Lesson4:
Descriptive Modelling of Similarity of Text
Unit3:
Vector space models for similarity

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

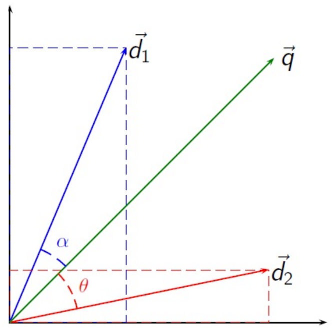
WeST People and Knowledge Networks

Completing this unit you should …

- Be familiar with the the vector space model for text documents
- Be aware of term frequency and (inverse) document frequency
- Have reviewed the definitions of base and dimension
- Realize that the angle between two vectors can be seen as a similarity measure

A model based on vector spaces

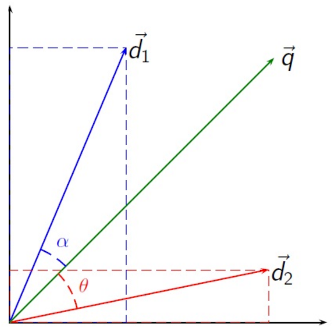
1) Model documents as vectors of words



A model based on vector spaces

Simple English
WIKIPEDIA

1) Model documents as vectors of words



2) calculating

Vector space Model

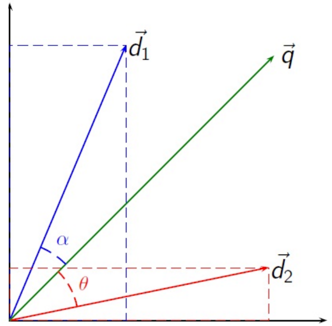
Distance between vectors

A model based on vector spaces

Matthew	Luke
δων δε πολλους των	ελεγεν ουν τοις
φαρισαιων και σαδδουκαιων	εκπορευομενοις οχλοις
ερχομενους επι το βαπτισμα	βαπτισθηναι υπ
αυτου ειπεν αυτοις	αυτου
γεννηματα εχιδνων τις	γεννηματα εχιδνων τις
υπεδειξεν υμιν φυγειν απο	υπεδειξεν υμιν φυγειν απο
της μελλουσης οργης	της μελλουσης οργης
ποιησατε ουν καρπους	ποιησατε ουν καρπους
αξιους της μετανοιας	αξιους της μετανοιας
και μη δοξητε λεγειν εν	και μη αρξησθε λεγειν εν
εαυτοις πατερα εχομεν τον	εαυτοις πατερα εχομεν τον
αβρααμ λεγω γαρ υμιν οτι	αβρααμ λεγω γαρ υμιν οτι
δυναται ο θεος εκ των λιθων	δυναται ο θεος εκ των λιθων
τουτων εγειραι τεκνα τω	τουτων εγειραι τεκνα τω
αβρααμ ηδη δε και η αξινη	αβρααμ ηδη δε και η αξινη
προς την ριζαν των δενδρων	προς την ριζαν των δενδρων
κειται παν ουν δενδρον μη	κειται παν ουν δενδρον μη
ποιουν καρπον καλον	ποιουν καρπον καλον
εκκοπτεται και εις πυρ	εκκοπτεται και εις πυρ
βαλλεται	βαλλεται

Simple English Wikipedia

1) Model documents as vectors of words



3) interpreting

2) calculating

Vector space Model

Distance between vectors

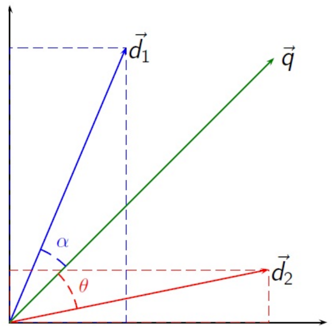
A model based on vector spaces

Simple English Wikipedia

Be able to make statements about similarity of documents

Matthew	Luke
δων δε πολλους των	ελεγεν ουν τοις
φαρισαιων και σαδδουκαιων	εκπορευομενοις οχλοις
ερχομενους επι το βαπτισμα	βαπτισθηναι υπ
αυτου ειπεν αυτοις	αυτου
γεννηματα εχιδνων τις	γεννηματα εχιδνων τις
υπεδειξεν υμιν φυγειν απο	υπεδειξεν υμιν φυγειν απο
της μελλουσης οργης	της μελλουσης οργης
ποιησατε ουν καρπους	ποιησατε ουν καρπους
αξιους της μετανοιας	αξιους της μετανοιας
και μη δοξητε λεγειν εν	και μη αρξησθε λεγειν εν
εαυτοις πατερα εχομεν τον	εαυτοις πατερα εχομεν τον
αβρααμ λεγω γαρ υμιν οτι	αβρααμ λεγω γαρ υμιν οτι
δυναται ο θεος εκ των λιθων	δυναται ο θεος εκ των λιθων
τουτων εγειραι τεκνα τω	τουτων εγειραι τεκνα τω
αβρααμ ηδη δε και η αξινη	αβρααμ ηδη δε και η αξινη
προς την ριζαν των δενδρων	προς την ριζαν των δενδρων
κειται παν ουν δενδρον μη	κειται παν ουν δενδρον μη
ποιουν καρπον καλον	ποιουν καρπον καλον
εκκοπτεται και εις πυρ	εκκοπτεται και εις πυρ
βαλλεται	βαλλεται

1) Model documents as vectors of words



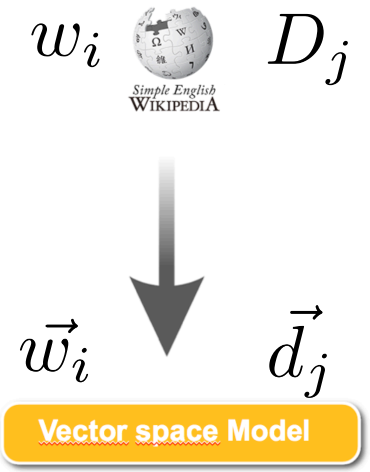
3) interpreting

2) calculating

Vector space Model

Distance between vectors

Pay close attention to notation



- Words $W = \{w_1, w_2, \ldots, w_n\}$
- Word Vectors $V = < \vec{w_1}, \vec{w_2}, \ldots, \vec{w_n} >$
- Document $D_j \in W^*$ are a sequence of words
  - So $D_j = w_{i_1} w_{i_2} \ldots w_{i_m}$ has a length of m
- Document vector $\vec{d_j} = \sum_{i=1}^{n} tf(w_i, D_j)\vec{w_i}$

Usually tf-idf is considered instead of tf!

- The document frequency is defined as
  $df(w_i) = |\{D_j | w_i \ in D_j\}|$
- Inverse document frequency is defined as
- $idf(w_i) = \log \frac{|D|}{df(w_i)}$ resulting in

$$tfidf(w_i, D_j) = tf(w_i, D_j) \times \log \frac{|D|}{df(w_i)}$$

- In the videos and slides for simplicity of numbers we will only use the term frequency

Example (generic language)

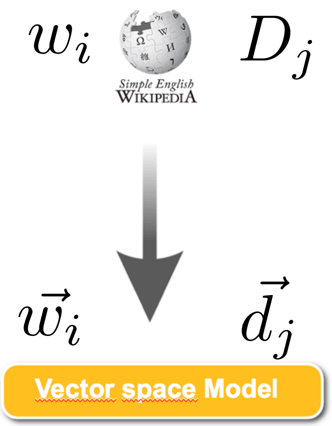
Be able to make statements about similarity of documents

- Let us assume 3 documents
  - D_1 = a a a b b a a b a b
  - D_2 = a b b
  - D_3 = a a b

- In our artificial language we have just two words “a” and “b”

- Which documents are similar to each other?

Choose a vector space and base



- Let $V = < \vec{a}, \vec{b} >$ be the vector space spanned by the words "a" and "b"

- $\vec{a} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is a base vector for word "a"

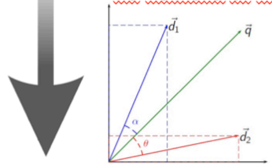
- Similarly for word "b" we have $\vec{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

- Chosing the base was a modelling choice!

Calculate the modelled document vectors d_1

- D_1 = a a a b b a a b a b
  - $tf(a, D_1) = 6$
  - $tf(b, D_1) = 4$

1) Model documents as vectors of words

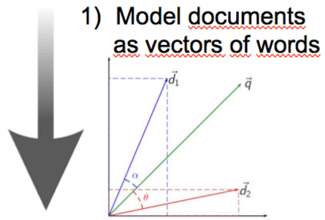


- Let us now create the document vectors

$$\vec{d_1} = \sum_{i=1}^{2} tf(w_i, D_1)\vec{w_i} = tf(a, D_1)\vec{a} + tf(b, D_1)\vec{b}$$

$$= 6\begin{pmatrix}1\\0\end{pmatrix} + 4\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}6\\4\end{pmatrix}$$

Calculate the modelled document vectors d_2



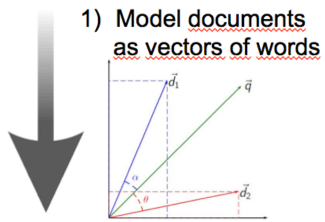
- D_2 = a b b
  - $tf(a, D_2) = 1$
  - $tf(b, D_2) = 2$
- Let us now create the document vectors

$$\vec{d_2} = \sum_{i=1}^{2} tf(w_i, D_2)\vec{w_i} = tf(a, D_2)\vec{a} + tf(b, D_2)\vec{b}$$

$$= 1\begin{pmatrix}1\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}1\\2\end{pmatrix}$$

Calculate the modelled document vectors d_3

- D_3 = a a b
  - $tf(a, D_3) = 2$
  - $tf(b, D_3) = 1$



- Let us now create the document vectors

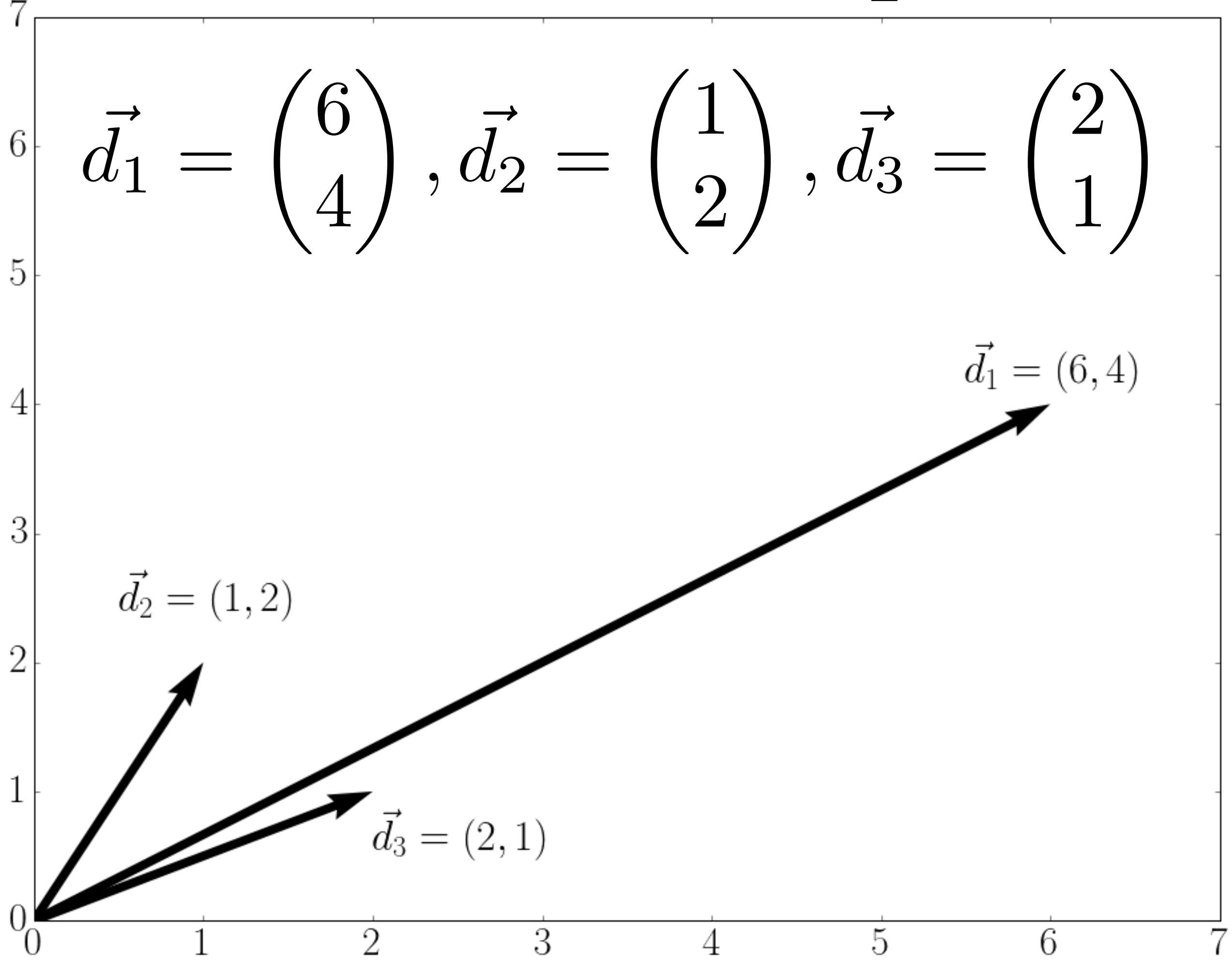
$$\vec{d_3} = \sum_{i=1}^{2} tf(w_i, D_3)\vec{w_i} = tf(a, D_3)\vec{a} + tf(b, D_3)\vec{b}$$

$$= 2\begin{pmatrix}1\\0\end{pmatrix} + 1\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}2\\1\end{pmatrix}$$

Digression: Drawing the document vectors

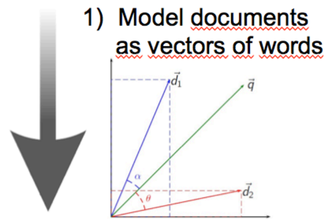
- D_1= a a a b b a a b a b, D_2= a b b, D_3= a a b

$$\vec{d_1} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}, \vec{d_2} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \vec{d_3} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$



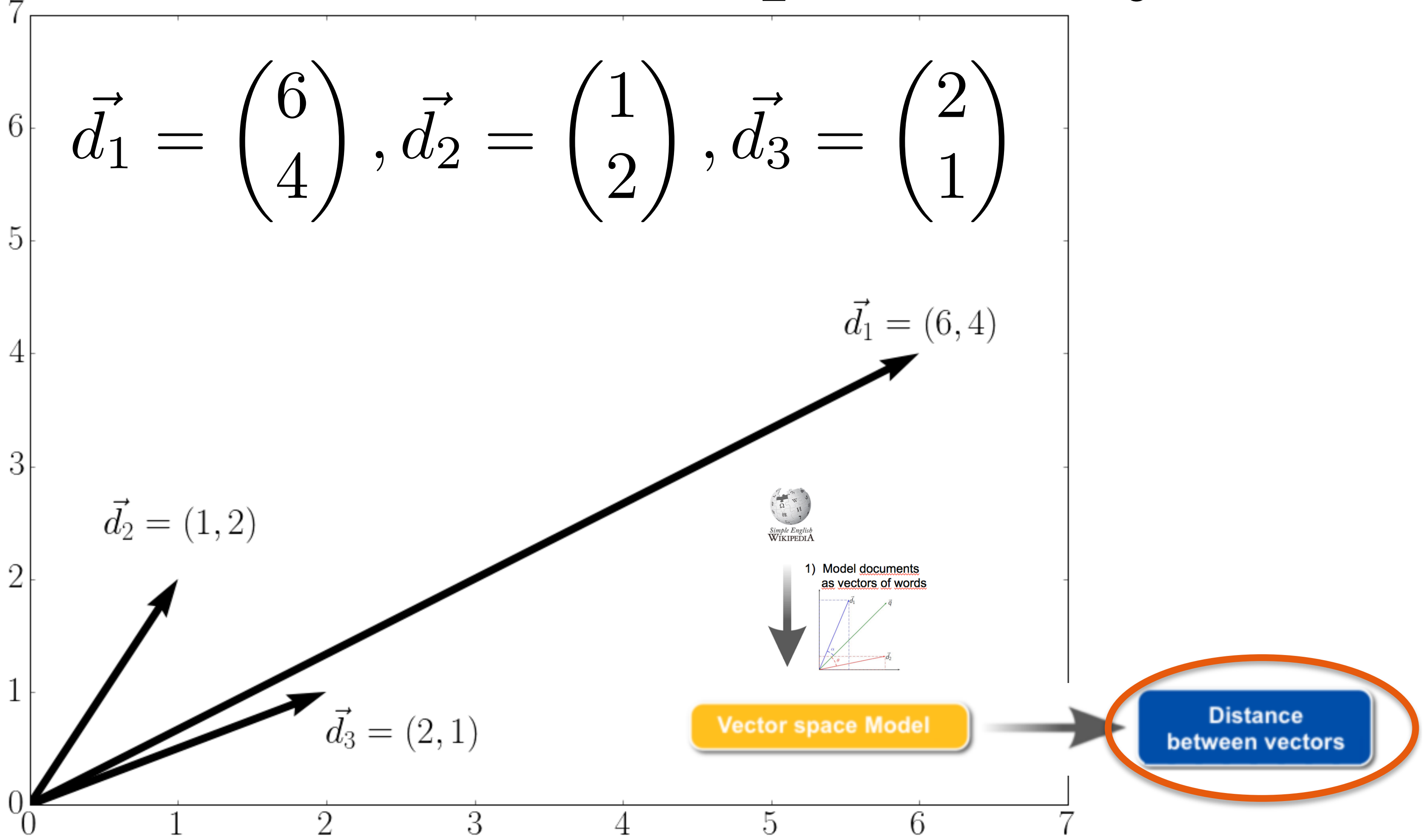
Simple English Wikipedia

1) Model documents as vectors of words



Which vectors are closest too each other?

- D_1= a a a b b a a b a b, D_2= a b b, D_3= a a b



Two ways of calculating the distance between two vectors d_i and d_j?

- **Euclidean** distance
  - "Take a rule and measure"
  - Take difference: $\vec{d} = \vec{d}_i - \vec{d}_j$
  - Calculate length of difference $||\vec{d}||$

- **Cosine** distance
  - "Take the angle between vectors"

Calculate Euclidean distance between two document vectors d_i and d_j

- Take the difference $\vec{d} = \vec{d}_i - \vec{d}_j$
- Calculate the length of the difference

$$||\vec{d}||^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((\vec{d}_i)_k - (\vec{d}_j)_k)^2$$

Calculate Euclidean distance between two document vectors d_i and d_j

- Take the difference $\vec{d} = \vec{d_i} - \vec{d_j}$
- Calculate the length of the difference

$$||\vec{d}||^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((\vec{d_i})_k - (\vec{d_j})_k)^2$$

The k-th component of vector $\vec{d_i}$
$(\vec{d_i})_k = tf(w_k, D_i)$ by definition

Calculate Euclidean distance between two document vectors d_i and d_j

- Take the difference $\vec{d} = \vec{d_i} - \vec{d_j}$
- Calculate the length of the difference

$$||\vec{d}||^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((\vec{d_i})_k - (\vec{d_j})_k)^2$$

$$= \sum_{k=1}^{n} (tf(w_k, D_i) - tf(w_k, D_j))^2$$

- For every word w_k we compare how often it appears in document D_i and document D_j

Calculate Euclidean distance between two document vectors d_i and d_j

- Take the difference $\vec{d} = \vec{d_i} - \vec{d_j}$
- Calculate the length of the difference

$$||\vec{d}||^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((\vec{d_i})_k - (\vec{d_j})_k)^2$$

$$= \sum_{k=1}^{n} (tf(w_k, D_i) - tf(w_k, D_j))^2$$

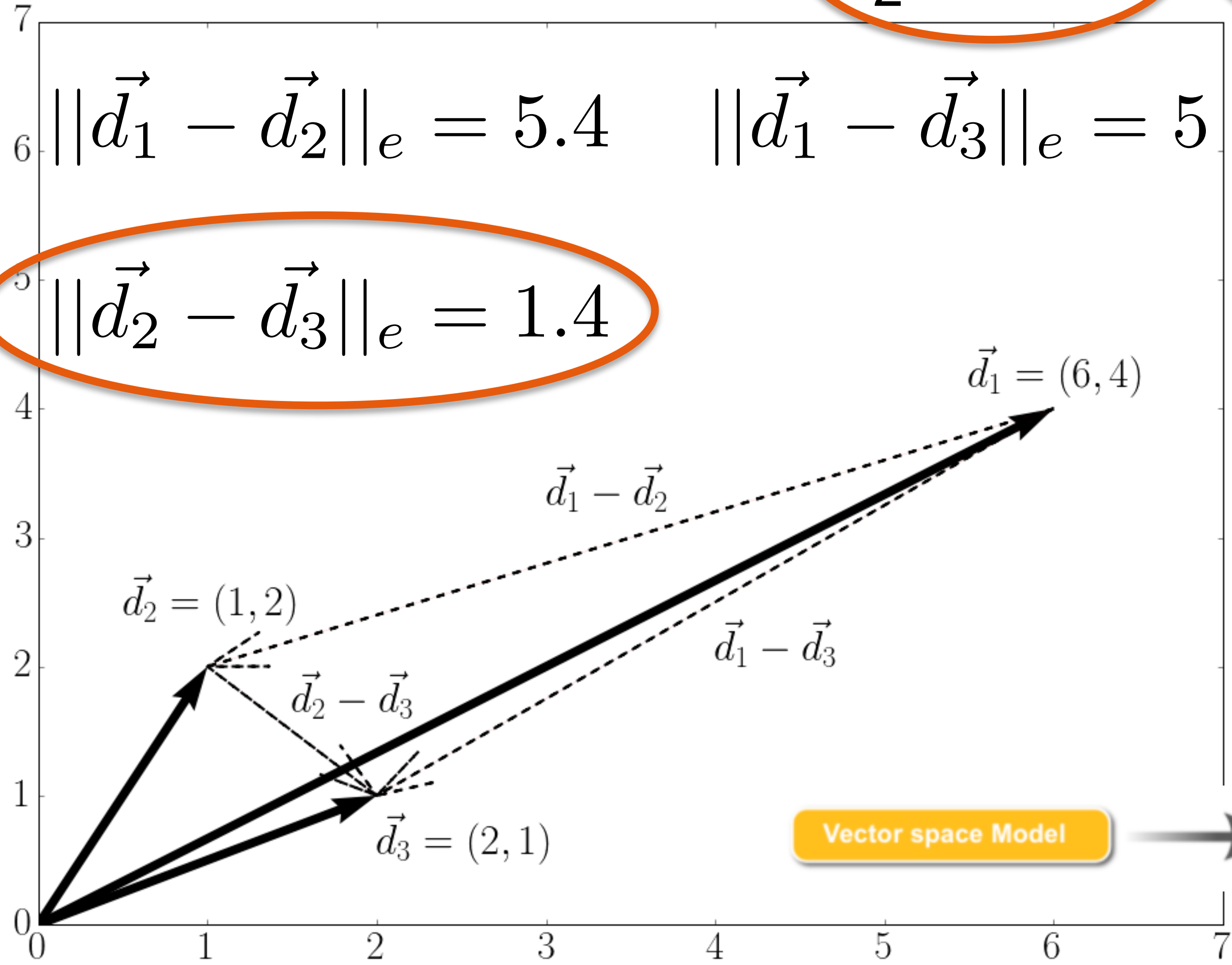
- For every word w_k we compare how often it appears in document D_i and document D_j

Euclidean distances for our example

- D_1= a a a b b a a b a b, D_2= a b b, D_3= a a b

$$||\vec{d_1} - \vec{d_2}||_e = 5.4 \quad ||\vec{d_1} - \vec{d_3}||_e = 5$$

$$||\vec{d_2} - \vec{d_3}||_e = 1.4$$



Vector space Model → Distance between vectors

Calculate Cosine distance between two document vectors d_i and d_j

- Calculate the scalar product $s = < \vec{d_i}, \vec{d_j} >$

$$< \vec{d_i}, \vec{d_j} > = \sum_{k=1}^{n} (\vec{d_i})_k (\vec{d_j})_k$$

$$= \sum_{k=1}^{n} tf(w_k, D_i) tf(w_k, D_j)$$

Remember: $(\vec{d_i})_k = tf(w_k, D_i)$
is zero most of the time.

Calculate Cosine distance between two document vectors d_i and d_j

- Calculate the scalar product $s =< \vec{d_i}, \vec{d_j} >$
- Divide it by the product of lengths of both vectors (length as in Euclidean distance)

$$\cos(\theta) = \frac{< \vec{d_i}, \vec{d_j} >}{||\vec{d_i}|| * ||\vec{d_j}||}$$

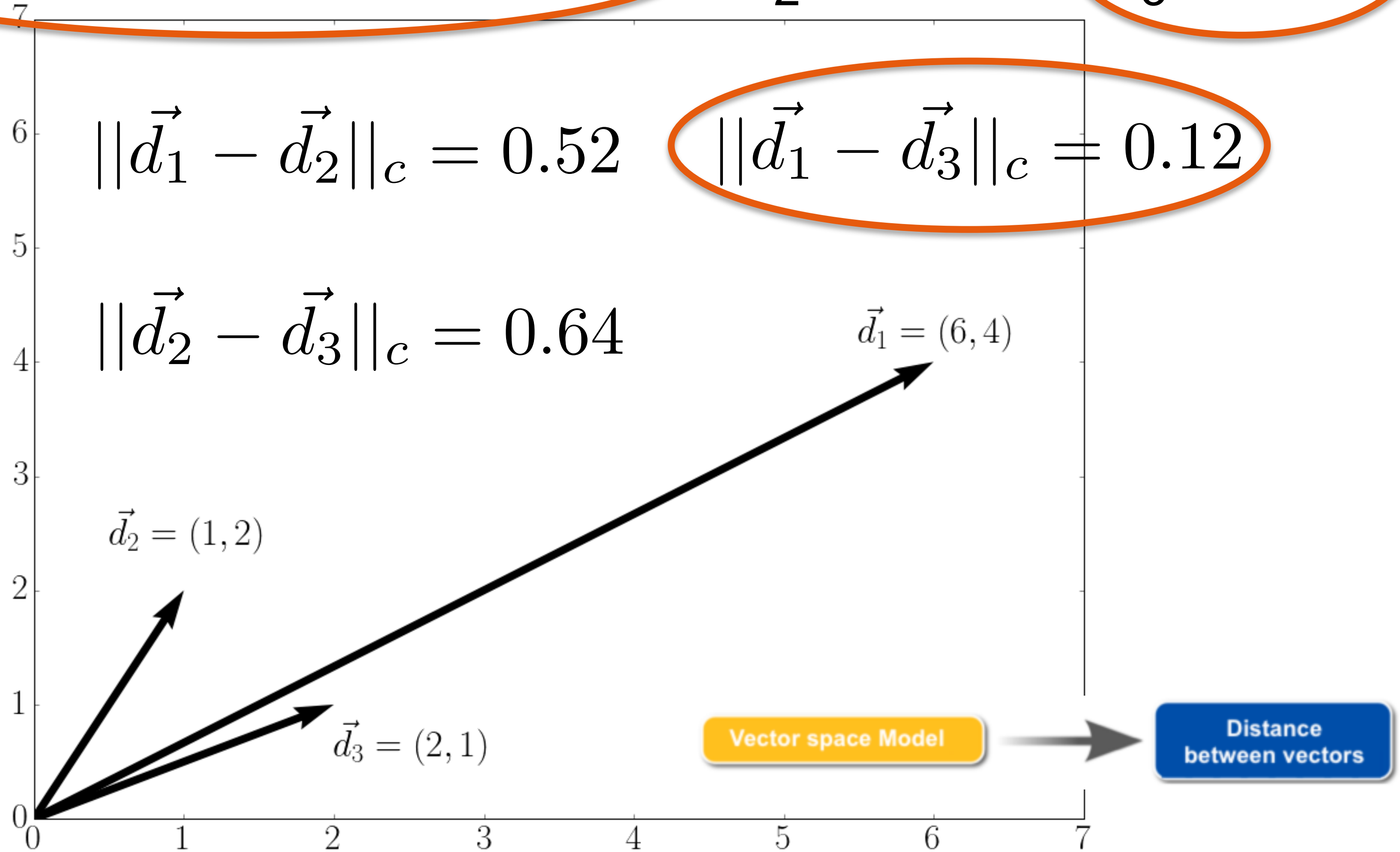
$$\Rightarrow \theta = \cos^{-1}\left(\frac{< \vec{d_i}, \vec{d_j} >}{||\vec{d_i}|| * ||\vec{d_j}||}\right)$$

Cosine distances for our example

- D_1= a a a b b a a b a b, D_2= a b b, D_3= a a b

$$||\vec{d_1} - \vec{d_2}||_c = 0.52 \qquad ||\vec{d_1} - \vec{d_3}||_c = 0.12$$

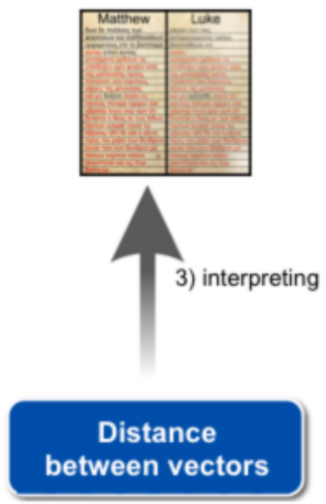
$$||\vec{d_2} - \vec{d_3}||_c = 0.64$$



Comparing cosine and Euclidean distance

- D_1= a a a b b a a b a b, D_2= a b b, D_3= a a b

	Euklid	Cosine
$\|\|\vec{d_1} - \vec{d_2}\|\|$	5.4	0.52
$\|\|\vec{d_2} - \vec{d_3}\|\|$	**1.4**	0.64
$\|\|\vec{d_1} - \vec{d_3}\|\|$	5	**0.12**



- Different choices of metric can yield very different results
- Choice of metric is part of the model!
- **Usually cosine distance is considered**

Thank you for your attention!

Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright: