# Introduction to Machine Learning for Social Science

## Class 12: Text Similarity & Distance

Rochelle Terman

Postdoctoral Fellow
Center for International Security Cooperation
Stanford University

February 22, 2018

# Announcements

## Group projects

- Instructions on canvas
- Fill out preference form by March 1.

# Announcements

## Group projects

- Instructions on canvas
- Fill out preference form by March 1.

## Midterm

- Pick up graded exam
- Class average: 28.04/35

# Announcements

Group projects

- Instructions on canvas
- Fill out preference form by March 1.

Midterm

- Pick up graded exam
- Class average: 28.04/35

Extra Credit (+10 points)

1. Find distinctive words using ‘text.no.noun‘ column.
2. Provide a brief qualitative analysis of the distinctive terms. What do these suggest about the ways American media portray women in the West vs. Middle East?
3. Classify documents using these terms. Identify whether the results are better or worse than the previous analysis, and briefly explain why you think this happened.

# Plan for the day

1. Loose ends: Evaluating dictionary methods and distinctive methods.
2. Introducing unsupervised learning.
3. Text similarity and distance.

# Evaluation and validity for Dictionary Methods & Distinctive Words

- How do we choose between different “distinctive word” metrics?

# Evaluation and validity for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?

# Evaluation and validity for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we validate our findings?

# Evaluation and validity for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we validate our findings?

Three validation strategies

# Evaluation and validity for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we validate our findings?

Three validation strategies

- Face validity (do these results make sense?)

# Evaluation and validity for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we validate our findings?

Three validation strategies

- Face validity (do these results make sense?)
- Convergence (do different metrics lead to the same result?)

# Evaluation and validity for Dictionary Methods & Distinctive Words

- How do we choose between different "distinctive word" metrics?
- How do we choose between dictionaries?
- How do we validate our findings?

Three validation strategies

- Face validity (do these results make sense?)
- Convergence (do different metrics lead to the same result?)
- "Gold Standard" (do our results align with human coding?)

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.
- Clear set of tools: multiple regression, logit, LASSO, etc.

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.
- Clear set of tools: multiple regression, logit, LASSO, etc.
- Clear understanding of how to assess the quality the results: test MSE, cross-validation.

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.
- Clear set of tools: multiple regression, logit, LASSO, etc.
- Clear understanding of how to assess the quality the results: test MSE, cross-validation.

Unsupervised learning: We observe only the inputs, but no measure for the outputs. Our task is to learn relationships and structures from such data.

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.
- Clear set of tools: multiple regression, logit, LASSO, etc.
- Clear understanding of how to assess the quality the results: test MSE, cross-validation.

Unsupervised learning: We observe only the inputs, but no measure for the outputs. Our task is to learn relationships and structures from such data.

- No clear goal: exploratory data analysis.

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.
- Clear set of tools: multiple regression, logit, LASSO, etc.
- Clear understanding of how to assess the quality the results: test MSE, cross-validation.

Unsupervised learning: We observe only the inputs, but no measure for the outputs. Our task is to learn relationships and structures from such data.

- No clear goal: exploratory data analysis.
- No clear way to check our work (because we don't know the true answer.)

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.
- Clear set of tools: multiple regression, logit, LASSO, etc.
- Clear understanding of how to assess the quality the results: test MSE, cross-validation.

Unsupervised learning: We observe only the inputs, but no measure for the outputs. Our task is to learn relationships and structures from such data.

- No clear goal: exploratory data analysis.
- No clear way to check our work (because we don't know the true answer.)
- Still important and useful!

# Supervised v. Unsupervised Learning

Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.

- Clear goal: predict a response variable.
- Clear set of tools: multiple regression, logit, LASSO, etc.
- Clear understanding of how to assess the quality the results: test MSE, cross-validation.

Unsupervised learning: We observe only the inputs, but no measure for the outputs. Our task is to learn relationships and structures from such data.

- No clear goal: exploratory data analysis.
- No clear way to check our work (because we don't know the true answer.)
- Still important and useful!

Supervised learning and Unsupervised learning are not competitors!

*An online shopping site is creating three advertisements in order to market themselves to potential customers. To do so, they want to divide customers into groups that share certain characteristics, like age, gender, and zip code. Then they can design advertisements that appeal to each group.*

*An online shopping site is creating three advertisements in order to market themselves to potential customers. To do so, they want to divide customers into groups that share certain characteristics, like age, gender, and zip code. Then they can design advertisements that appeal to each group.*

$\rightsquigarrow$Cluster analysis / Clustering

*An online shopping site is creating three advertisements in order to market themselves to potential customers. To do so, they want to divide customers into groups that share certain characteristics, like age, gender, and zip code. Then they can design advertisements that appeal to each group.*

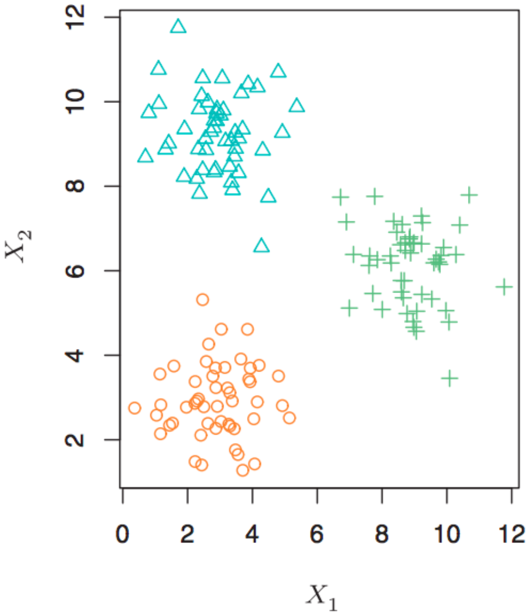
## $\rightsquigarrow$ Cluster analysis / Clustering

- Goal is to ascertain, on the basis of $x_1, x_2, ..., x_n$, whether the observations fall into relatively distinct groups.

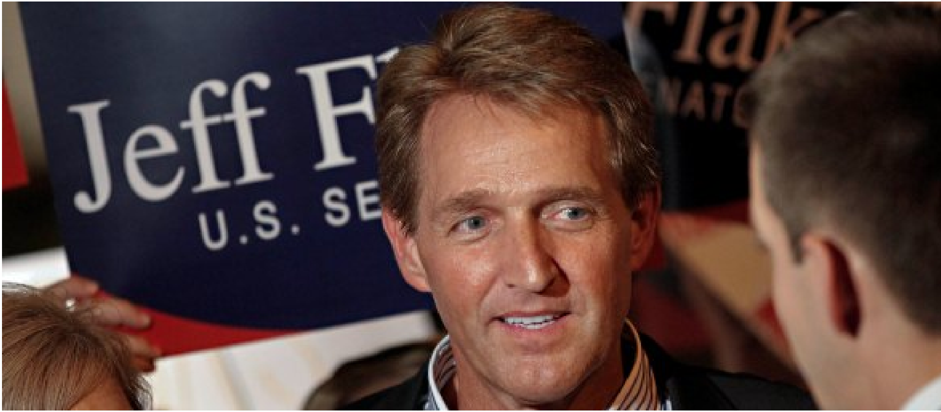
*An online shopping site is creating three advertisements in order to market themselves to potential customers. To do so, they want to divide customers into groups that share certain characteristics, like age, gender, and zip code. Then they can design advertisements that appeal to each group.*

## $\rightsquigarrow$Cluster analysis / Clustering

- Goal is to ascertain, on the basis of $x_1, x_2, ..., x_n$, whether the observations fall into relatively distinct groups.
- These groups are interesting because the may correspond to some category or quantity of interest.



12
10
8
6
4
2
$X_2$
0
2
4
6
8
10
12
$X_1$


Jeff F
U.S. SE
NAT

Today (and Tuesday): Cluster press releases

Goal: partition documents such that:

- similar documents are together
- dissimilar documents are apart

Method: Clustering methods

Game Plan:

1) What makes two data points (i.e. documents) similar?
2) How do we find a good partition?
3) How do we interpret the clusters?

Key Terms:

- (Multidimensional) Space
- Distance
- Euclidean Distance
- Cosine Distance
- Cluster Analysis / Clustering
- K-means
- Centroid

What makes two documents similar?

What makes two documents similar?

- Similar use of language $\rightsquigarrow$ complicated

What makes two documents similar?

- Similar use of language $\rightsquigarrow$ complicated
- Similar word count vectors $\rightsquigarrow$ simple

What makes two documents similar?

- Similar use of language $\rightsquigarrow$ complicated
- Similar word count vectors $\rightsquigarrow$ simple

Similar $=$ Geometrically Close

Dissimilar $=$ Geometrically Distant

# 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

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

By transforming our text into a word count vector, we are representing it as a point in a multidimensional space

# Texts and Geometry

Consider a document-term matrix

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

By transforming our text into a word count vector, we are representing it as a point in a multidimensional space

- Provides a geometry

# Texts and Geometry

Consider a document-term matrix

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

By transforming our text into a word count vector, we are representing it as a point in a multidimensional space

- Provides a geometry
- Natural notions of distance and similarity

# Texts and Geometry

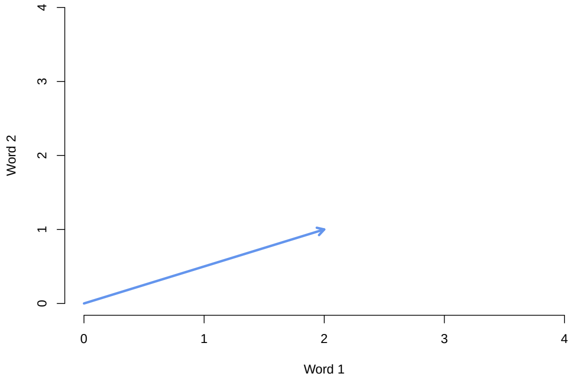
Consider a document-term matrix

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

By transforming our text into a word count vector, we are representing it as a point in a multidimensional space

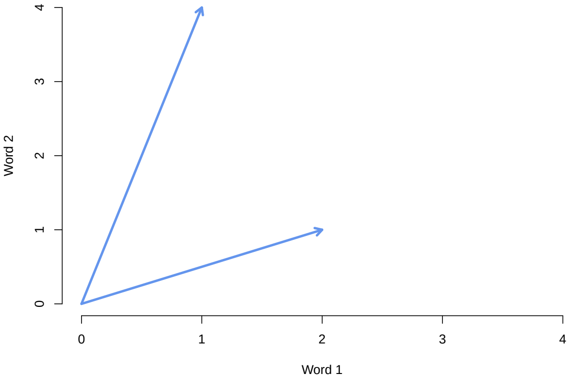
- Provides a geometry
- Natural notions of distance and similarity
- Tools from linear algebra to calculate distances mathematically.

# Texts in Space



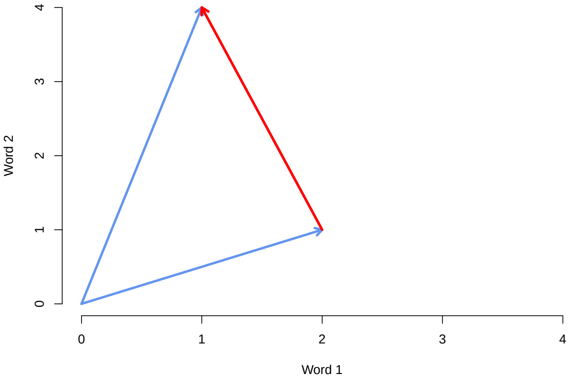

Doc1 = "Wait? No wait." $\rightsquigarrow (2, 1)$

# Texts in Space




Doc1 = "Wait? No wait." $\rightsquigarrow (2,1)$
Doc2 = "No, wait! No, no, no!" $\rightsquigarrow (1,4)$

# Texts in Space




Doc1 = "Wait? No wait." $\rightsquigarrow (2, 1)$
Doc2 = "No, wait! No, no, no!" $\rightsquigarrow (1, 4)$

Suppose $\boldsymbol{X}_1 = (1, 4)$ and $\boldsymbol{X}_2 = (2, 1)$.

The Euclidean distance (aka norm) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (or from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is the length of the line segment connecting them.

Suppose $\boldsymbol{X}_1 = (1, 4)$ and $\boldsymbol{X}_2 = (2, 1)$.

The Euclidean distance (aka norm) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (or from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is the length of the line segment connecting them.

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) = d(\boldsymbol{X}_2, \boldsymbol{X}_1) \quad = \quad \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2}$$

Suppose $\boldsymbol{X}_1 = (1, 4)$ and $\boldsymbol{X}_2 = (2, 1)$.

The Euclidean distance (aka norm) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (or from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is the length of the line segment connecting them.

$$\begin{aligned} d(\boldsymbol{X}_1, \boldsymbol{X}_2) = d(\boldsymbol{X}_2, \boldsymbol{X}_1) &= \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2} \\ &= \sqrt{(1-2)^2 + (4-1)^2} \end{aligned}$$

Suppose $\boldsymbol{X}_1 = (1, 4)$ and $\boldsymbol{X}_2 = (2, 1)$.

The Euclidean distance (aka norm) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (or from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is the length of the line segment connecting them.

$$\begin{aligned} d(\boldsymbol{X}_1, \boldsymbol{X}_2) = d(\boldsymbol{X}_2, \boldsymbol{X}_1) &= \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2} \\ &= \sqrt{(1-2)^2 + (4-1)^2} \\ &= \sqrt{10} \end{aligned}$$

Suppose $\boldsymbol{X}_1 = (1, 4)$ and $\boldsymbol{X}_2 = (2, 1)$.

The Euclidean distance (aka norm) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (or from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is the length of the line segment connecting them.

$$
\begin{aligned}
d(\boldsymbol{X}_1, \boldsymbol{X}_2) = d(\boldsymbol{X}_2, \boldsymbol{X}_1) &= \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2} \\
&= \sqrt{(1-2)^2 + (4-1)^2} \\
&= \sqrt{10}
\end{aligned}
$$

This generalizes beyond 2 dimensions!

Suppose $\boldsymbol{X}_1 = (1,4)$ and $\boldsymbol{X}_2 = (2,1)$.

The Euclidean distance (aka norm) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (or from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is the length of the line segment connecting them.

$$\begin{aligned} d(\boldsymbol{X}_1, \boldsymbol{X}_2) = d(\boldsymbol{X}_2, \boldsymbol{X}_1) &= \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2} \\ &= \sqrt{(1-2)^2 + (4-1)^2} \\ &= \sqrt{10} \end{aligned}$$

This generalizes beyond 2 dimensions!

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2 + \cdots + (x_{1,p} - x_{2,p})^2}$$

Suppose $\boldsymbol{X}_1 = (1, 4)$ and $\boldsymbol{X}_2 = (2, 1)$.

The Euclidean distance (aka norm) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (or from $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is the length of the line segment connecting them.

$$\begin{aligned} d(\boldsymbol{X}_1, \boldsymbol{X}_2) = d(\boldsymbol{X}_2, \boldsymbol{X}_1) &= \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2} \\ &= \sqrt{(1-2)^2 + (4-1)^2} \\ &= \sqrt{10} \end{aligned}$$

This generalizes beyond 2 dimensions!

$$\begin{aligned} d(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \sqrt{(x_{1,1} - x_{2,1})^2 + (x_{1,2} - x_{2,2})^2 + \cdots + (x_{1,p} - x_{2,p})^2} \\ &= \sqrt{\sum_{p=1}^{P} (x_{1p} - x_{2p})^2} \end{aligned}$$

# Test your knowledge

The Euclidean distance between any documents $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is:

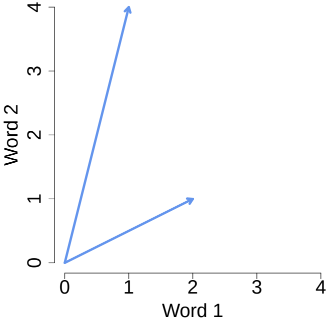
$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) \;=\; \sqrt{\sum_{p=1}^{P} (x_{1p} - x_{2p})^2}$$

Suppose

- $\boldsymbol{X}_1 =$ Oh na na na.
- $\boldsymbol{X}_2 =$ Oh, me? Na.

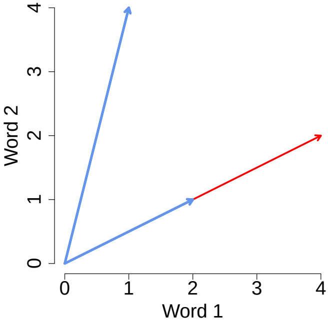
Calculate the euclidean distance between these two documents.

# Problem(?) with Euclidean Distance




$$
\begin{aligned}
\boldsymbol{X}_1 &= (2, 1) \\
\boldsymbol{X}_2 &= (1, 4) \\
d(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \sqrt{(1-2)^2 + (4-1)^2} \\
&= \sqrt{10}
\end{aligned}
$$

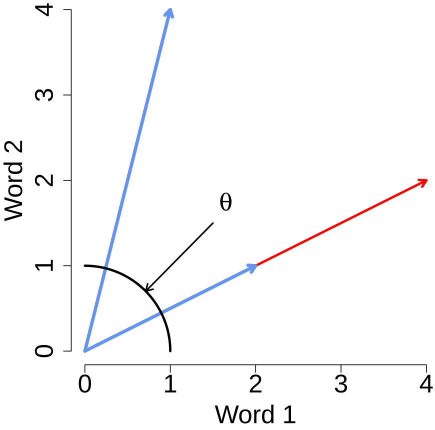
# Problem(?) with Euclidean Distance




$$\begin{aligned} \boldsymbol{X}_1 &= (2, 1) \\ \boldsymbol{X}_2 &= (1, 4) \\ \boldsymbol{X}_3 &= 2\boldsymbol{X}_1 = (4, 2) \\ d(\boldsymbol{X}_3, \boldsymbol{X}_2) &= \sqrt{(4-1)^2 + (2-4)^2} \\ &= \sqrt{13} \end{aligned}$$

Euclidean distance depends on document-length.

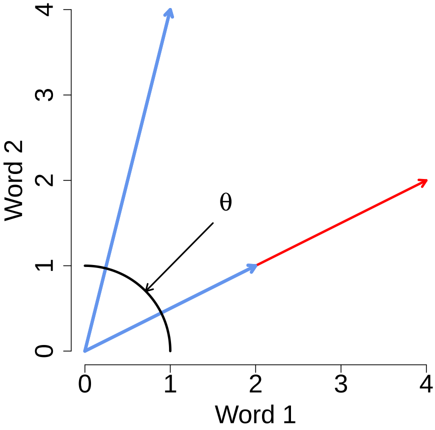
# Cosine Similarity




Cosine Similarity

- Takes into consideration documents length.

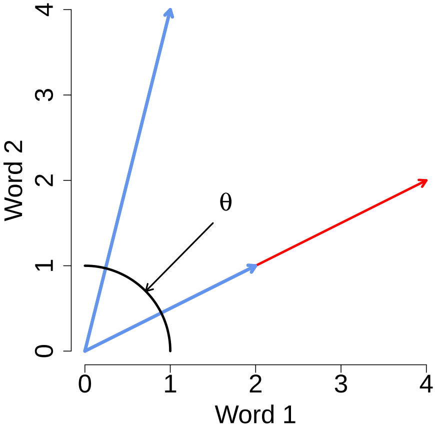
# Cosine Similarity




Cosine Similarity

- Takes into consideration documents length.
- Measures cosine of the angle ($\theta$) between vectors.

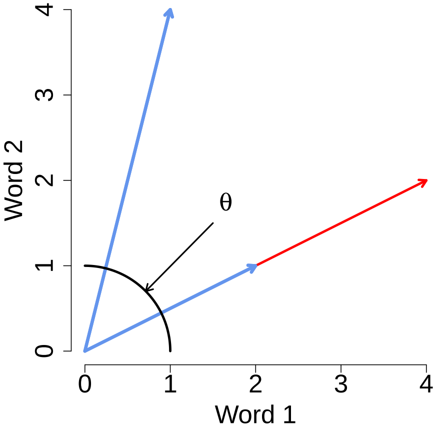
# Cosine Similarity




Cosine Similarity

- Takes into consideration documents length.
- Measures cosine of the angle ($\theta$) between vectors.
- Measure of similarity (rather than distance) ranging between 0 and 1.

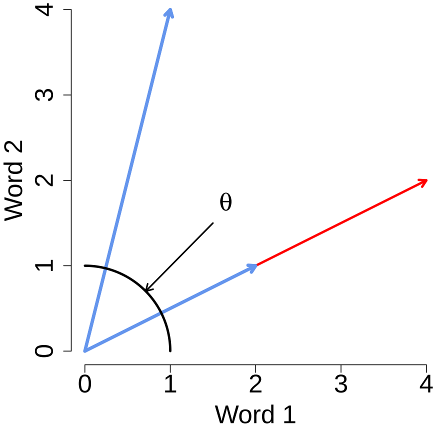
# Cosine Similarity




Cosine Similarity

- Takes into consideration documents length.
- Measures cosine of the angle ($\theta$) between vectors.
- Measure of similarity (rather than distance) ranging between 0 and 1.
- To convert to distance (or dissimilarity), take $1 - \cos\theta$ .

# Cosine Similarity




Cosine Similarity

- Takes into consideration documents length.
- Measures cosine of the angle ($\theta$) between vectors.
- Measure of similarity (rather than distance) ranging between 0 and 1.
- To convert to distance (or dissimilarity), take $1 - \cos\theta$ .

What makes two data points (i.e. documents) similar?

## What makes two data points (i.e. documents) similar?

- Similar = Geometrically close
- Euclidean distance
- Cosine distance
- Many more! (as always...)

What makes two data points (i.e. documents) similar?

- Similar = Geometrically close
- Euclidean distance
- Cosine distance
- Many more! (as always...)

Why do we care?

- Distances $\rightsquigarrow$ clustering.
- Other applications
  - Plagiarism,
  - Diffusion of policy

## What makes two data points (i.e. documents) similar?

- Similar = Geometrically close
- Euclidean distance
- Cosine distance
- Many more! (as always...)

Why do we care?

- Distances $\rightsquigarrow$ clustering.
- Other applications
  - Plagiarism,
  - Diffusion of policy

## Tomorrow

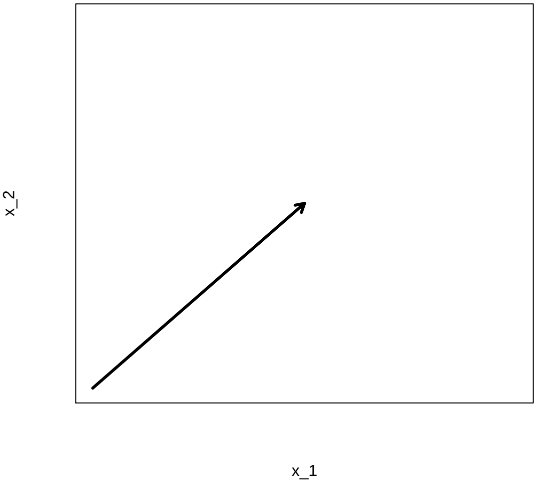
- How do we find a good partition?
- How do we interpret the clusters?

To the `R` code!

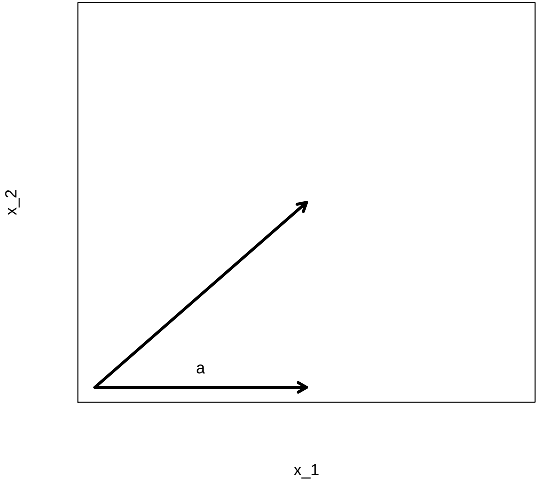
# Bonus Slides

For those who heart math.

# Vector Length

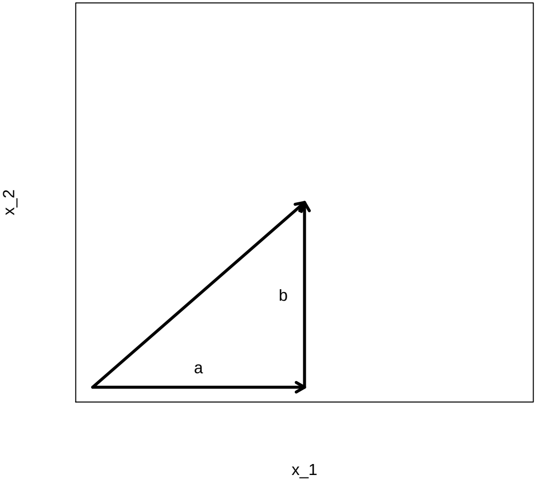


# Vector Length



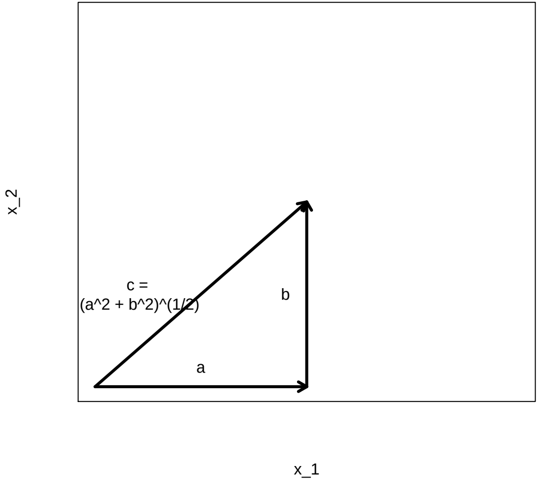

- Pythogorean Theorem: Side with length $a$

# Vector Length



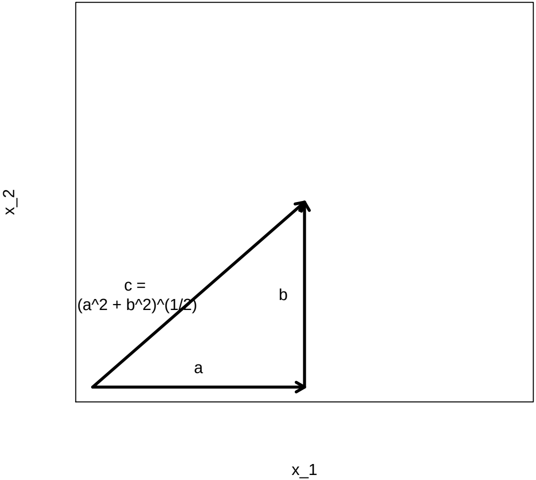

- Pythogorean Theorem: Side with length *a*
- Side with length *b* and right triangle

# Vector Length




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

# Vector Length




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

# Vector (Euclidean) Length

Suppose $\boldsymbol{X}_i$ is a document (row from an $N \times K$ document-term matrix).

Then, we will define its length as

$$
\begin{aligned}
||\boldsymbol{X}_i|| &= \sqrt{(\boldsymbol{X}_i \cdot \boldsymbol{X}_i)} \\
&= \sqrt{(X_{i1}^2 + X_{i2}^2 + X_{i3}^2 + \ldots + X_{iK}^2)} \\
&= \sqrt{\sum_{k=1}^{K} X_{ik}^2}
\end{aligned}
$$

# Cosine Similarity

# Cosine Similarity

$$\cos\theta \quad = \quad \left(\frac{X_1}{||X_1||}\right) \cdot \left(\frac{X_2}{||X_2||}\right)$$

# Cosine Similarity

$$
\begin{aligned}
\cos\theta &= \left(\frac{X_1}{||X_1||}\right) \cdot \left(\frac{X_2}{||X_2||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45)
\end{aligned}
$$

# Cosine Similarity

$$
\begin{aligned}
\cos\theta &= \left(\frac{X_1}{||X_1||}\right) \cdot \left(\frac{X_2}{||X_2||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45) \\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45)
\end{aligned}
$$

# Cosine Similarity

$$
\begin{aligned}
\cos\theta &= \left(\frac{X_1}{||X_1||}\right)\cdot\left(\frac{X_2}{||X_2||}\right)\\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45)\\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45)\\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97)
\end{aligned}
$$

# Cosine Similarity

$$
\begin{aligned}
\cos\theta &= \left(\frac{X_1}{||X_1||}\right)\cdot\left(\frac{X_2}{||X_2||}\right)\\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45)\\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45)\\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97)\\
(0.89, 0.45)\cdot(0.24, 0.97) &= 0.65
\end{aligned}
$$

# Cosine Similarity

$$\begin{aligned}
\cos\theta &= \left(\frac{X_1}{||X_1||}\right)\cdot\left(\frac{X_2}{||X_2||}\right)\\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45)\\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45)\\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97)\\
(0.89, 0.45)\cdot(0.24, 0.97) &= 0.65\\
\text{cos dissimilarity} &= 1-\cos\theta
\end{aligned}$$

# Cosine Similarity

$$
\begin{aligned}
\cos\theta &= \left(\frac{X_1}{||X_1||}\right) \cdot \left(\frac{X_2}{||X_2||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45) \\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45) \\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97) \\
(0.89, 0.45) \cdot (0.24, 0.97) &= 0.65 \\
\text{cos dissimilarity} &= 1 - \cos\theta
\end{aligned}
$$

# Cosine Similarity

$$
\begin{aligned}
\cos\theta &= \left(\frac{X_1}{||X_1||}\right) \cdot \left(\frac{X_2}{||X_2||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45) \\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45) \\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97) \\
(0.89, 0.45) \cdot (0.24, 0.97) &= 0.65 \\
\text{cos dissimilarity} &= 1 - \cos\theta
\end{aligned}
$$