 + \ldots + 5 \times 1 = 7 \]
Texts in Space

\[ \text{Doc1} = (1, 1, 3, \ldots, 5) \]
Texts in Space

\[ \text{Doc1} = (1, 1, 3, \ldots, 5) \]
\[ \text{Doc2} = (2, 0, 0, \ldots, 1) \]
Texts in Space

\[
\begin{align*}
\text{Doc1} & = (1, 1, 3, \ldots, 5) \\
\text{Doc2} & = (2, 0, 0, \ldots, 1) \\
\text{Doc1, Doc2} & \in \mathbb{R}^M
\end{align*}
\]
Texts in Space

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

$$\text{Doc1, Doc2} \in \mathbb{R}^M$$

Provides many operations that will be useful
Texts in Space

\[ \text{Doc1} = (1, 1, 3, \ldots, 5) \]
\[ \text{Doc2} = (2, 0, 0, \ldots, 1) \]

Doc1, Doc2 ∈ \( \mathbb{R}^M \)

Provides many operations that will be useful

Inner Product between documents:
Texts in Space

\[
\text{Doc1} = (1, 1, 3, \ldots, 5) \\
\text{Doc2} = (2, 0, 0, \ldots, 1)
\]

**Doc1, Doc2 \in \mathbb{R}^M**

Provides many operations that will be useful

**Inner Product** between documents:

\[
\text{Doc1} \cdot \text{Doc2} = (1, 1, 3, \ldots, 5)'(2, 0, 0, \ldots, 1)
\]
Texts in Space

\[ \text{Doc1} = (1, 1, 3, \ldots, 5) \]
\[ \text{Doc2} = (2, 0, 0, \ldots, 1) \]
\[ \text{Doc1, Doc2} \in \mathbb{R}^M \]

Provides many operations that will be useful

**Inner Product** between documents:

\[ \text{Doc1} \cdot \text{Doc2} = (1, 1, 3, \ldots, 5)'(2, 0, 0, \ldots, 1) \]
\[ = 1 \times 2 + 1 \times 0 + 3 \times 0 + \ldots + 5 \times 1 \]
Texts in Space

\[
\begin{align*}
\text{Doc1} & = (1, 1, 3, \ldots, 5) \\
\text{Doc2} & = (2, 0, 0, \ldots, 1) \\
\text{Doc1, Doc2} & \in \mathbb{R}^M
\end{align*}
\]

Provides many operations that will be useful

**Inner Product** between documents:

\[
\text{Doc1} \cdot \text{Doc2} = (1, 1, 3, \ldots, 5)'(2, 0, 0, \ldots, 1) = 1 \times 2 + 1 \times 0 + 3 \times 0 + \ldots + 5 \times 1 = 7
\]
Length of document:
Length of document:

\[ \| \text{Doc1} \| \equiv \sqrt{\text{Doc1} \cdot \text{Doc1}} \]
\[ = \sqrt{(1, 1, 3, \ldots, 5)'(1, 1, 3, \ldots, 5)} \]
\[ = \sqrt{1^2 + 1^2 + 3^2 + 5^2} \]
\[ = 6 \]
**Length of document:**

\[
\|\text{Doc1}\| \equiv \sqrt{\text{Doc1} \cdot \text{Doc1}}
\]

\[
= \sqrt{(1, 1, 3, \ldots, 5)'(1, 1, 3, \ldots, 5)}
\]

\[
= \sqrt{1^2 + 1^2 + 3^2 + 5^2}
\]

\[
= 6
\]

**Cosine of the angle between documents:**
Length of document:

$$\|\text{Doc1}\| \equiv \sqrt{\text{Doc1} \cdot \text{Doc1}}$$

$$= \sqrt{(1, 1, 3, \ldots, 5)'(1, 1, 3, \ldots, 5)}$$

$$= \sqrt{1^2 + 1^2 + 3^2 + 5^2}$$

$$= 6$$

Cosine of the angle between documents:

$$\cos \theta \equiv \left( \frac{\text{Doc1}}{\|\text{Doc1}\|} \right) \cdot \left( \frac{\text{Doc2}}{\|\text{Doc2}\|} \right)$$

$$= \frac{7}{6 \times 2.24}$$

$$= 0.52$$
Measuring Similarity

Documents in space → 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
- Symmetric: $(a, b) = (b, a)$. 
Measuring Similarity

Documents in space → 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

Documents in space → measure similarity/dissimilarity
What properties should similarity measure have?

- Maximum: document with itself
Measuring Similarity

Documents in space → 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 → 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

Documents in space → 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
- ? \( s(a, b) = s(b, a) \).
Measure 1: Inner product
Measuring Similarity

Measure 1: Inner product

\[(2, 1)' \cdot (1, 4) = 6\]
Problem: length dependent

\[ (4, 2) \cdot (1, 4) = 12 \]

\[ a \cdot b = ||a|| \times ||b|| \times \cos \theta \]
Problem(?): length dependent
Problem(?): length dependent

\[(4, 2)'(1, 4) = 12\]
Problem(?): length dependent

\[(4, 2)'(1, 4) = 12\]

\[a \cdot b = ||a|| \times ||b|| \times \cos \theta\]
Cosine Similarity

\[ \cos \theta : \text{removes document length from similarity measure} \]
Cosine Similarity

cos \theta: \text{ removes document length from similarity measure}

\[
\cos \theta = \left( \frac{a}{\|a\|} \right) \cdot \left( \frac{b}{\|b\|} \right)
\]
Cosine Similarity

\[ \cos \theta: \text{removes document length from similarity measure} \]

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

\[
\frac{(4, 2)}{\| (4, 2) \|} = (0.89, 0.45)
\]
Cosine Similarity

\[ \cos \theta : \text{removes document length from similarity measure} \]

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

\[
\frac{(4, 2)}{\| (4, 2) \|} = (0.89, 0.45)
\]

\[
\frac{(2, 1)}{\| (2, 1) \|} = (0.89, 0.45)
\]
Cosine Similarity

\[ \cos \theta : \text{removes document length from similarity measure} \]

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

\[
\frac{(4, 2)}{\|(4, 2)\|} = (0.89, 0.45)
\]

\[
\frac{(2, 1)}{\|(2, 1)\|} = (0.89, 0.45)
\]

\[
\frac{(1, 4)}{\|(1, 4)\|} = (0.24, 0.97)
\]
Cosine Similarity

\[ \cos \theta : \text{removes document length from similarity measure} \]

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

\[
\frac{(4, 2)}{\| (4, 2) \|} = (0.89, 0.45)
\]

\[
\frac{(2, 1)}{\| (2, 1) \|} = (0.89, 0.45)
\]

\[
\frac{(1, 4)}{\| (1, 4) \|} = (0.24, 0.97)
\]

\[
(0.89, 0.45)'(0.24, 0.97) = 0.65
\]
Cosine Similarity

\[ \cos \theta: \text{removes document length from similarity measure} \]
Cosine Similarity

\[ \cos \theta : \text{removes document length from similarity measure} \]

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

\[ \cos \theta : \text{removes document length from similarity measure} \]

Project onto Hypersphere

\[ \cos \theta \rightarrow \text{Inverse distance on Hypersphere} \]

von Mises Fisher distribution : distribution on sphere surface
Measures of Dissimilarity

Euclidean distance:
$$||a - b|| = \sqrt{ (a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_M - b_M)^2 }$$

For example:
$$|| (1, 4) - (2, 1) || = \sqrt{ (1 - 2)^2 + (4 - 1)^2 } = \sqrt{ 10 }$$
Measures of Dissimilarity

Measure **distance** or **dissimilarity** between documents
Measures of Dissimilarity

Measure distance or dissimilarity between documents

Euclidean distance:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_M - b_M)^2}$$
Measures of Dissimilarity

Measure distance or dissimilarity between documents

Euclidean distance:

\[ \|a - b\| = \sqrt{(a_1 - b_1)^2 + (a_2 + b_2)^2 + \ldots + (a_M - b_M)^2} \]

\[ \|(1, 4) - (2, 1)\| = \sqrt{(1 - 2)^2 + (4 - 1)^2} \]

\[ = \sqrt{10} \]
Measures of Dissimilarity

Many, Many Measures.

\[ d_{\text{Man.}}((a, b), (c, d)) = |a - c| + |b - d| \]

\[ d_p((a, b), (c, d)) = \left( \sum_{i=1}^{n} |a_i - c_i|^p \right)^{1/p} \]
Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here
Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here
Manhattan metric

\[\begin{align*}
\text{Manhattan metric} & \quad d_{\text{Man.}}((a, b)) = \sum_{i=1}^{n} |a_i - b_i| \\
& \quad d_{\text{Man.}}((1, 4), (2, 1)) = |1 - 2| + |3 - 1| = 4
\end{align*}\]
Measures of Dissimilarity

**Many, Many Measures.** Cover Minkowski family here

**Manhattan metric**

\[ d_{\text{Man.}}(a, b) = \sum_{i=1}^{M} |a_i - b_i| \]
Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here

Manhattan metric

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

\[ d_{\text{Man.}}((1, 4), (2, 1)) = |1| + |3| = 4 \]
Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here

Manhattan metric

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

\[ d_{\text{Man.}}((1, 4), (2, 1)) = |1| + |3| = 4 \]

Minkowski (p) metric
Measures of Dissimilarity

Many, Many Measures. Cover Minkowski family here

Manhattan metric

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

\[ d_{\text{Man.}}((1, 4), (2, 1)) = |1| + |3| = 4 \]

Minkowski (p) metric

\[ d_p(\mathbf{a}, \mathbf{b}) = \left( \sum_{i=1}^{M} (a_i - b_i)^p \right)^{1/p} \]

\[ d_p((1, 4), (2, 1)) = ((1 - 2)^p + (4 - 1)^p)^{1/p} \]
What Does $p$ Do?

Increasing $p \Rightarrow \text{greater importance of coordinates with largest differences}$

If we let $p \to \infty$

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

Mapping Cosine similarity to dissimilarity $d_{\text{cos}}(a, b) = 1 - \cos \theta_{a, b}$

Quick proof that this makes sense
- Restricted to nonnegative entries on documents
- Implies $\cos \theta \geq 0$
- $\cos \theta \leq 1$ (Cauchy-Schwartz)
- $\cos \theta = 1 \iff a = b$
What Does $p$ Do?

Increasing $p$ $\leadsto$ greater importance of coordinates with largest differences
What Does $p$ Do?

Increasing $p \sim$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain maximum-metric
What Does $p$ Do?

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

If we let $p \to \infty$ Obtain maximum-metric

$$d_\infty(a, b) = \max_{i=1}^M |a_i - b_i|$$
What Does $p$ Do?

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

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

Mapping Cosine similarity to dissimilarity
What Does $p$ Do?

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

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

Mapping Cosine similarity to dissimilarity

$$d_{\cos}(a, b) = 1 - \cos \theta_{a,b}$$
What Does $p$ Do?

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

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

Mapping Cosine similarity to dissimilarity

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

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

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

Mapping Cosine similarity to dissimilarity

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

Quick proof that this makes sense
- Restricted to nonnegative entries on documents
- Implies $\cos \theta \geq 0$
What Does $p$ Do?

Increasing $p \rightarrow \infty$ implies greater importance of coordinates with largest differences. If we let $p \rightarrow \infty$ we obtain maximum-metric

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

Mapping Cosine similarity to dissimilarity

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

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

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

Mapping Cosine similarity to dissimilarity

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

Quick proof that this makes sense
- Restricted to nonnegative entries on documents
- Implies $\cos \theta \geq 0$
- $\cos \theta \leq 1$ (Cauchy-Schwartz)
- $\cos \theta = 1 \iff a = b$
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?

Inverse document frequency:

\[
\text{idf}_j = \log \frac{N}{n_j}
\]

where:

- \( N \) is the total number of documents
- \( n_j \) is the number of documents in which word \( j \) occurs
Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?
- Used frequently

Inverse document frequency:

$$\text{idf}_j = \log \frac{N}{n_j}$$

where

- $N$ is the total number of documents
- $n_j$ is the number of documents in which word $j$ occurs

$$\text{idf} = (\text{idf}_1, \text{idf}_2, \ldots, \text{idf}_M)$$
Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?
- Used frequently
- But not too frequently

Inverse document frequency:

\[ \text{idf}_j = \log \frac{N}{n_j} \]

where:

- \( N \) is the number of documents in the collection
- \( n_j \) is the number of documents in which word \( j \) 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
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:

\[ n_j = \text{No. documents in which word } j \text{ occurs} \]
\[ \text{idf}_j = \log \frac{N}{n_j} \]
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:

\[
\begin{align*}
  n_j &= \text{No. documents in which word } j \text{ occurs} \\
  \text{idf}_j &= \log \frac{N}{n_j} \\
  \text{idf} &= (\text{idf}_1, \text{idf}_2, \ldots, \text{idf}_M)
\end{align*}
\]
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

\[
a_{\text{idf}} \equiv \left( a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M \right)
\]

\[
b_{\text{idf}} \equiv \left( b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M \right)
\]

How Does This Matter For Measuring Similarity/Dissimilarity?

\[
\mathbf{a}_{\text{idf}} \cdot \mathbf{b}_{\text{idf}} = \left( a_1 \times \text{idf}_1 \right)' \left( b_1 \times \text{idf}_1 \right) + \left( a_2 \times \text{idf}_2 \right)' \left( b_2 \times \text{idf}_2 \right) + \ldots + \left( a_M \times \text{idf}_M \right)' \left( b_M \times \text{idf}_M \right)
\]
Weighting Words: TF-IDF

\[ a_{idf} \equiv \underbrace{a_{tf}} \times \text{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M) \]
Weighting Words: TF-IDF

\[ a_{idf} \equiv a_{tf} \times \text{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M) \]

\[ b_{idf} \equiv b \times \text{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M) \]
Weighting Words: TF-IDF

\[ a_{idf} \equiv a_{\text{tf}} \times \text{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M) \]

\[ b_{idf} \equiv b \times \text{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M) \]

How Does This Matter For Measuring Similarity/Dissimilarity?
Weighting Words: TF-IDF

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

\[
b_{\text{idf}} \equiv b \times \text{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M)
\]

How Does This Matter For Measuring Similarity/Dissimilarity?

Inner Product
Weighting Words: TF-IDF

\[ a_{idf} \equiv \left( a \times \text{idf} \right) = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M) \]

\[ b_{idf} \equiv \left( b \times \text{idf} \right) = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M) \]

How Does This Matter For Measuring Similarity/Dissimilarity?

Inner Product

\[ a_{idf} \cdot b_{idf} = (a \times \text{idf})' (b \times \text{idf}) \]
Weighting Words: TF-IDF

\[ a_{\text{idf}} \equiv a_{\text{tf}} \times \text{idf} = (a_1 \times \text{idf}_1, a_2 \times \text{idf}_2, \ldots, a_M \times \text{idf}_M) \]

\[ b_{\text{idf}} \equiv b \times \text{idf} = (b_1 \times \text{idf}_1, b_2 \times \text{idf}_2, \ldots, b_M \times \text{idf}_M) \]

How Does This Matter For Measuring Similarity/Dissimilarity?

**Inner Product**

\[ a_{\text{idf}} \cdot b_{\text{idf}} = (a \times \text{idf})' (b \times \text{idf}) \]

\[ = (\text{idf}_1^2 \times a_1 \times b_1) + (\text{idf}_2^2 \times a_2 \times b_2) + \ldots + (\text{idf}_M^2 \times a_M \times b_M) \]
Weighting Words: Inner Product

Define:

\[
\Sigma = \begin{bmatrix}
idf_2 & 1 & 0 & 0 & \ldots & 0 \\
1 & idf_2 & 0 & \ldots & 0 \\
0 & 0 & idf_2 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & idf_M
\end{bmatrix}
\]

We can then define the new inner product as

\[a' \Sigma b = a_{idf} \cdot b_{idf}\]
Weighting Words: Inner Product

Define:

\[ \Sigma = \begin{pmatrix} \text{idf}_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \text{idf}_2^2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \text{idf}_M^2 \end{pmatrix} \]
Weighting Words: Inner Product

Define:

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

We can then define the new inner product as
Weighting Words: Inner Product

Define:

\[ \Sigma = \begin{pmatrix}
idf_1^2 & 0 & 0 & \ldots & 0 \\
0 & idf_2^2 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & idf_M^2
\end{pmatrix} \]

We can then define the new inner product as

\[ a^\prime \Sigma b = a_{idf} \cdot b_{idf} \]
Weighting Words: Inner Product

Define:

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

We can then define the new inner product as

\[
a^\top \Sigma b = a_{\text{idf}} \cdot b_{\text{idf}}
\]
Weighting Words: Inner Product

Why is this important?
Weighting Words: Inner Product

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

If, for all $x, y \in \mathbb{R}^M$, $x' \Sigma y \geq 0$
Then $\Sigma$ defines a valid geometry

You can use $\Sigma$ to modify similarity measures
Inferences will depend upon choice of $\Sigma$
Weighting Words: Inner Product

Why is this important?
Suggests general use of $\Sigma$
If, for all $x, y \in \mathbb{R}^M$

$\Sigma$ defines a valid geometry
Weighting Words: Inner Product

Why is this important?
Suggests general use of $\Sigma$
If, for all $x, y \in \mathbb{R}^M$

$$x^T \Sigma y \geq 0$$
Weighting Words: Inner Product

Why is this important?
Suggests general use of $\Sigma$
If, for all $x, y \in \mathbb{R}^M$

$$x' \Sigma y \geq 0$$

Then $\Sigma$ defines a valid geometry
Why is this important?
Suggests general use of $\Sigma$
If, for all $x, y \in \mathbb{R}^M$

$$x' \Sigma y \geq 0$$

Then $\Sigma$ defines a valid geometry
$\rightsquigarrow$ You can use $\Sigma$ to modify similarity measures
Some Intuition: The Unit Circle

\[ \sum = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]
Some Intuition: The Unit Circle

\[ \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix} \]
Some Intuition: The Unit Circle

\[ \Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} \]
Some Intuition: The Unit Circle

\[ \Sigma = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix} \]
Some Intuition: The Unit Circle

\[ \Sigma = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix} \]

Remember: Define inner product, define all other operations  
\( \Sigma \) will be useful next week when clustering
Some Intuition: The Unit Circle

\[ \Sigma = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix} \]

Remember: Define inner product, define all other operations
\( \Sigma \) will be useful next week when clustering
Multidimensional Scaling and Projection

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

Define $D$ as an $N \times N$ distance matrix:

$$D = \begin{bmatrix}
0 & d(1, 2) & d(1, 3) & \cdots & d(1, N) \\
d(2, 1) & 0 & d(2, 3) & \cdots & d(2, N) \\
d(3, 1) & d(3, 2) & 0 & \cdots & d(3, N) \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
d(N, 1) & d(N, 2) & d(N, 3) & \cdots & 0
\end{bmatrix}$$

The lower triangle contains unique information $N(N - 1)/2$.
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.
Multidimensional Scaling and Projection

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

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

Lower Triangle contains unique information $N(N-1)/2$
Multidimensional Scaling and Projection

Set of $N$ documents, with $M$ features.

Use distance metric $d(\cdot, \cdot)$ to measure dissimilarities.

Define $D$ as $N \times N$ distance matrix

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

Lower Triangle contains unique information $N(N - 1)/2$
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$.
Multidimensional Scaling and Projection

Learning low-dimensional structure of $D$. (Or: Machine Learning, 101)
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

- **Assume**: Documents reside in $\mathbb{R}^M$

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

What are the set of points in $\mathbb{R}^J$ that "best" approximate points in $\mathbb{R}^M$?
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

- **Assume**: Documents reside in $\mathbb{R}^M$
- Hard to visualize

Key question in Manifold learning (low-dimensional representation of high-dimensional data): What are the set of points in $\mathbb{R}^J$ that "best" approximate points in $\mathbb{R}^M$?
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

- **Assume**: Documents reside in $\mathbb{R}^M$
- Hard to visualize
- **Project** into $\mathbb{R}^J$, $J << M$
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

- **Assume**: Documents reside in $\mathbb{R}^M$
- Hard to visualize
- **Project** into $\mathbb{R}^J$, $J << M$
- **Key point**: we will lose information
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

- **Assume:** Documents reside in $\mathbb{R}^M$
- Hard to visualize
- **Project** into $\mathbb{R}^J$, $J << M$
- **Key point:** we will lose information
  - Distances between points in $\mathbb{R}^J$ will not equal distances in $\mathbb{R}^M$
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

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

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

- **Assume**: Documents reside in $\mathbb{R}^M$
- Hard to visualize
- **Project** into $\mathbb{R}^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\mathbb{R}^J$ will not equal distances in $\mathbb{R}^M$
- Why Project:
  - Identify **systematic** characteristics of data
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

- **Assume**: Documents reside in $\mathbb{R}^M$
- Hard to visualize
- **Project** into $\mathbb{R}^J$, $J << M$
- **Key point**: we will lose information
  - Distances between points in $\mathbb{R}^J$ will not equal distances in $\mathbb{R}^M$
- **Why Project**:
  - Identify **systematic** characteristics of data
  - Visualize **proximity**
Multidimensional Scaling and Projection

Learning low-dimensional structure of $\mathbf{D}$. (Or: Machine Learning, 101)

- **Assume:** Documents reside in $\mathbb{R}^M$
- Hard to visualize
- **Project** into $\mathbb{R}^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\mathbb{R}^J$ will not equal distances in $\mathbb{R}^M$
- 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 $\mathbb{R}^M$
- Hard to visualize
- **Project** into $\mathbb{R}^J$, $J << M$
- Key point: we will lose information
  - Distances between points in $\mathbb{R}^J$ will not equal distances in $\mathbb{R}^M$
- 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 $\mathbb{R}^J$ that “best” approximate points in $\mathbb{R}^M$?
Classic Multidimensional Scaling Algorithms

Begin: set of observations Doc1, Doc2, ..., DocN ∈ ℜM

Goal: identify x1, x2, ..., xN ∈ ℜJ that are "closest".

Classic MDS objective function

\[ \text{Stress}(x) = \sum_{j=2}^{N} \sum_{i<j} (d(Doc_j, Doc_i) - d(x_j, x_i))^2 \]

Identify x∗ that minimizes the Stress

cmdscale command in R
Classic Multidimensional Scaling Algorithms

Begin: set of observations $\text{Doc}_1, \text{Doc}_2, \ldots, \text{Doc}_N \in \mathbb{R}^M$
Classic Multidimensional Scaling Algorithms

Begin: set of observations $\text{Doc}1, \text{Doc}2, \ldots, \text{Doc}N \in \mathbb{R}^M$
Goal: identify $x_1, x_2, \ldots, x_N \in \mathbb{R}^J$ that are “closest”.

Classic MDS objective function

$$\text{Stress}(x) = \sum_{j=2}^{N} \sum_{i<j} \left( d(\text{Doc}j, \text{Doc}i) - d(x_j, x_i) \right)^2$$

Identify $x^*$ that minimizes the Stress

$\text{cmdscale}$ command in \texttt{R}
Classic Multidimensional Scaling Algorithms

Begin: set of observations $\text{Doc}_1, \text{Doc}_2, \ldots, \text{Doc}_N \in \mathbb{R}^M$
Goal: identify $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbb{R}^J$ that are “closest”.
Classic MDS objective function

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

Begin: set of observations $\text{Doc1, Doc2, \ldots, DocN} \in \mathbb{R}^M$

Goal: identify $x_1, x_2, \ldots, x_N \in \mathbb{R}^J$ that are “closest”.

Classic MDS objective function

$$\text{Stress}(x) = \sum_{j=2}^{N} \sum_{i<j} (d(\text{Doc}_j, \text{Doc}_i) - d(x_j, x_i))^2$$
Classic Multidimensional Scaling Algorithms

Begin: set of observations $\text{Doc}_1, \text{Doc}_2, \ldots, \text{Doc}_N \in \mathbb{R}^M$

Goal: identify $x_1, x_2, \ldots, x_N \in \mathbb{R}^J$ that are “closest”.

Classic MDS objective function

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

Identify $x^*$ that minimizes the Stress
Classic Multidimensional Scaling Algorithms

Begin: set of observations $\text{Doc1, Doc2, ... , DocN} \in \mathbb{R}^M$

Goal: identify $x_1, x_2, \ldots, x_N \in \mathbb{R}^J$ that are “closest”.

Classic MDS objective function

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

Identify $x^*$ that minimizes the Stress

cmdscale command in R
Classic MDS

If $x^*$ minimize stress then all $x^{**}$ that are rotations, translations, or shifts of $x^*$ also minimize stress. Why?

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

\( \mathbf{x}^* \) is not unique.
Classic MDS

\( x^* \) is not unique.

If \( x^* \) minimize stress then all \( x^{**} \) that are rotations, translations, or shifts of \( x^* \) also minimize stress.
Classic MDS

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

\( \mathbf{x}^\ast \) is not unique.

If \( \mathbf{x}^\ast \) minimize stress then all \( \mathbf{x}^{\ast\ast} \) that are rotations, translations, or shifts of \( \mathbf{x}^\ast \) 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
"The intolerance and discrimination we have seen from the Bush administration against gay and lesbian Americans is astounding, and anything but compassionate,"

"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

What can we infer?
- Conditional on model, variance explained by factors

What can’t we infer?
- True Dimensionality

Justin Grimmer (Stanford University)

Text as Data

May 4th, 2011 25 / 34
What can we infer?

- Conditional on model, variance explained by factors
- True Dimensionality
What can we infer?
- Conditional on model, variance explained by factors
What can we infer?
- Conditional on model, variance explained by factors

What can’t we infer?

What can we infer?
- Conditional on model, variance explained by factors

What can’t we infer?
- True Dimensionality
Many ways to infer low-dimensional structure from dissimilarities.
Many ways to infer low-dimensional structure from dissimilarities. Consider one other method: Sammon Scaling
Many ways to infer low-dimensional structure from dissimilarities. Consider one other method: **Sammon Scaling**. Classic MDS minimizes **global** stress.
Many ways to infer low-dimensional structure from dissimilarities. Consider one other method: **Sammon Scaling**

Classic MDS minimizes *global* stress

\[
\text{Stress}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j}^{N} (d(\text{Doc}_j, \text{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2
\]
Many ways to infer low-dimensional structure from dissimilarities. Consider one other method: **Sammon Scaling**

Classic MDS minimizes **global** stress

\[
\text{Stress}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} \left( d(Doc_j, Doc_i) - d(x_j, x_i) \right)^2
\]

Often, we want a good approximation of **neighborhoods** (close to points), but don’t care about far away distances.
Many ways to infer low-dimensional structure from dissimilarities. Consider one other method: **Sammon Scaling**

Classic MDS minimizes **global** stress

\[
\text{Stress}(x) = \sum_{j=2}^{N} \sum_{i<j} (d(\text{Doc}_j, \text{Doc}_i) - d(x_j, x_i))^2
\]

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

**Sammon Scaling**
Sammon MDS

\[ \text{Algorithm "cares" more about small distances} \]

\[ \text{⇝ prioritizes approximations for small distances} \]

\[
\text{library}(\text{MASS}) \\
\text{sammon}
\]

Pro tip: For all document \( j \neq k \) \( d(\text{doc} j, \text{doc} k) > 0 \).
Sammon MDS

\[
\text{Stress}_{\text{Sammon}}(x) = \sum_{j=2}^{N} \sum_{i<j} \frac{(d(\text{Doc}_j, \text{Doc}_i) - d(x_j, x_i))^2}{d(\text{Doc}_j, \text{Doc}_i)}
\]
Sammon MDS

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

Algorithm “cares” more about small distances \(\rightsquigarrow\) prioritizes approximations for small distances
Sammon MDS

$$\text{Stress}_{\text{Sammon}}(x) = \sum_{j=2}^{N} \sum_{i<j} \frac{(d(Doc_j, Doc_i) - d(x_j, x_i))^2}{d(Doc_j, Doc_i)}$$

Algorithm “cares” more about small distances ⇾ prioritizes approximations for small distances

library(MASS)
Sammon MDS

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

Algorithm “cares” more about small distances $\iff$ prioritizes approximations for small distances
library(MASS)
sammon
Sammon MDS

$$\text{Stress}_{\text{Sammon}}(\mathbf{x}) = \sum_{j=2}^{N} \sum_{i<j} (d(\text{Doc}_j, \text{Doc}_i) - d(\mathbf{x}_j, \mathbf{x}_i))^2 \frac{1}{d(\text{Doc}_j, \text{Doc}_i)}$$

Algorithm “cares” more about small distances $\Rightarrow$ 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

![Graph comparing Sammon and Classic MDS](image-url)
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

**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~

quite useful

Peace Between Us

Analyzes K-substrings
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→ quite useful

Peace Between Us
Analyzes K-substrings
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~→ 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~→ quite useful

Peace Between Us

Analyzes K-substrings
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

quite useful
Peace Between Us
Analyzes K-substrings
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

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 → quite useful

Peace *Between Us*

*Analyzes K-substrings*
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

quite useful

Peace Between Us

Analyzes K-substrings
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, 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 ~⇒ quite useful

Peace Between Us

Analyzes K-substrings
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→ 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
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⇝
quite useful
Peace Between\textcolor{red}{n} Us
\textbf{Analyzes K-substrings}
Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously

- Kernel Trick (Linear Algebra, 101):
  \[ a = (a_1, a_2, ..., a_K) \]
  \[ b = (b_1, b_2, ..., b_K) \]
  \[ a \cdot b = a_1 \times b_1 + a_2 \times b_2 + ... + a_K \times b_K \]
- 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

- **Kernel Methods**: Represent texts, measure similarity

- Kernel Trick (Linear Algebra, 101): 

  \[ \mathbf{a} = (a_1, a_2, ..., a_K) \]

  \[ \mathbf{b} = (b_1, b_2, ..., b_K) \]

  \[ \mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + ... + a_K \times b_K \]

- 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

- **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):

\[ a = (a_1, a_2, \ldots, a_K) \quad b = (b_1, b_2, \ldots, b_K) \]

- 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

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

\[ a = (a_1, a_2, \ldots, a_K) \quad b = (b_1, b_2, \ldots, b_K) \]

\[ a \cdot b = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K \]
Kernel Trick

- **Kernel Methods**: Represent texts, measure similarity simultaneously
- **Kernel Trick (Linear Algebra, 101)**:
  \[
  \mathbf{a} = (a_1, a_2, \ldots, a_K) \quad \mathbf{b} = (b_1, b_2, \ldots, b_K)
  \]
  \[
  \mathbf{a} \cdot \mathbf{b} = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K
  \]

- 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):**
  \[ a = (a_1, a_2, \ldots, a_K) \quad b = (b_1, b_2, \ldots, b_K) \]
  \[ a \cdot b = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K \]

- 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):

\[
a = (a_1, a_2, \ldots, a_K) \quad b = (b_1, b_2, \ldots, b_K)
\]

\[
a \cdot b = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K
\]

- 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):

\[ a = (a_1, a_2, \ldots, a_K) \quad b = (b_1, b_2, \ldots, b_K) \]

\[ a \cdot b = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K \]

- 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):

\[
a = (a_1, a_2, \ldots, a_K) \quad b = (b_1, b_2, \ldots, b_K)
\]

\[
a \cdot b = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K
\]

- 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):

\[ a = (a_1, a_2, \ldots, a_K) \quad b = (b_1, b_2, \ldots, b_K) \]

\[ a \cdot b = a_1 \times b_1 + a_2 \times b_2 + \ldots + a_K \times b_K \]

- 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):

  \[
  \mathbf{a} = (a_1, a_2, \ldots, a_K) \quad \mathbf{b} = (b_1, b_2, \ldots, b_K)
  \]

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

- 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 $\rightarrow$ 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