Text as Data

Justin Grimmer

Professor
Department of Political Science
Stanford University

May 23rd, 2019

Discovery and Measurement

What is the research process? (Grimmer, Roberts, and Stewart 2019)

1) Discovery: a hypothesis or view of the world
2) Measurement according to some organization
3) Causal Inference: effect of some intervention

Text as data methods assist at each stage of research process

Text as Data Methods for Discovery

Text as Data Methods for Discovery Goal: Automatically Discover Organization (Similar Groups)

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 \\ 0 & 0 & 3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 3 \end{pmatrix}$$

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 \\ 0 & 0 & 3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 3 \end{pmatrix}$$

Suppose documents live in a space

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 \\ 0 & 0 & 3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 \\ 0 & 0 & 3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry $\rightsquigarrow$ modify with word weighting

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 \\ 0 & 0 & 3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry $\rightsquigarrow$ modify with word weighting
- Natural notions of distance

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \ldots & 0 \\ 0 & 0 & 3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 3 \end{pmatrix}$$

Suppose documents live in a space $\leadsto$ rich set of results from linear algebra

- Provides a geometry $\leadsto$ modify with word weighting
- Natural notions of distance
- Building block for clustering, supervised learning, and scaling

Texts in Space

Texts in Space

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

Texts in Space

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

Texts in Space

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

Texts in Space

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

Inner Product between documents:

Texts in Space

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

Inner Product between documents:

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

Texts in Space

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

Inner Product between documents:

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

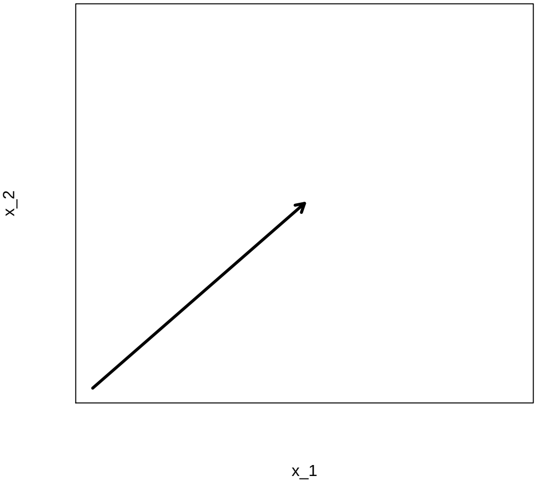
Texts in Space

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

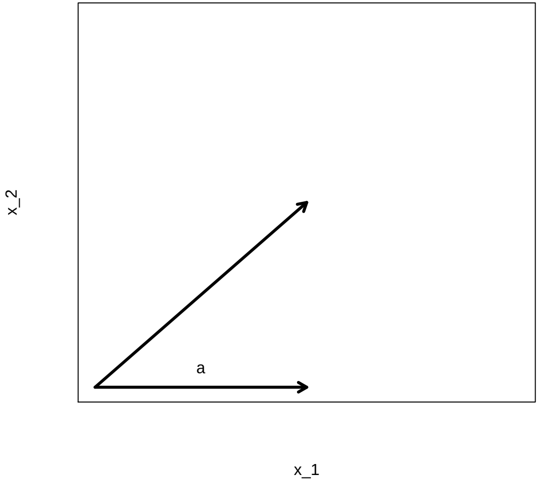
Inner Product between documents:

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

Vector Length

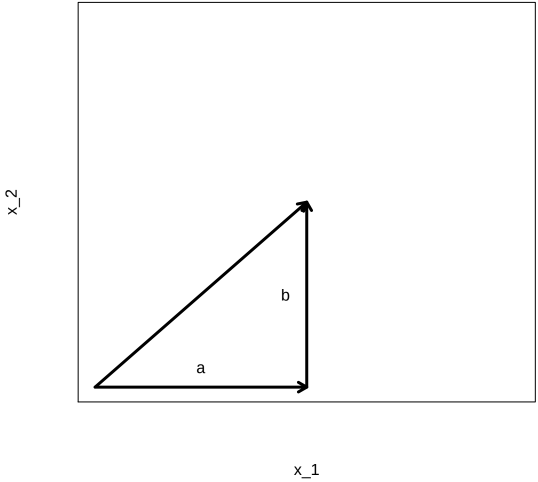


Vector Length



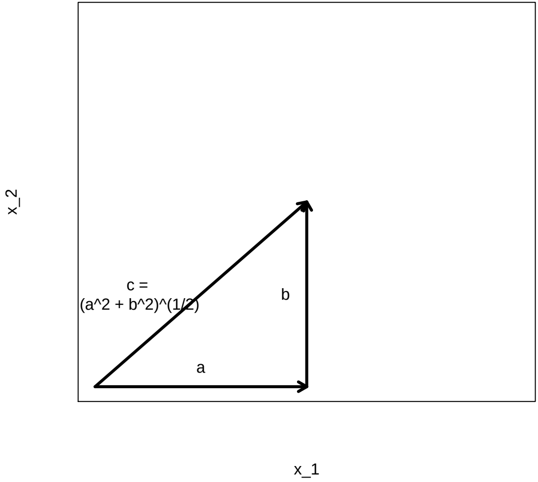
- Pythagorean Theorem: Side with length a

Vector Length



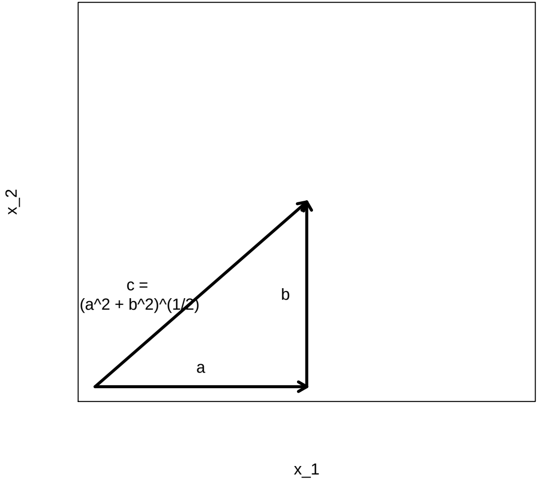
- Pythagorean Theorem: Side with length *a*
- Side with length *b* and right triangle

Vector Length



- Pythagorean Theorem: Side with length a
- Side with length b and right triangle
- $c = \sqrt{a^2 + b^2}$

Vector Length



- Pythagorean Theorem: Side with length a
- Side with length b and right triangle
- $c = \sqrt{a^2 + b^2}$
- This is generally true

Vector (Euclidean) Length

Definition

Suppose $\boldsymbol{v} \in \Re^J$. *Then, we will define its* *length* *as*

$$
\begin{aligned}
||\boldsymbol{v}|| &= (\boldsymbol{v} \cdot \boldsymbol{v})^{1/2} \\
&= (v_1^2 + v_2^2 + v_3^2 + \ldots + v_J^2)^{1/2}
\end{aligned}
$$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess $\rightsquigarrow$ Distance metrics
Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess $\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$
4) $d(\boldsymbol{X}_i, \boldsymbol{X}_k) \leq d(\boldsymbol{X}_i, \boldsymbol{X}_j) + d(\boldsymbol{X}_j, \boldsymbol{X}_k)$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\leadsto$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$
4) $d(\boldsymbol{X}_i, \boldsymbol{X}_k) \leq d(\boldsymbol{X}_i, \boldsymbol{X}_j) + d(\boldsymbol{X}_j, \boldsymbol{X}_k)$

Explore distance functions to compare documents$\leadsto$

Measures of Dissimilarity

Initial guess $\rightsquigarrow$ Distance metrics

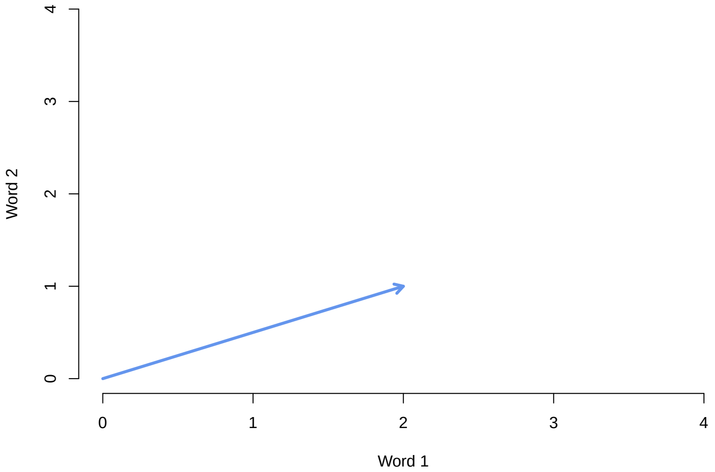
Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$
4) $d(\boldsymbol{X}_i, \boldsymbol{X}_k) \leq d(\boldsymbol{X}_i, \boldsymbol{X}_j) + d(\boldsymbol{X}_j, \boldsymbol{X}_k)$

Explore distance functions to compare documents $\rightsquigarrow$ Do we want additional assumptions/properties?

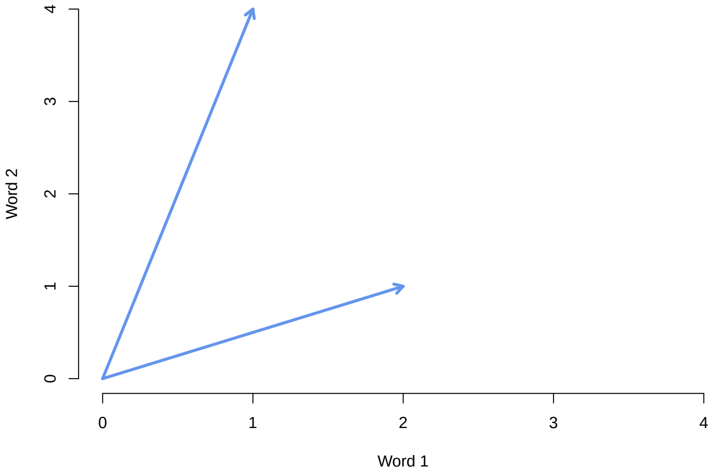
Measuring the Distance Between Documents

Euclidean Distance



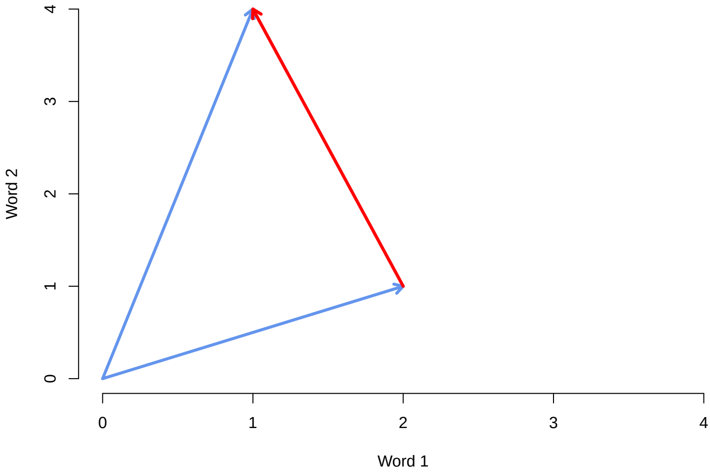
Measuring the Distance Between Documents

Euclidean Distance



Measuring the Distance Between Documents

Euclidean Distance



Measuring the Distance Between Documents

Definition

The Euclidean distance between documents $\boldsymbol{X}_i$ *and* $\boldsymbol{X}_j$ *as*

$$||\boldsymbol{X}_i - \boldsymbol{X}_j|| = \sqrt{\sum_{m=1}^{J} (x_{im} - x_{jm})^2}$$

Measuring the Distance Between Documents

Definition

The Euclidean distance between documents $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ as

$$||\boldsymbol{X}_i - \boldsymbol{X}_j|| = \sqrt{\sum_{m=1}^{J} (x_{im} - x_{jm})^2}$$

Suppose $\boldsymbol{X}_i = (1, 4)$ and $\boldsymbol{X}_j = (2, 1)$. The distance between the documents is:

$$\begin{aligned} ||(1,4) - (2,1)|| &= \sqrt{(1-2)^2 + (4-1)^2} \\ &= \sqrt{10} \end{aligned}$$

Measuring Similarity (and removing document length)

Measuring Similarity (and removing document length)

What properties should similarity measure have?

Measuring Similarity (and removing document length)

What properties should similarity measure have?

- Maximum: document with itself

Measuring Similarity (and removing document length)

What properties should similarity measure have?

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

Measuring Similarity (and removing document length)

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 (and removing document length)

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
- ? $s(a, b) = s(b, a)$.

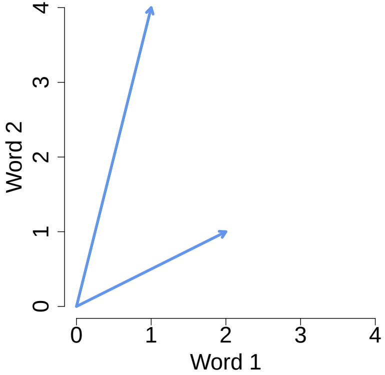
Measuring Similarity (and removing document length)

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
- ? $s(a, b) = s(b, a)$.

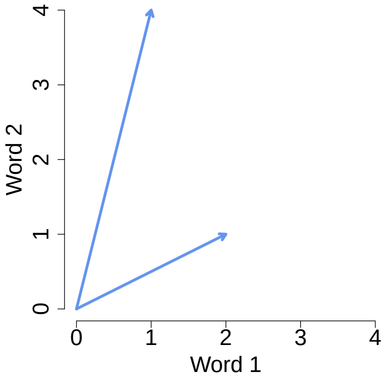
How should additional words be treated?

Measuring Similarity



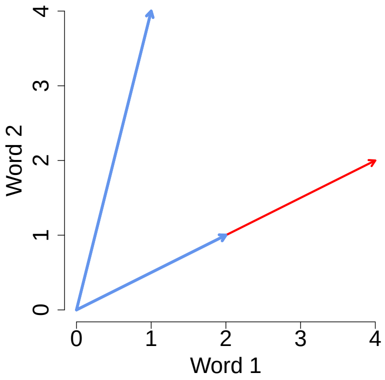
Measure 1: Inner product

Measuring Similarity

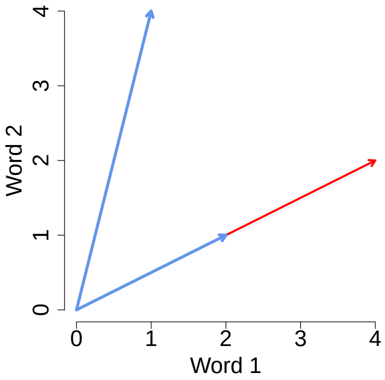


Measure 1: Inner product

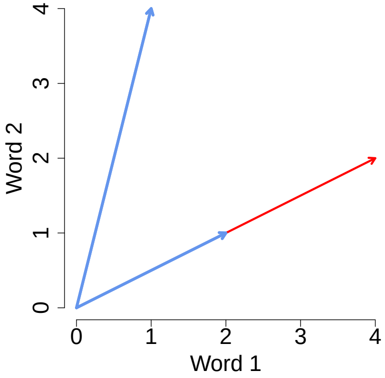
$$(2, 1)' \cdot (1, 4) \quad = \quad 6$$



Word 2
0
1
2
3
4
Word 1

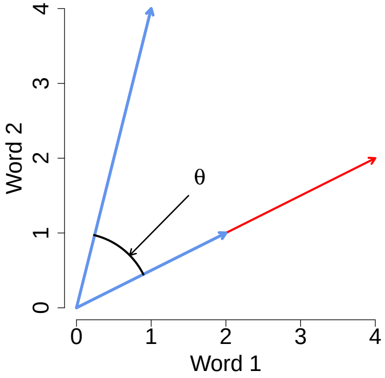


Problem(?): length dependent



Problem(?): length dependent

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



Problem(?): length dependent

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

Cosine Similarity

Cosine Similarity

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

Cosine Similarity

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

Cosine Similarity

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

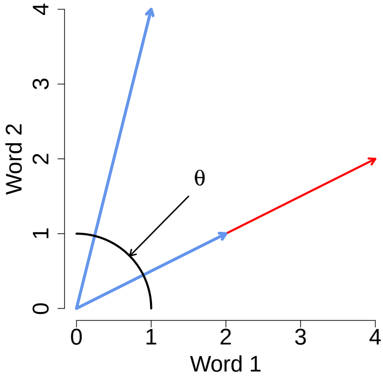
Cosine Similarity

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

Cosine Similarity

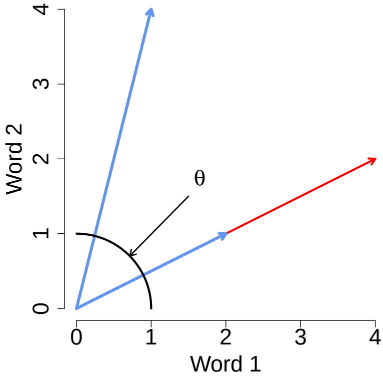
$$
\begin{aligned}
\cos\theta &= \left(\frac{a}{||a||}\right)\cdot\left(\frac{b}{||b||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45) \\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45) \\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97) \\
(0.89, 0.45)^{'}(0.24, 0.97) &= 0.65
\end{aligned}
$$

Cosine Similarity



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

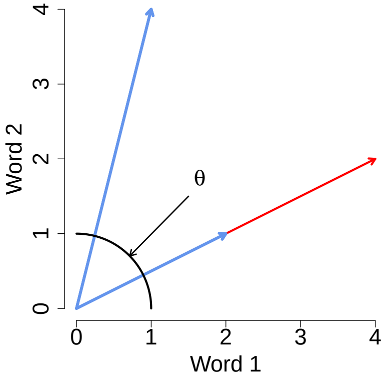
Cosine Similarity



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

Projects texts to unit length representation $\rightsquigarrow$ onto sphere

Cosine Similarity



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

Projects texts to unit length representation$\rightsquigarrow$ onto sphere

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:

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

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:

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

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

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} \quad = \quad (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \textbf{idf} \quad = \quad (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \textbf{idf} \quad = \quad (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \textbf{idf} = (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \textbf{idf} = (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \textbf{idf} = (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \textbf{idf} = (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

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

Weighting Words: TF-IDF

$$
\begin{aligned}
\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} &= (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J) \\
\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \mathbf{idf} &= (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)
\end{aligned}
$$

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

$$
\mathbf{X}_{i,\text{idf}} \cdot \mathbf{X}_{j,\text{idf}} = (\mathbf{X}_i \times \mathbf{idf})'(\mathbf{X}_j \times \mathbf{idf})
$$

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} = (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \mathbf{idf} = (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Inner Product

$$\begin{aligned} \mathbf{X}_{i,\text{idf}} \cdot \mathbf{X}_{j,\text{idf}} &= (\mathbf{X}_i \times \mathbf{idf})^{\prime}(\mathbf{X}_j \times \mathbf{idf}) \\ &= (\text{idf}_1^2 \times X_{i1} \times X_{j1}) + (\text{idf}_2^2 \times X_{i2} \times X_{j2}) + \\ &\quad \ldots + (\text{idf}_J^2 \times X_{iJ} \times X_{jJ}) \end{aligned}$$

Weighting Words: Inner Product

Define:

Weighting Words: Inner Product

Define:

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

Weighting Words: Inner Product

Define:

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

If we use tf-idf for our documents, then

Weighting Words: Inner Product

Define:

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

If we use tf-idf for our documents, then

$$\begin{aligned} d_2(\boldsymbol{X}_i, \boldsymbol{X}_j) &= \sqrt{\sum_{m=1}^{J} (x_{im,\text{idf}} - x_{jm,\text{idf}})^2} \\ &= \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_j)' \boldsymbol{\Sigma} (\boldsymbol{X}_i - \boldsymbol{X}_j)} \end{aligned}$$

Final Product

Applying some measure of distance, similarity (if symmetric) yields:

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

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

Clustering

Fully Automated Clustering

1) Distance metric $\rightsquigarrow$ when are documents close?
2) Objective function $\rightsquigarrow$ how do we summarize distances?
3) Optimization method $\rightsquigarrow$ how do we find optimal clustering?

THERE IS NO A PRIORI OPTIMAL METHOD

Computer Assisted Clustering (Grimmer and King, 2011)

- crucial to combine human and computer insights

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

K-Means$\leadsto$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\leadsto$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k \quad = \quad (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

K-Means$\leadsto$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\leadsto$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

$\boldsymbol{\theta}_k$ = exemplar for cluster k

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k$ = exemplar for cluster k

2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k =$ exemplar for cluster k

2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.
   If observation i is from cluster k, then

K-Means$\leadsto$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\leadsto$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k =$ exemplar for cluster k

2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.
   If observation i is from cluster k, then

$$\boldsymbol{\tau}_i = (0, 0, \ldots, 0, \underbrace{1}_{k^{th}}, 0, \ldots, 0)$$

K-Means$\rightsquigarrow$ Objective Function

N documents $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ (normalized)
Goal$\rightsquigarrow$ Partition documents into K clusters.
Two parameters to estimate

1) $K \times J$ matrix of cluster centers Θ.
   Cluster k has center

$$\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{Jk})$$

   $\boldsymbol{\theta}_k$ = exemplar for cluster k

2) $\boldsymbol{T}$ is an $N \times K$ matrix. Each row is an indicator vector.
   If observation i is from cluster k, then

$$\boldsymbol{\tau}_i = (0, 0, \ldots, 0, \underbrace{1}_{k^{th}}, 0, \ldots, 0)$$

Hard Assignment

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

K-Means $\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
    - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)

K-Means$\leadsto$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J}(x_{ij} - \theta_{kj})^2\right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$
  - If $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sigma^2$

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$
  - If $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sigma^2$
    - Each observation in same cluster

K-Means$\rightsquigarrow$ Objective Function

Assume squared euclidean distance

$$f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \overbrace{\tau_{ik}}^{\text{cluster indicator}} \underbrace{\left(\sum_{j=1}^{J} (x_{ij} - \theta_{kj})^2 \right)}_{\text{Squared Euclidean Distance}}$$

- Calculate squared euclidean distance from center
- Only for the assigned cluster
- Two trivial solutions
  - If $K = N$ then $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = 0$ (Minimum)
    - Each observation in its own cluster
    - $\boldsymbol{\theta}_i = \boldsymbol{x}_i$
  - If $K = 1$, $f(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\Theta}) = N \times \sigma^2$
    - Each observation in same cluster
    - $\boldsymbol{\theta}_1 =$ Average across documents

K-Means $\rightsquigarrow$ Optimization

Coordinate descent

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

K-Means $\rightsquigarrow$ Optimization

Coordinate descent $\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

K-Means$\rightsquigarrow$ Optimization

Coordinate descent$\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^t$

K-Means $\rightsquigarrow$ Optimization

Coordinate descent $\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^{t}$
- Conditional on $\boldsymbol{T}^{t}$, choose $\boldsymbol{\Theta}^{t}$

K-Means $\rightsquigarrow$ Optimization

Coordinate descent $\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^t$
- Conditional on $\boldsymbol{T}^t$, choose $\boldsymbol{\Theta}^t$

Repeat until convergence $\rightsquigarrow$ as measured as change in f dropping below threshold ϵ

K-Means $\rightsquigarrow$ Optimization

Coordinate descent $\rightsquigarrow$ iterate between labels and centers.

Iterative algorithm: each iteration t

- Conditional on $\boldsymbol{\Theta}^{t-1}$ (from previous iteration), choose $\boldsymbol{T}^t$
- Conditional on $\boldsymbol{T}^t$, choose $\boldsymbol{\Theta}^t$

Repeat until convergence $\rightsquigarrow$ as measured as change in f dropping below threshold ϵ

$$\text{Change} = f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta}^t) - f(\boldsymbol{X}, \boldsymbol{T}^{t-1}, \boldsymbol{\Theta}^{t-1})$$

K-Means $\rightsquigarrow$ Optimization

K-Means$\rightsquigarrow$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

K-Means$\rightsquigarrow$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

2) Choose $\boldsymbol{T}^t$

K-Means$\rightsquigarrow$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

2) Choose $\boldsymbol{T}^t$

$$\tau_{im}^t = \begin{cases} 1 \text{ if } m = \arg\min_k \sum_{j=1}^{J} (x_{ij} - \theta_{kj}^t)^2 \\ 0 \text{ otherwise }, \end{cases} .$$

K-Means$\rightsquigarrow$ Optimization

1) initialize K cluster centers $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \ldots, \boldsymbol{\theta}_K^t$.

2) Choose $\boldsymbol{T}^t$

$$\tau_{im}^t = \begin{cases} 1 \text{ if } m = \arg\min_k \sum_{j=1}^J (x_{ij} - \theta_{kj}^t)^2 \\ 0 \text{ otherwise }, \end{cases} .$$

In words: Assign each document $\boldsymbol{x}_i$ to the closest center $\boldsymbol{\theta}_m^t$

K-Means$\rightsquigarrow$ Optimization

K-Means$\rightsquigarrow$ Optimization

3) Choose $\Theta^t \rightsquigarrow$ Focus on the center for cluster k

K-Means $\rightsquigarrow$ Optimization

3) Choose Θ^t $\rightsquigarrow$ Focus on the center for cluster k

$$f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k \quad = \quad \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right)$$

K-Means$\rightsquigarrow$ Optimization

3) Choose $\Theta^t \rightsquigarrow$ Focus on the center for cluster k

$$\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk} \right)
\end{aligned}$$

K-Means $\rightsquigarrow$ Optimization

3) Choose Θ^t $\rightsquigarrow$ Focus on the center for cluster k

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^{t} \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^{t} \left(x_{ij} - \theta_{jk} \right) \\
0 &= -2 \sum_{i=1}^{N} \tau_{ij}^{t} \left(x_{ij} - \theta_{jk}^{*} \right)
\end{aligned}
$$

K-Means$\rightsquigarrow$ Optimization

3) Choose $\Theta^t \rightsquigarrow$ Focus on the center for cluster k

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk} \right) \\
0 &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk}^* \right) \\
&= \sum_{i=1}^{N} \tau_{ij}^t x_{ij} - \theta_{jk}^* \sum_{i=1}^{N} \tau_{ij}^t
\end{aligned}
$$

K-Means$\rightsquigarrow$ Optimization

3) Choose $\Theta^t \rightsquigarrow$ Focus on the center for cluster k

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k &= \sum_{i=1}^{N} \tau_{ik}^t \left(\sum_{j=1}^{J} (x_{ij} - \theta_{jk})^2 \right) \\
\frac{\partial f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta})_k}{\partial \theta_{kj}} &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk} \right) \\
0 &= -2 \sum_{i=1}^{N} \tau_{ij}^t \left(x_{ij} - \theta_{jk}^* \right) \\
&= \sum_{i=1}^{N} \tau_{ij}^t x_{ij} - \theta_{jk}^* \sum_{i=1}^{N} \tau_{ij}^t \\
\frac{\sum_{i=1}^{N} \tau_{ik}^t x_{ij}}{\sum_{i=1}^{N} \tau_{ik}^t} &= \theta_{jk}^*
\end{aligned}
$$

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}}$$

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

K-Means $\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} \;=\; \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.

K-Means$\leadsto$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

K-Means$\leadsto$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} = \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers
- Do until converged:

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers
- Do until converged:
  - For each document, find closest center$\rightsquigarrow$ $\boldsymbol{\tau}_i^t$

K-Means$\rightsquigarrow$ Optimization

$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

- Initialize centers
- Do until converged:
  - For each document, find closest center$\rightsquigarrow$ $\boldsymbol{\tau}_i^t$
  - For each center, take average of assigned documents$\rightsquigarrow$ $\boldsymbol{\theta}_k^t$

K-Means$\rightsquigarrow$ Optimization

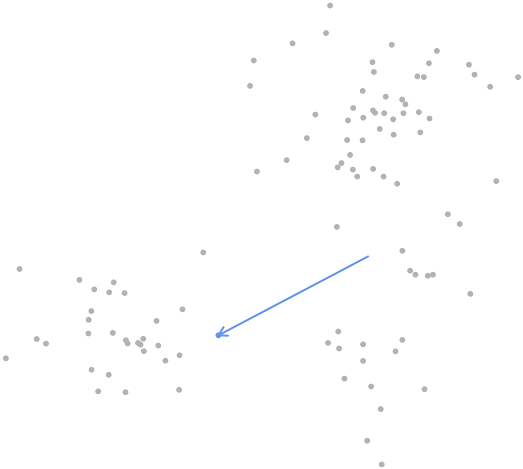
$$\boldsymbol{\theta}^{t+1} \quad = \quad \frac{\sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \tau_{ik}} \propto \sum_{i=1}^{N} \tau_{ik} \boldsymbol{x}_i$$

In words: $\boldsymbol{\theta}^{t+1}$ is the average of the documents assigned to k.
Optimization algorithm:

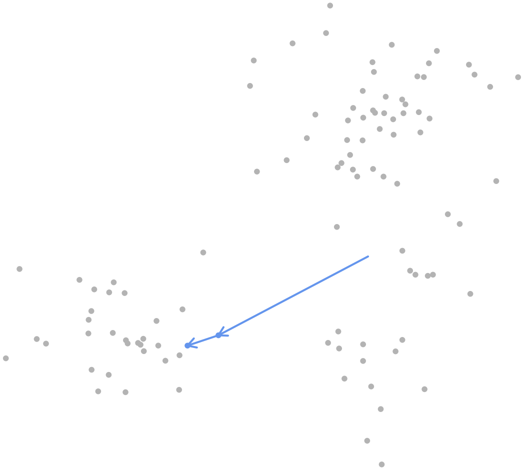
- Initialize centers
- Do until converged:
  - For each document, find closest center$\rightsquigarrow$ $\boldsymbol{\tau}_i^t$
  - For each center, take average of assigned documents$\rightsquigarrow$ $\boldsymbol{\theta}_k^t$
  - Update change $f(\boldsymbol{X}, \boldsymbol{T}^t, \boldsymbol{\Theta}^t) - f(\boldsymbol{X}, \boldsymbol{T}^{t-1}, \boldsymbol{\Theta}^{t-1})$

Visual Example

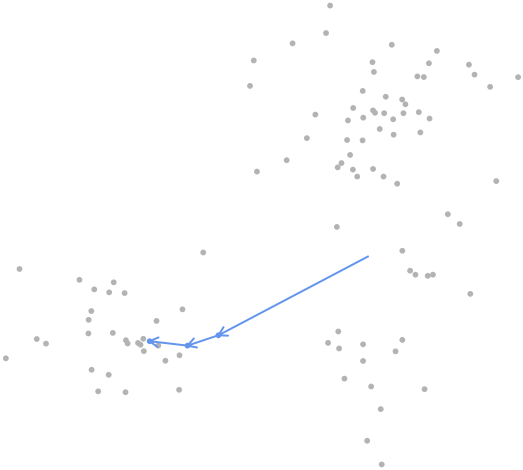
Visual Example



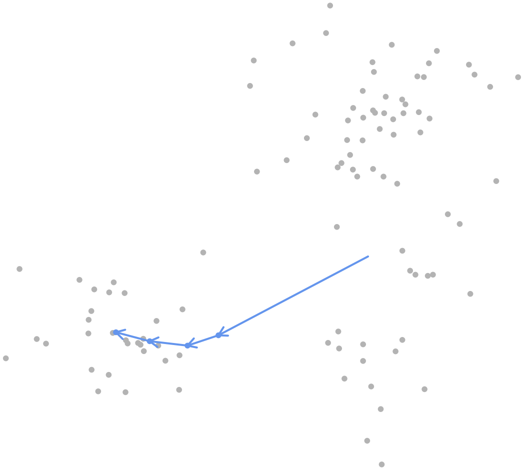
Visual Example



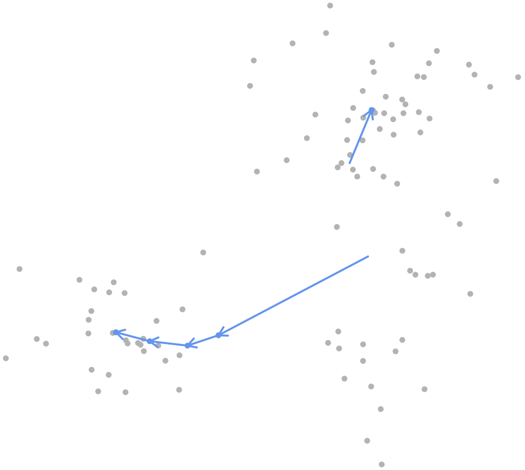
Visual Example



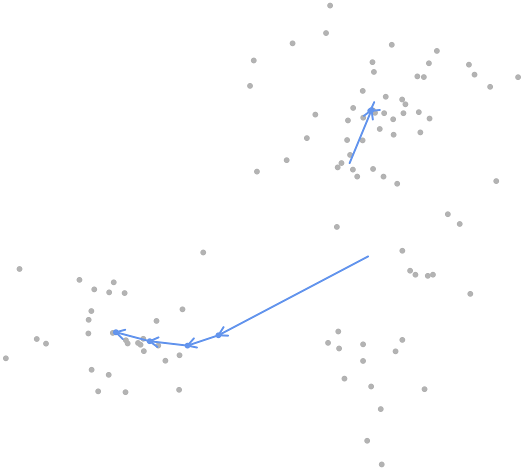
Visual Example



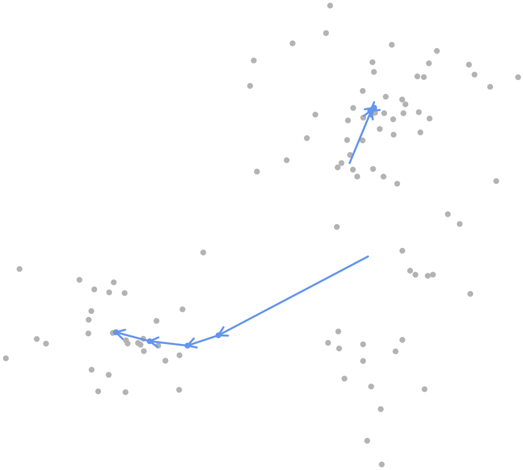
Visual Example



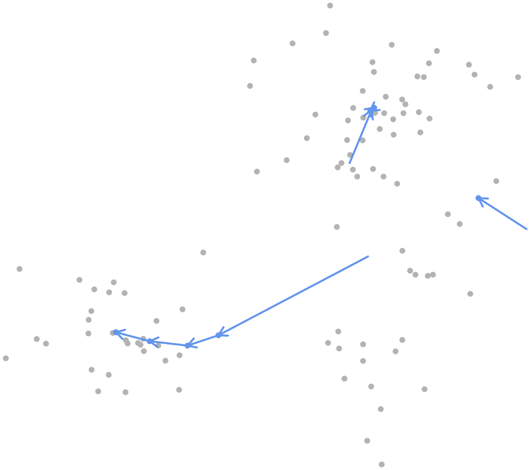
Visual Example



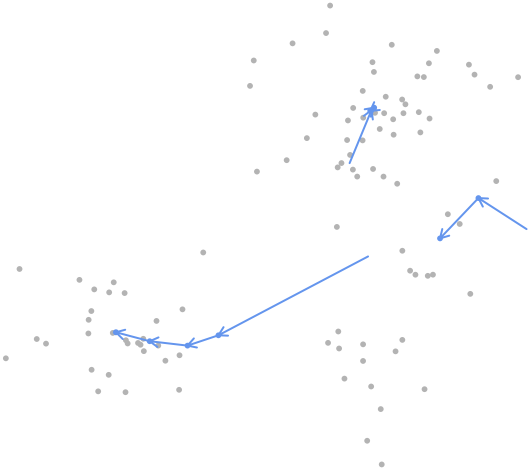
Visual Example



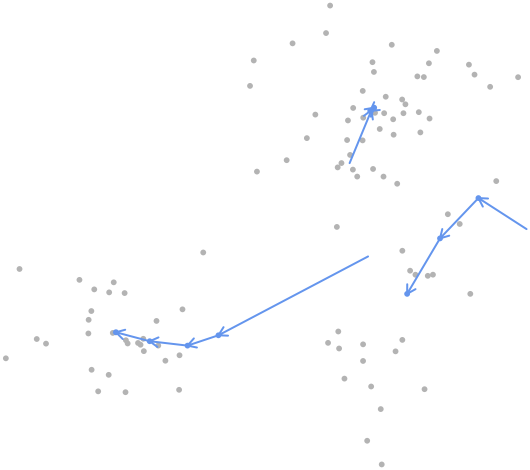
Visual Example



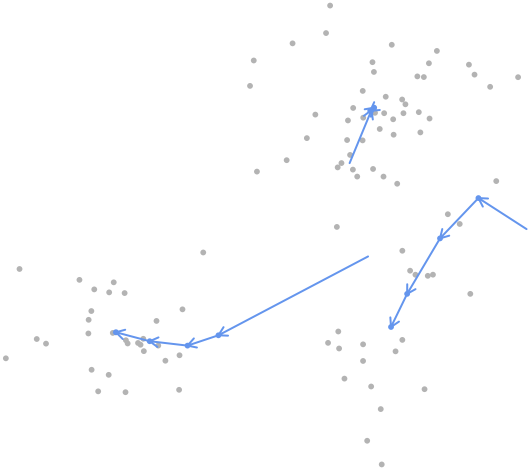
Visual Example



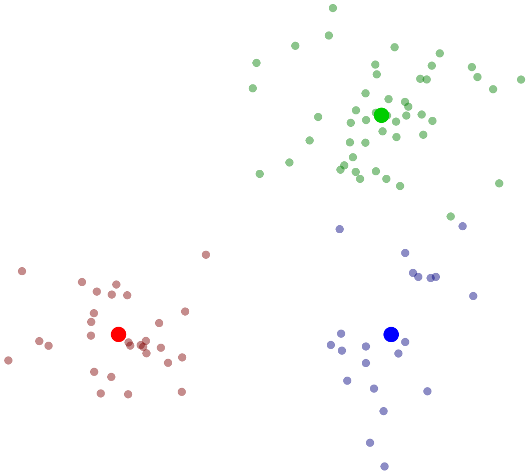
Visual Example



Visual Example



Visual Example



An Example: Jeff Flake

`To the R Code!`

Interpreting Cluster Components

Unsupervised methods

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster
        - Read documents
        - Assign cluster label
    - Automatic identification

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
    - Manual identification (Quinn et al 2010)
        - Sample set of documents from same cluster
        - Read documents
        - Assign cluster label
    - Automatic identification
        - Know label classes

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean

Interpreting Cluster Components

Unsupervised methods $\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean
  - Provide documents + organizations

Interpreting Cluster Components

Unsupervised methods$\rightsquigarrow$ low startup costs, high post-model costs

- Apply clustering methods, we have groups of documents
- How to interpret the groups?
- Two (broad) methods:
  - Manual identification (Quinn et al 2010)
    - Sample set of documents from same cluster
    - Read documents
    - Assign cluster label
  - Automatic identification
    - Know label classes
    - Use methods to identify separating words
    - Use these to help infer differences across clusters
- Transparency
  - Debate what clusters are
  - Debate what they mean
  - Provide documents + organizations

`back to the R code!`

How Do We Choose K?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistic + manual search

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistic + manual search$\rightsquigarrow$discuss statistical methods/experimental methods on Thursday
- Humans should be the final judge

How Do We Choose K?

- Previous Analysis Assumed We Know Number of Clusters
- How Do We Choose Cluster Number?
- Cannot Compare f across clusters
  - Sum squared errors decreases as K increases
  - Trivial answer: each document in own cluster (useless)
  - Modeling problem: Fit often increases with features
- How do we choose number of clusters?

Think!

- No one statistic captures how you want to use your data
- But, can help guide your selection
- Combination statistic + manual search $\rightsquigarrow$ discuss statistical methods/experimental methods on Thursday
- Humans should be the final judge
  - Compare insights across clusterings

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications
Single distribution data generating process:

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\boldsymbol{\tau}_i|\boldsymbol{\pi} \sim \text{Multinomial}(1, \boldsymbol{\pi})$$

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models $\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{\tau}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | \tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{\tau}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i|\tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

In words:

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned}\boldsymbol{\tau}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i|\tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k)\end{aligned}$$

In words:

- Draw a cluster label

Mixture of Unigram Models (Mixture of Multinomials)

Mixture models$\rightsquigarrow$ wide range of applications
Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{\tau}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i|\tau_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

In words:

- Draw a cluster label
- Given distribution, draw realization

Mixture of Unigram Models (Mixture of Multinomials)

A mixture of unigram-language models

$$
\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\tau}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\
\boldsymbol{x}_i | \tau_{ik} = 1, \boldsymbol{\theta}_k &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_k)
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$p(\boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\pi} | \boldsymbol{X})$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) \propto \overbrace{p(\pi)p(\theta)}^{1} \underbrace{p(\boldsymbol{X}, \boldsymbol{T} | \pi, \theta)}_{\text{Complete data likelihood}}$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) & \propto \overbrace{p(\boldsymbol{\pi}) p(\boldsymbol{\theta})}^{1} \quad \underbrace{p(\boldsymbol{X}, \boldsymbol{T} | \pi, \boldsymbol{\theta})}_{\text{Complete data likelihood}} \\
& \propto \underbrace{\prod_{i=1}^{N} p(\boldsymbol{\tau}_i, \boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi})}_{\text{Complete data likelihood}}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) &\propto \overbrace{p(\boldsymbol{\pi}) p(\boldsymbol{\theta})}^{1} \underbrace{p(\boldsymbol{X}, \boldsymbol{T} | \pi, \boldsymbol{\theta})}_{\text{Complete data likelihood}} \\
&\propto \underbrace{\prod_{i=1}^{N} p(\boldsymbol{\tau}_i, \boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi})}_{\text{Complete data likelihood}} \\
&\propto \prod_{i=1}^{N} p(\boldsymbol{\tau}_i | \boldsymbol{\pi}) p(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\tau}_i)
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

This implies the following posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{T}, \boldsymbol{\Theta}, \pi | \boldsymbol{X}) &\propto \overbrace{p(\boldsymbol{\pi})p(\boldsymbol{\theta})}^{1} \quad \underbrace{p(\boldsymbol{X}, \boldsymbol{T}|\pi, \boldsymbol{\theta})}_{\text{Complete data likelihood}} \\
&\propto \underbrace{\prod_{i=1}^{N} p(\boldsymbol{\tau}_i, \boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi})}_{\text{Complete data likelihood}} \\
&\propto \prod_{i=1}^{N} p(\boldsymbol{\tau}_i|\boldsymbol{\pi})p(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{\tau}_i) \\
&\propto \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}}\right]^{\tau_{ik}}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates$\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N} \sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N} \sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i | \boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\mathrm{E}[\log \text{Complete data} | \boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N} \sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t | \boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t | \boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

$$\begin{aligned} \text{Change} &= \mathrm{E}[\log \text{Complete data} | \boldsymbol{\Theta}^{t+1}, \boldsymbol{\pi}^{t+1}] \\ &\quad - \mathrm{E}[\log \text{Complete data} | \boldsymbol{\Theta}^{t}, \boldsymbol{\pi}^{t}] \end{aligned}$$

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k) p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

$$\begin{aligned} \text{Change} &= \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t+1}, \boldsymbol{\pi}^{t+1}] \\ &\quad - \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t}, \boldsymbol{\pi}^{t}] \end{aligned}$$

Our update steps will be strikingly similar to the K-Means algorithm

Mixture of Unigram Models (Mixture of Multinomials)

Obtain MAP estimates $\rightsquigarrow$ EM Algorithm

1) Initialize parameters $\boldsymbol{\Theta}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute $p(\boldsymbol{\tau}_i|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$
3) Maximization step: maximize with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}_k, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} \log p(\boldsymbol{x}_i, \tau_{ik}^t|\boldsymbol{\theta}_k, \pi_k)p(\tau_{ik}^t|\boldsymbol{\Theta}, \pi_k)$$

Obtain $\boldsymbol{\Theta}^{t+1}$, $\boldsymbol{\pi}^{t+1}$

4) Assess change

$$\begin{aligned}\text{Change} &= \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t+1}, \boldsymbol{\pi}^{t+1}] \\ &\quad - \text{E}[\log \text{Complete data}|\boldsymbol{\Theta}^{t}, \boldsymbol{\pi}^{t}]\end{aligned}$$

Our update steps will be strikingly similar to the K-Means algorithm

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X})$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \quad = \quad \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^{K}\left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}}$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^{K}\left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^{J}(\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^{K}\left(\pi_m^t \prod_{j=1}^{J}(\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^K \left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^J (\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^K \left(\pi_m^t \prod_{j=1}^J (\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Define:

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^{K}\left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^{J}(\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^{K}\left(\pi_m^t \prod_{j=1}^{J}(\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Define:

$$
r_{ik}^t \equiv \frac{\pi_k^t \prod_{j=1}^{J}(\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^{K}\left(\pi_m^t \prod_{j=1}^{J}(\theta_{jm}^t)^{x_{ij}}\right)}
$$

Mixture of Unigram Models (Mixture of Multinomials)

1) Initialize parameters $\boldsymbol{\Theta}^t$ and $\boldsymbol{\pi}^t$
2) E-Step

$$
\begin{aligned}
p(\tau_{ik}|\boldsymbol{\Theta}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) &= \overbrace{\frac{p(\tau_{ik}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_k^t)}{\sum_{m=1}^{K}\left(p(\tau_{im}|\boldsymbol{\pi}^t)p(\boldsymbol{x}_i|\boldsymbol{\theta}_m^t)\right)}}^{\text{general form}} \\
&= \frac{\pi_k^t \prod_{j=1}^{J}(\theta_{jk}^t)^{x_{ij}}}{\sum_{m=1}^{K}\left(\pi_m^t \prod_{j=1}^{J}(\theta_{jm}^t)^{x_{ij}}\right)}
\end{aligned}
$$

Define: Avoid underflow

$$
r_{ik}^t = \left[1 + \sum_{k' \neq k} \frac{\pi_{k'} \prod_{j=1}^{J}(\theta_{jk'}^t)^{x_{ij}}}{\pi_k \prod_{j=1}^{J}(\theta_{jk}^t)^{x_{ij}}}\right]^{-1}
$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right)$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N} \sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N}\sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E[log Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N}\sum_{k=1}^{K} E[\tau_{ik}] \log\left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

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

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N} \sum_{k=1}^{K} E[\tau_{ik}] \log \left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}} \right) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

$$
\begin{aligned}
\pi_k^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t}}{N} \\
\theta_{jk}^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t} x_{ij}}{\sum_{m=1}^{J} \sum_{i=1}^{N} r_{ik}^{t} x_{im}}
\end{aligned}
$$

Mixture of Unigram Models (Mixture of Multinomials)

3) M-Step:

$$
\begin{aligned}
\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] &= \sum_{i=1}^{N}\sum_{k=1}^{K} E[\tau_{ik}] \log\left(\pi_k \prod_{j=1}^{J} \theta_{jk}^{x_{ik}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^{t} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{J} r_{ik}^{t} x_{ij} \log \theta_{jk}
\end{aligned}
$$

Introducing constraints, differentiating, setting equal to zero and algebra yields:

$$
\begin{aligned}
\pi_k^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t}}{N} \\
\theta_{jk}^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^{t} x_{ij}}{\sum_{m=1}^{J}\sum_{i=1}^{N} r_{ik}^{t} x_{im}} \propto \sum_{i=1}^{N} r_{ik}^{t} \boldsymbol{x}_i
\end{aligned}
$$

Example: Jeff Flake Again!

```
To the R Code!
```

Fully Automated Clustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:

Fully Automated Clustering

- Notion of similarity and “good” partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering

Fully Automated Clustering

- Notion of similarity and “good” partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents

Fully Automated Clustering

- Notion of similarity and "good" partition$\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality$\rightsquigarrow$ Thursday

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

YOU DON'T!

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

YOU DON'T! $\rightsquigarrow$ **And never will**

Fully Automated Clustering

- Notion of similarity and “good” partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the “right” model?

YOU DON’T! $\rightsquigarrow$ And never will $\rightsquigarrow$ but still useful(!!!!)

Fully Automated Clustering

- Notion of similarity and "good" partition $\rightsquigarrow$ clustering
- Many clustering methods:
  - Spectral clustering
  - Affinity Propagation
  - Non-parametric statistical models
  - Hierarchical clustering
  - Biclustering
  - ...
- How do we know we have something useful?
  - Validation: read the documents
  - Validation: experiments to assess cluster quality $\rightsquigarrow$ Thursday
  - Validation: model based fit statistics
- How do we know we have the "right" model?

YOU DON'T! $\rightsquigarrow$ **And never will** $\rightsquigarrow$ **but still useful(!!!!)**

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1

Doc 2

Doc 3

$\vdots$

Doc N

Cluster 1

Cluster 2

$\vdots$

Cluster K

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1	
	Cluster 1
Doc 2 $\rightarrow$ Cluster 1	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Topic and Mixed Membership Models

Clustering
Document $\rightsquigarrow$ One Cluster

Doc 1
Doc 2
Doc 3
$\vdots$
Doc N

Cluster 1
Cluster 2
$\vdots$
Cluster K

Topic and Mixed Membership Models

Clustering

Document $\rightsquigarrow$ One Cluster

Doc 1	
	Cluster 1
Doc 2	
	Cluster 2
Doc 3	
	$\vdots$
$\vdots$	
	Cluster K
Doc N	

Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters

Doc 1

Cluster 1

Doc 2

Cluster 2

Doc 3

$\vdots$

$\vdots$

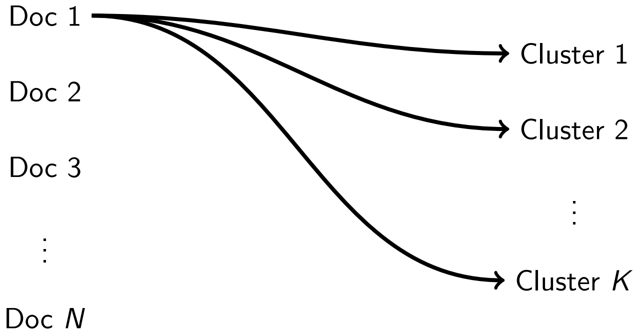
Cluster K

Doc N

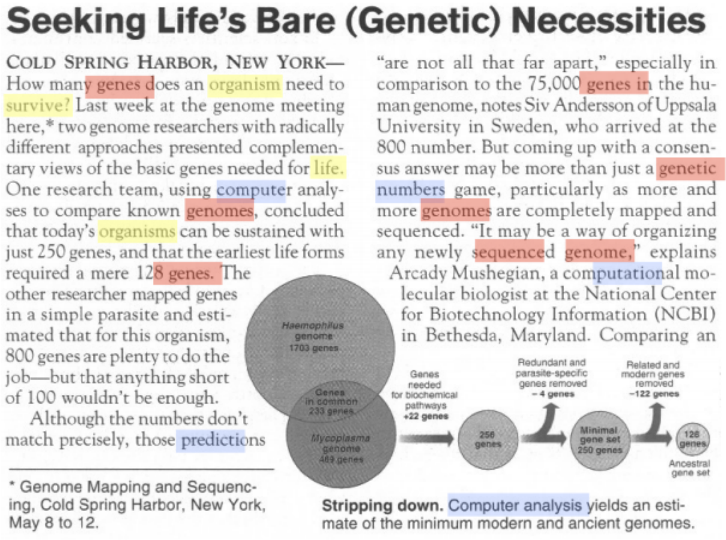
Topic and Mixed Membership Models

Topic Models (Mixed Membership)

Document $\rightsquigarrow$ Many clusters



A Statistical Highlighter (With Many Colors)



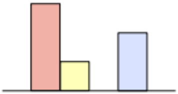
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



AdMixture: A General Approach to Modeling

AdMixture: A General Approach to Modeling

Latent Dirichlet Allocation: an admixture model

AdMixture: A General Approach to Modeling

Latent Dirichlet Allocation: an admixture model

Intuition: each unit is a mixture of latent components

AdMixture: A General Approach to Modeling

Latent Dirichlet Allocation: an admixture model

Intuition: each unit is a mixture of latent components

Mixture Model

$$
\begin{aligned}
\boldsymbol{\pi} & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\tau_i & \sim \text{Multinomial}(1, \boldsymbol{\pi}) \\
\boldsymbol{x}_i | \tau_{ik} = 1 & \sim \text{Distribution}(\boldsymbol{\theta}_k)
\end{aligned}
$$

AdMixture: A General Approach to Modeling

Latent Dirichlet Allocation: an admixture model
Intuition: each unit is a mixture of latent components
Mixture Model

$$\begin{aligned} \boldsymbol{\pi} & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \tau_i & \sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | \tau_{ik} = 1 & \sim \text{Distribution}(\boldsymbol{\theta}_k) \end{aligned}$$

AdMixture Model

$$\begin{aligned} \boldsymbol{\pi}_i & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \tau_{im} & \sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_{im} | \tau_{imk} = 1 & \sim \text{Distribution}(\boldsymbol{\theta}_k) \end{aligned}$$

AdMixture: A General Approach to Modeling

Latent Dirichlet Allocation: an admixture model
Intuition: each unit is a mixture of latent components
Mixture Model

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \tau_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | \tau_{ik} = 1 &\sim \text{Distribution}(\boldsymbol{\theta}_k) \end{aligned}$$

AdMixture Model

$$\begin{aligned} \boldsymbol{\pi}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \tau_{im} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_{im} | \tau_{imk} = 1 &\sim \text{Distribution}(\boldsymbol{\theta}_k) \end{aligned}$$

Vanilla Latent Dirichlet Allocation $\leadsto$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

*Notice: this is a different representation than a document-term matrix. x_{im} is a number that says which of the J words are used. The difference is for clarity and we'll this representation is closely related to document-term matrix

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$\boldsymbol{\pi}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\pi}_i | \boldsymbol{\alpha} & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i & \sim \text{Multinomial}(1, \boldsymbol{\pi}_i)
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, ($i = 1, 2, \ldots, N$).
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$\boldsymbol{\theta}_k \sim \text{Dirichlet}(\mathbf{1})$$

$$\begin{aligned}
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

- Consider document i, $(i = 1, 2, \ldots, N)$.
- Suppose there are M_i total words and $\boldsymbol{x}_i$ is an $M_i \times 1$ vector, where x_{im} describes the m^{th} word used in the document*.

$$
\begin{aligned}
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\pi}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{\tau}_{im} | \boldsymbol{\pi}_i &\sim \text{Multinomial}(1, \boldsymbol{\pi}_i) \\
x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\theta}_k)
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\alpha)p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\theta, \boldsymbol{T})$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi}|\alpha) p(\boldsymbol{T}|\boldsymbol{\pi}) p(\boldsymbol{X}|\theta, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi}|\alpha) p(\boldsymbol{T}|\boldsymbol{\pi}) p(\boldsymbol{X}|\theta, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) & \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T}) \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i | \boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im} | \boldsymbol{\pi}) p(x_{im} | \boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{T}|\boldsymbol{\pi}) p(\boldsymbol{X}|\theta, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) &\propto p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{T}|\boldsymbol{\pi}) p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation $\rightsquigarrow$ Find "closest" distribution

Vanilla Latent Dirichlet Allocation $\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha}|\boldsymbol{X}) &\propto p(\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{T}|\boldsymbol{\pi})p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation $\rightsquigarrow$ Find "closest" distribution
- Gibbs sampling $\rightsquigarrow$ MCMC algorithm to approximate posterior

Vanilla Latent Dirichlet Allocation$\rightsquigarrow$ Objective Function

Together the model implies the following posterior:

$$
\begin{aligned}
p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) & \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{T}|\boldsymbol{\pi}) p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[p(\boldsymbol{\pi}_i|\boldsymbol{\alpha}) \prod_{m=1}^{M_i} p(\boldsymbol{\tau}_{im}|\boldsymbol{\pi}) p(x_{im}|\boldsymbol{\theta}_k, \tau_{imk} = 1) \right] \\
& \propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{ik}^{\alpha_k - 1} \prod_{m=1}^{M} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{j=1}^{J} \theta_{jk}^{x_{imj}} \right]^{\tau_{ikm}} \right]
\end{aligned}
$$

Optimization:

- Variational Approximation$\rightsquigarrow$ Find "closest" distribution
- Gibbs sampling $\rightsquigarrow$ MCMC algorithm to approximate posterior

Described in the slides appendix

Running a Topic Model with STM

`to the STM Code`

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word$_1$	Word$_2$	$\ldots$	Word$_J$
Doc$_1$	0	1	$\ldots$	0
Doc$_2$	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc$_N$	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?
Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$
Allows: measure of correlation of term usage across documents
(heuristically: partition words, based on usage in documents)

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$

Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)

Latent Semantic Analysis: Reduce information in matrix using linear algebra (provides similar results, difficult to generalize)

Why does this work $\rightsquigarrow$ Co-occurrence

Where's the information for each word's topic?

Reconsider document-term matrix

	Word_1	Word_2	$\ldots$	Word_J
Doc_1	0	1	$\ldots$	0
Doc_2	2	0	$\ldots$	3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Doc_N	0	1	$\ldots$	1

Inner product of Documents (rows): $\textbf{Doc}_i^{'}\textbf{Doc}_l$

Inner product of Terms (columns): $\textbf{Word}_j^{'}\textbf{Word}_k$

Allows: measure of correlation of term usage across documents (heuristically: partition words, based on usage in documents)

Latent Semantic Analysis: Reduce information in matrix using linear algebra (provides similar results, difficult to generalize)

Biclustering: Models that partition documents and words simultaneously

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})$$

Why does this work $\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\alpha} | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) p(\boldsymbol{T} | \boldsymbol{\pi}) \underbrace{p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \rightsquigarrow$ Greater weight on terms that occur together

Why does this work$\leadsto$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) p(\boldsymbol{\pi} | \boldsymbol{\alpha}) \underbrace{p(\boldsymbol{T} | \boldsymbol{\pi})}_{2} \underbrace{p(\boldsymbol{X} | \boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \leadsto$ Greater weight on terms that occur together
2) $\boldsymbol{\pi} \leadsto$ Greater weight on indicators that appear more regularly

Why does this work$\rightsquigarrow$ Co-occurrence logic (h/t Colorado Reed Tutorial)

$$p(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\Theta}, \alpha | \boldsymbol{X}) \quad \propto \quad p(\boldsymbol{\alpha}) \underbrace{p(\boldsymbol{\pi}|\boldsymbol{\alpha})}_{3} \underbrace{p(\boldsymbol{T}|\pi)}_{2} \underbrace{p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T})}_{1}$$

1) $\boldsymbol{\theta} \rightsquigarrow$ Greater weight on terms that occur together
2) $\boldsymbol{\pi} \rightsquigarrow$ Greater weight on indicators that appear more regularly
3) $\boldsymbol{\alpha} \rightsquigarrow$ Emphasis on $\boldsymbol{\pi}$ with greater weight

Validation $\rightsquigarrow$ Topic Intrusion

Discussed several validations

- Labeling paragraphs
  - Identify separating words automatically
  - Label topics manually (read!)
- Statistical methods
  1) Entropy
  2) Exclusivity
  3) Cohesiveness
- Experiment Based Methods
  - Word intrusion $\rightsquigarrow$ topic validity
  - Topic intrusion $\rightsquigarrow$ model fit

Validation $\rightsquigarrow$ Topic Intrusion

1) Ask research assistant to read paragraph
2) Construct experiment
    - For the document, select top three topics
    - Select a fourth topic
    - Show participant, ask her/him to identify intruder

Higher identification $\rightsquigarrow$ topics are a better model of text

Measuring Cluster Performance: Out of Sample Prediction

How well does our model perform?

Measuring Cluster Performance: Out of Sample Prediction

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

Measuring Cluster Performance: Out of Sample Prediction

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

Measuring Cluster Performance: Out of Sample Prediction

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

Measuring Cluster Performance: Out of Sample Prediction

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

What's Prediction Got to Do With It?

- Prediction $\rightsquigarrow$ One Task

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

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

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

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

- Compare perplexity with human based evaluations

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

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

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

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

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

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

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

(Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

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

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

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

- Statistics: measure cohesiveness and exclusivity (Roberts, et al 2017 Forthcoming)

What's Prediction Got to Do With It?

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

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

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

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

- Statistics: measure cohesiveness and exclusivity (Roberts, et al 2017 Forthcoming)
- Experiments: measure topic and cluster quality

Experimental Approaches

Mathematical approaches

Experimental Approaches

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

Experimental Approaches

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

Experimental Approaches

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

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

Humans$\rightsquigarrow$ read texts

Humans$\rightsquigarrow$ use cluster output

Experimental Approaches

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

Experimental Approaches

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

1) Topic Quality

Experimental Approaches

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

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

Humans $\rightsquigarrow$ read texts

Humans $\rightsquigarrow$ use cluster output

Do humans think the model is performing well?

1) Topic Quality
2) Cluster Quality

Experimental Approaches

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Deploy on Mechanical Turk

Stylometry⇝ Who Wrote Disputed Federalist Papers?

Federalist papers ⇝ Mosteller and Wallace (1963)

- Persuade citizens of New York State to adopt constitution
- Canonical texts in study of American politics
- 77 essays
  - Published from 1787-1788 in Newspapers
  - And under the name Publius, anonymously

Who Wrote the Federalist papers?

- Jay wrote essays 2, 3, 4,5, and 64
- Hamilton: wrote 43 papers
- Madison: wrote 12 papers

Disputed: Hamilton or Madison?

- Essays: 49-58, 62, and 63
- Joint Essays: 18-20

Task: identify authors of the disputed papers.

Task: Classify papers as Hamilton or Madison using dictionary methods

Setting up the Analysis

Training $\rightsquigarrow$ papers Hamilton, Madison are known to have authored
Test $\rightsquigarrow$ unlabeled papers
Preprocessing:

- Hamilton/Madison both discuss similar issues
- Differ in extent they use stop words
- Focus analysis on the stop words

Setting up the Analysis

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N) = (\text{Hamilton}, \text{Hamilton}, \text{Madison}, ..., \text{Hamilton})$
  $N \times 1$ matrix with author labels
- Define the number of words in federalist paper i as num_i

$$\boldsymbol{X} = \begin{pmatrix} \frac{1}{\text{num}_1} & \frac{2}{\text{num}_1} & \frac{0}{\text{num}_1} & \cdots & \frac{3}{\text{num}_1} \\ \frac{0}{\text{num}_2} & \frac{1}{\text{num}_2} & \frac{0}{\text{num}_2} & \cdots & \frac{0}{\text{num}_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{0}{\text{num}_N} & \frac{0}{\text{num}_N} & \frac{1}{\text{num}_N} & \cdots & \frac{0}{\text{num}_N} \end{pmatrix}$$

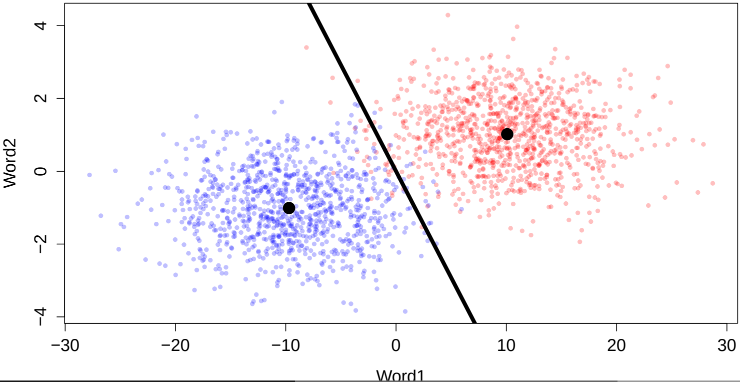
  $N \times J$ counting stop word usage rate
- $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$
  Word weights.

Objective Function

Heuristically: find $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \ldots, \theta_J^*)$ used to create score

$$p_i = \sum_{j=1}^{J} \theta_j^* X_{ij}$$

that maximally discriminates between categories



Objective Function

Define:

$$
\begin{aligned}
\boldsymbol{\mu}_{\text{Madison}} &= \frac{1}{N_{\text{Madison}}} \sum_{i=1}^{N} I(Y_i = \text{Madison}) \boldsymbol{X}_i \\
\boldsymbol{\mu}_{\text{Hamilton}} &= \frac{1}{N_{\text{Hamilton}}} \sum_{i=1}^{N} I(Y_i = \text{Hamilton}) \boldsymbol{X}_i
\end{aligned}
$$

Objective Function

We can then define functions that describe the "projected" mean and variance for each author

$$
\begin{aligned}
g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison}) &= \frac{1}{N_{\text{Madison}}} \sum_{i=1}^{N} I(Y_i = \text{Madison})\boldsymbol{\theta}^{'}\boldsymbol{X}_i = \boldsymbol{\theta}^{'}\boldsymbol{\mu}_{\text{Madison}} \\
g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) &= \frac{1}{N_{\text{Hamilton}}} \sum_{i=1}^{N} I(Y_i = \text{Hamilton})\boldsymbol{\theta}^{'}\boldsymbol{X}_i = \boldsymbol{\theta}^{'}\boldsymbol{\mu}_{\text{Hamilton}} \\
s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison}) &= \sum_{i=1}^{N} I(Y_i = \text{Madison})(\boldsymbol{\theta}^{'}\boldsymbol{X}_i - \boldsymbol{\theta}^{'}\boldsymbol{\mu}_{\text{Madison}})^2 \\
s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) &= \sum_{i=1}^{N} I(Y_i = \text{Hamilton})(\boldsymbol{\theta}^{'}\boldsymbol{X}_i - \boldsymbol{\theta}^{'}\boldsymbol{\mu}_{\text{Hamilton}})^2
\end{aligned}
$$

Objective Function$\leadsto$ Optimization

$$
\begin{aligned}
f(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}) &= \frac{(g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) - g(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison}))^2}{s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Hamilton}) + s(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}, \text{Madison})} \\
&= \frac{\left(\boldsymbol{\theta}^{'}(\boldsymbol{\mu}_{\text{Hamilton}} - \boldsymbol{\mu}_{\text{Madison}})\right)^2}{\text{Scatter}_{\text{Hamilton}} + \text{Scatter}_{\text{Madison}}}
\end{aligned}
$$

Optimization$\leadsto$ find $\boldsymbol{\theta}^{*}$ to maximize $f(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y})$, assuming independence across dimensions.
(Fisher's) Linear Discriminant Analysis

Optimization$\rightsquigarrow$ Word Weights

For each word j, construct weight θ_j^*,

$$
\begin{aligned}
\mu_{j,\text{Hamilton}} &= \frac{\sum_{i=1}^{N} I(Y_i = \text{Hamilton})X_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N} I(Y_i = \text{Hamilton})X_{ij}} \\
\mu_{j,\text{Madison}} &= \frac{\sum_{i=1}^{N} I(Y_i = \text{Madison})X_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N} I(Y_i = \text{Madison})X_{ij}} \\
\sigma^2_{j,\text{Hamilton}} &= \text{Var}(X_{i,j}|\text{Hamilton}) \\
\sigma^2_{j,\text{Madison}} &= \text{Var}(X_{i,j}|\text{Madison})
\end{aligned}
$$

We can then generate weight θ_j^* as

$$
\theta_j^* = \frac{\mu_{j,\text{Hamilton}} - \mu_{j,\text{Madison}}}{\sigma^2_{j,\text{Hamilton}} + \sigma^2_{j,\text{Madison}}}
$$

Optimization$\rightsquigarrow$ Trimming the Dictionary

- Trimming weights: Focus on discriminating words (very simple regularization)
- Cut off: For all $|\theta_j^*| < 0.025$ set $\theta_j^* = 0$.

Classification ⇝ Determining Authorship

For each disputed document i, compute discrimination statistic

$$p_i = \sum_{j=1}^{J} \theta_j^{*} X_{ij}$$

$p_i \rightsquigarrow$ classification (linear discriminator)

- Above midpoint in training set $\rightarrow$ Hamilton text
- Below midpoint in training set $\rightarrow$ Madison text

Findings: Madison is the author of the disputed federalist papers.

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

- Stylometry $\rightsquigarrow$ Classify Authors

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

- Stylometry $\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

- Stylometry $\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\rightsquigarrow$ Customized to particular setting

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

- Stylometry $\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

- Stylometry $\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language

Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)

Inferring Separating Words

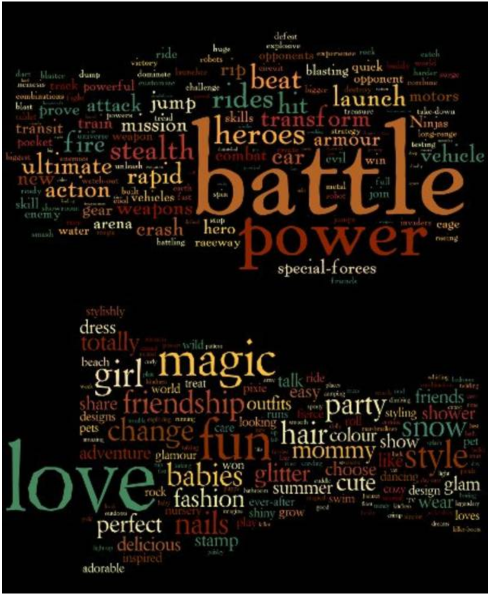
Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising

Inferring Separating Words



Inferring Separating Words

Classification$\rightsquigarrow$ Custom Dictionaries

- Stylometry$\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification$\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
- Difference in Language across groups$\rightsquigarrow$ Labeling output from Clustering/Topic Models

Inferring Separating Words

Classification $\rightsquigarrow$ Custom Dictionaries

- Stylometry $\rightsquigarrow$ Classify Authors
- Dictionary based classification $\rightsquigarrow$ Gentzkow and Shapiro (2010) and measures of media slant
- Dictionary based classification $\rightsquigarrow$ Customized to particular setting

Fictitious Prediction Problem $\rightsquigarrow$ Infer words that are indicative of some class/group

- Difference in Republican, Democratic language $\rightsquigarrow$ Partisan words
- Difference in Liberal, Conservative language $\rightsquigarrow$ Ideological Language
- Difference in Secret/Not Secret Language $\rightsquigarrow$ Secretive Language (Gill and Spirling 2014)
- Difference in Toy advertising
- Difference in Language across groups $\rightsquigarrow$ Labeling output from Clustering/Topic Models

Vague and Difficult to derive before hand

Congressional Language Across Sources

Congregational Language Across Sources

Congressional Press Releases and Floor Speeches

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
- One Answer: texts used for different purposes

Congressional Language Across Sources

Congressional Press Releases and Floor Speeches

- Collected 64,033 press releases
- Problem: are they distinct from floor statements (approx. 52,000 during same time)?
  - Yes: press releases have different purposes, targets, and need not relate to official business
  - No: press releases are just reactive to floor activity, will follow floor statements
- Deeper question: what does it mean for two text collections to be different?
- One Answer: texts used for different purposes
- Partial answer: identify words that distinguish press releases and floor speeches

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

Mutual Information

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release
  - Guess press release/floor statement

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty
        - Unrelated: Conditional uncertainty = uncertainty

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty
        - Unrelated: Conditional uncertainty = uncertainty
        - Perfect predictor: Conditional uncertainty = 0

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0
- Mutual information(X_j): uncertainty - conditional uncertainty (X_j)

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
  - Randomly sample a press release
  - Guess press release/floor statement
  - Uncertainty about guess
    - Maximum: No. press releases = No. floor statements
    - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
  - Condition on presence of word X_j
  - Randomly sample a press release
  - Guess press release/floor statement
  - Word presence reduces uncertainty
    - Unrelated: Conditional uncertainty = uncertainty
    - Perfect predictor: Conditional uncertainty = 0
- Mutual information(X_j): uncertainty - conditional uncertainty (X_j)
  - Maximum: Uncertainty $\rightarrow$ X_j is perfect predictor

A Method for Identifying Distinguishing Words

Mutual Information

- Unconditional uncertainty (entropy):
    - Randomly sample a press release
    - Guess press release/floor statement
    - Uncertainty about guess
        - Maximum: No. press releases = No. floor statements
        - Minimum : All documents in one category
- Conditional uncertainty (X_j) (conditional entropy)
    - Condition on presence of word X_j
    - Randomly sample a press release
    - Guess press release/floor statement
    - Word presence reduces uncertainty
        - Unrelated: Conditional uncertainty = uncertainty
        - Perfect predictor: Conditional uncertainty = 0
- Mutual information(X_j): uncertainty - conditional uncertainty (X_j)
    - Maximum: Uncertainty $\rightarrow X_j$ is perfect predictor
    - Minimum: $0 \rightarrow X_j$ fails to separate speeches and floor statements

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech

A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release
- $\Pr(\text{Speech}) \equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release
- $\Pr(\text{Speech}) \equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre}, \text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits

A Method for Identifying Distinguishing Words

- Pr(Press) $\equiv$ Probability selected document press release
- Pr(Speech) $\equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre}, \text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits
- Maximum: $\Pr(\text{Press}) = \Pr(\text{Speech}) = 0.5$

A Method for Identifying Distinguishing Words

- $\Pr(\text{Press}) \equiv$ Probability selected document press release
- $\Pr(\text{Speech}) \equiv$ Probability selected document speech
- Define entropy $H(\text{Doc})$

$$H(\text{Doc}) = - \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t) \log_2 \Pr(t)$$

- $\log_2$? Encodes bits
- Maximum: $\Pr(\text{Press}) = \Pr(\text{Speech}) = 0.5$
- Minimum: $\Pr(\text{Press}) \rightarrow 0$ (or $\Pr(\text{Press}) \rightarrow 1$)

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

$$H(\text{Doc}|X_j) \;=\; -\sum_{s=0}^{1} \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

$$H(\text{Doc}|X_j) \quad = \quad -\sum_{s=0}^{1} \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

- Maximum: X_j unrelated to Press Releases/Floor Speeches

A Method for Identifying Distinguishing Words

- Consider presence/absence of word X_j
- Define conditional entropy $H(\text{Doc}|X_j)$

$$H(\text{Doc}|X_j) \;=\; -\sum_{s=0}^{1} \sum_{t \in \{\text{Pre},\text{Spe}\}} \Pr(t, X_j = s) \log_2 \Pr(t|X_j = s)$$

- Maximum: X_j unrelated to Press Releases/Floor Speeches
- Minimum: X_j is a perfect predictor of press release/floor speech

A Method for Identifying Distinguishing Words

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) \quad = \quad H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$.

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) \quad = \quad H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$.

Bigger mutual information $\Rightarrow$ better discrimination

A Method for Identifying Distinguishing Words

- Define Mutual Information(X_j) as

$$\text{Mutual Information}(X_j) = H(\text{Doc}) - H(\text{Doc}|X_j)$$

- Maximum: entropy $\Rightarrow H(\text{Doc}|X_j) = 0$
- Minimum: $0 \Rightarrow H(\text{Doc}|X_j) = H(\text{Doc})$.

Bigger mutual information $\Rightarrow$ better discrimination

Objective function and optimization$\rightsquigarrow$ estimate probabilities that we then place in mutual information

A Method for Identifying Distinguishing Words

Formula for mutual information
(based on ML estimates of probabilities)

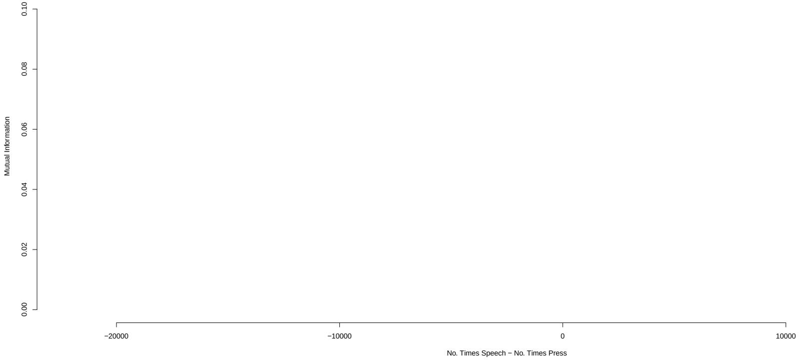
$$
\begin{aligned}
n_p &= \text{Number Press Releases} \\
n_s &= \text{Number of Speeches} \\
D &= n_p + n_s \\
n_j &= \sum_{i=1}^{D} X_{i,j} \qquad \text{(No. docs } X_j \text{ appears)} \\
n_{-j} &= \text{ No. docs } X_j \text{ does not appear} \\
n_{j,p} &= \text{No. press and } X_j \\
n_{j,s} &= \text{No. speech and } X_j \\
n_{-j,p} &= \text{No. press and not } X_j \\
n_{-j,s} &= \text{No. speech and not } X_j
\end{aligned}
$$

A Method for Identifying Distinguishing Words

Formula for Mutual Information

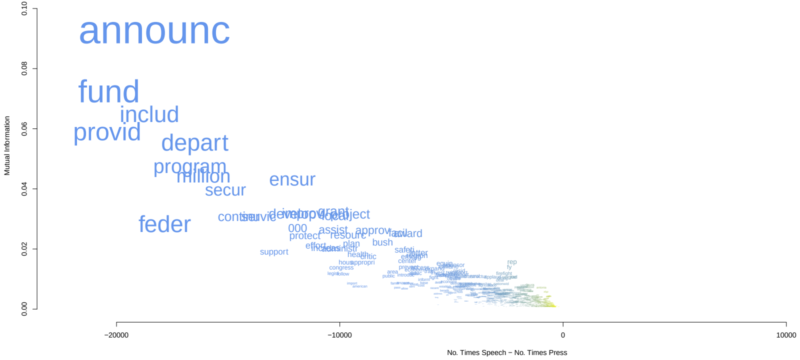
$$
\begin{aligned}
\mathrm{MI}(X_j) &= \frac{n_{j,p}}{D}\log_2\frac{n_{j,p}D}{n_j n_p} + \frac{n_{j,s}}{D}\log_2\frac{n_{j,s}D}{n_j n_s} \\
&\quad + \frac{n_{-j,p}}{D}\log_2\frac{n_{-j,p}D}{n_{-j} n_p} + \frac{n_{-j,s}}{D}\log_2\frac{n_{-j,s}D}{n_{-j} n_s}.
\end{aligned}
$$

What's Different About Press Releases



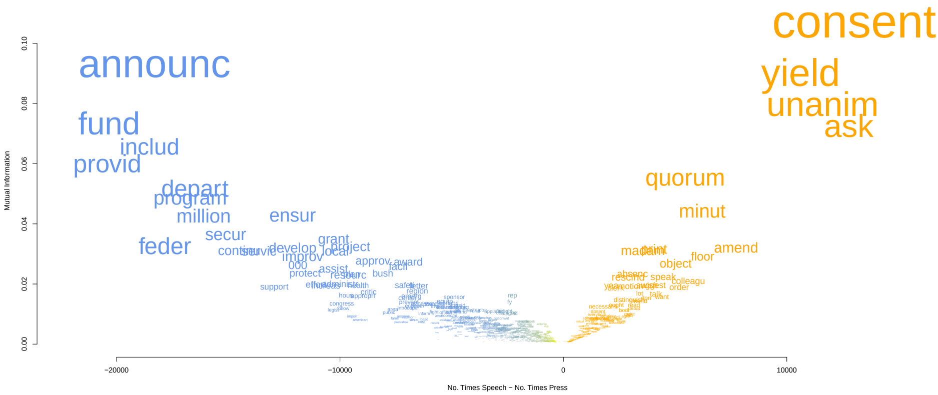
What's Different?

What's Different About Press Releases



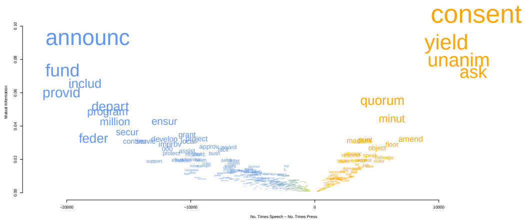
What's Different?

What's Different About Press Releases



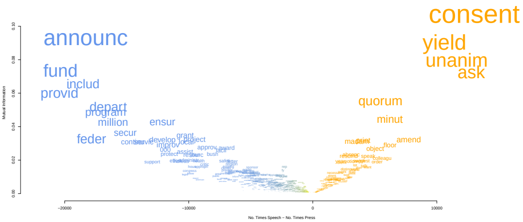
What's Different?

What's Different About Press Releases



What's Different?

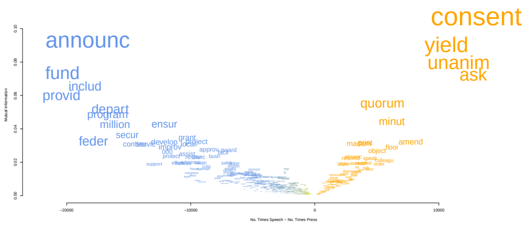
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming

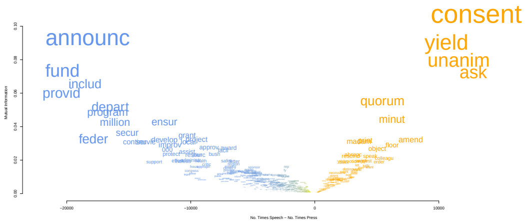
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words

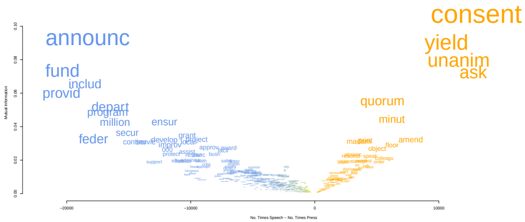
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification

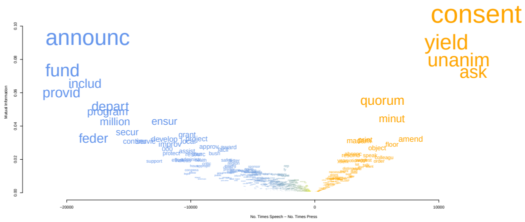
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches

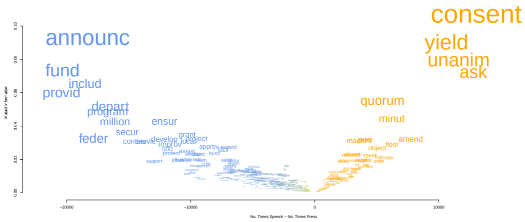
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
- Credit Claiming: 36% Press Releases, 4% Floor Speeches

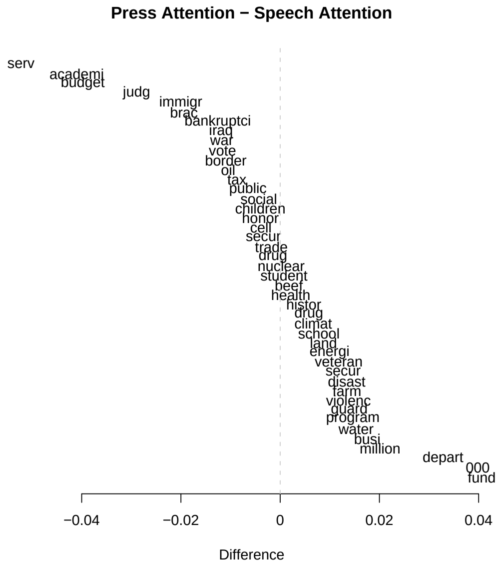
What's Different About Press Releases



What's Different?

- Press Releases: Credit Claiming
- Floor Speeches: Procedural Words
- Validate: Manual Classification
- Sample 500 Press Releases, 500 Floor Speeches
- Credit Claiming: 36% Press Releases, 4% Floor Speeches
- Procedural: 0% Press Releases, 44% Floor Speeches

What's Different About Press Releases



Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?

Fightin' Words $\leadsto$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\leadsto$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance

Fightin' Words$\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)$\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$P(E) = 1 - P(E^c)$$

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$\begin{aligned} P(E) &= 1 - P(E^c) \\ \text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \end{aligned}$$

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1-P(E))}}{\frac{P(F)}{1-P(F)}}
\end{aligned}$$

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$
\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1 - P(E))}}{\frac{P(F)}{1 - P(F)}} \\
\text{Log Odds Ratio}(E, F) &= \log\left(\frac{P(E)}{1 - P(E)}\right) - \log\left(\frac{P(F)}{1 - P(F)}\right)
\end{aligned}
$$

Fightin' Words$\leadsto$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009)$\leadsto$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$
\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1-P(E))}}{\frac{P(F)}{1-P(F)}} \\
\text{Log Odds Ratio}(E, F) &= \log\left(\frac{P(E)}{1 - P(E)}\right) - \log\left(\frac{P(F)}{1 - P(F)}\right)
\end{aligned}
$$

Strategy$\leadsto$ Construct objective function on *proportions* (and then calculate log-odds)

Fightin' Words $\rightsquigarrow$ An Introduction to Regularization

Monroe, Colaresi, and Quinn (2009) $\rightsquigarrow$ what makes a word partisan?
Argue for using Log Odds Ratio, weighted by variance
Recall: For some event E and F

$$\begin{aligned}
P(E) &= 1 - P(E^c) \\
\text{Odds}(E) &= \frac{P(E)}{1 - P(E)} \\
\text{Odds Ratio}(E, F) &= \frac{\frac{P(E))}{(1-P(E))}}{\frac{P(F)}{1-P(F)}} \\
\text{Log Odds Ratio}(E, F) &= \log\left(\frac{P(E)}{1 - P(E)}\right) - \log\left(\frac{P(F)}{1 - P(F)}\right)
\end{aligned}$$

Strategy $\rightsquigarrow$ Construct objective function on *proportions* (and then calculate log-odds)

Objective Function

Suppose we're interested in how a word separates partisan speech.
$\boldsymbol{Y} = (\text{Republican}, \text{Republican}, \text{Democrat}, \ldots, \text{Republican})$
$\boldsymbol{X} = \text{Unnormalized matrix of word counts } N \times J$
Define

$$\begin{aligned}
\boldsymbol{x}_{\text{Republican}} &= (\sum_{i=1}^{N} I(Y_i = \text{Republican})X_{i1}, \sum_{i=1}^{N} I(Y_i = \text{Republican})X_{i2}, \\
& \ldots, \sum_{i=1}^{N} I(Y_i = \text{Republican})X_{iJ})
\end{aligned}$$

with $N_{\text{Republican}} = \text{Total number of Republican words}$

Objective Function

Objective Function

$$\boldsymbol{\pi}_{\text{Republican}} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

Objective Function

$$\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}} | \boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}$$

Objective Function

$$
\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}
$$

This implies an objective function on $\boldsymbol{\pi}$,

Objective Function

$$\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) \quad \propto \quad p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y})$$

Objective Function

$$\begin{aligned}\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha})\\ \boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})\end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$\begin{aligned}p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) &\propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y})\\ &\propto \frac{\Gamma(\sum_{j=1}^{J}\alpha_j)}{\prod_j \Gamma(\alpha_j)}\prod_{j=1}^{J}\pi_j^{\alpha_j - 1}\pi_j^{x_{\text{Republican},j}}\end{aligned}$$

Objective Function

$$
\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} & \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} & \sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}
$$

This implies an objective function on $\boldsymbol{\pi}$,

$$
\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) & \propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y}) \\
& \propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \pi_j^{x_{\text{Republican},j}}
\end{aligned}
$$

$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y})$ is a Dirichlet distribution:

Objective Function

$$\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) &\propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y}) \\
&\propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \pi_j^{x_{\text{Republican},j}}
\end{aligned}$$

$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y})$ is a Dirichlet distribution:

$$\pi^{*}_{\text{Republican},j} = \frac{x_{\text{Republican},j} + \alpha_j}{N_{\text{Republican}} + \sum_{j=1}^{J} \alpha_j}$$

Objective Function

$$\begin{aligned}
\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}_{\text{Republican}} &\sim \text{Multinomial}(N_{\text{Republican}}, \boldsymbol{\pi}_{\text{Republican}})
\end{aligned}$$

This implies an objective function on $\boldsymbol{\pi}$,

$$\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y}) &\propto p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{x}_{\text{Republican}}|\boldsymbol{\pi}\boldsymbol{\alpha}, \boldsymbol{Y}) \\
&\propto \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \pi_j^{x_{\text{Republican},j}}
\end{aligned}$$

$p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{Y})$ is a Dirichlet distribution:

$$\pi^{*}_{\text{Republican},j} = \frac{x_{\text{Republican},j} + \alpha_j}{N_{\text{Republican}} + \sum_{j=1}^{J} \alpha_j}$$

Calculating Log Odds Ratio

Define $\log \text{Odds Ratio}_j$ as

$$\log \text{Odds Ratio}_j = \log\left(\frac{\pi_{\text{Republican},j}}{1-\pi_{\text{Republican},j}}\right) - \log\left(\frac{\pi_{\text{Democratic},j}}{1-\pi_{\text{Democratic},j}}\right)$$

$$\text{Var}(\log \text{Odds Ratio}_j) \approx \frac{1}{x_{jD}+\alpha_j} + \frac{1}{x_{jR}+\alpha_j}$$

$$\text{Std. Log Odds}_j = \frac{\log \text{Odds Ratio}_j}{\sqrt{\text{Var}(\log \text{Odds Ratio}_j)}}$$

Applying the Model

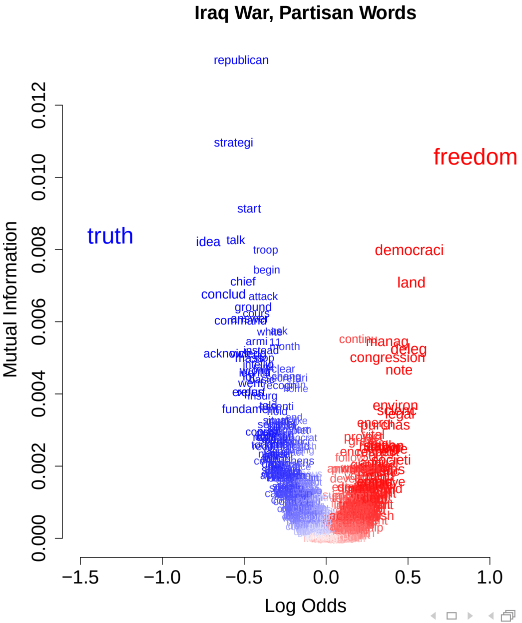
https://gist.github.com/thiagomarzagao/5851207

How do Republicans and Democrats differ in debate?

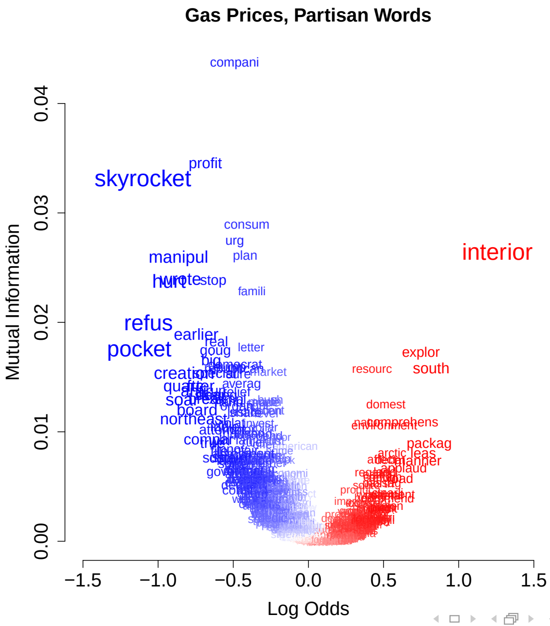
Condition on topic and examine word usage

- Press Releases (64,033)
- Topic Coded
- Given press release is about topic, what are the features that distinguish Republican and Democratic language?

Mutual Information, Standardized Log Odds

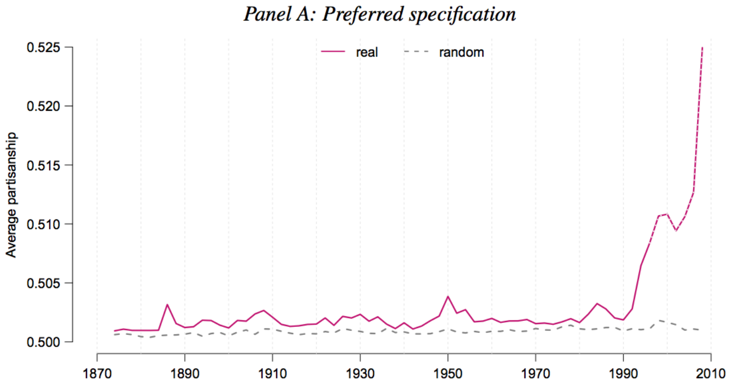


Mutual Information, Standardized Log Odds



Gentzkow, Shapiro, and Taddy (2017): Rhetorical Polarization

Figure 3: Average partisanship of speech, penalized estimates



Discovery

Where do concepts/ideas/questions come from?

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from
  does

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from
  does.not

Discovery

Where do concepts/ideas/questions come from?

- Text as Data (machine learning) methods can suggest idea
- Human in the loop: utility requires human presence
- Goal specific validation: once you have an organization it is yours and where it came from
  does.not.matter