Metody statystyczne w wyszukiwaniu informacji

Information Retrieval and Search

Mateusz Lango

Laboratory of Intelligent Decision Support Systems
Poznan University of Technology
mateusz.lango@cs.put.poznan.pl

Information Retrieval models

- Boolean model (BIR)
- vector model
- probabilistic model

Query likelihood model

assumption: uniform distribution

$$P(d|q) = \frac{P(q|d)\mathrm{P(d)}}{\mathrm{P(q)}} \propto P(q|d)$$

constant for a particular query

Query likelihood model

assumption: uniform distribution

$$P(d|q) = \frac{P(q|d)\mathrm{P(d)}}{\mathrm{P(q)}} \propto P(q|d)$$

constant for a particular query

Query likelihood model

assumption: uniform distribution

$$P(d|q) = \frac{P(q|d)\mathrm{P(d)}}{\cancel{\mathrm{P(q)}}} \propto P(q|d)$$

constant for a particular query

Query likelihood model

assumption: uniform distribution

$$P(d|q) = \frac{P(q|d)\cancel{\mathrm{P(d)}}}{\cancel{\mathrm{P(q)}}} \propto P(q|d)$$

constant for a particular query

Query likelihood model

assumption: uniform distribution

$$P(d|q) = \frac{P(q|d)\cancel{P(d)}}{\cancel{P(q)}} \propto P(q|d)$$

constant for a particular query

A user formulates a query based on an *imaginary relevant document*

Query likelihood model

assumption: uniform distribution

$$P(d|q) = \frac{P(q|d)\cancel{P(d)}}{\cancel{P(q)}} \propto P(q|d)$$

constant for a particular query

A user formulates a query based on an *imaginary relevant document*

How to model probability over *text*? $\Rightarrow$ Language model

How to compute $P(q|d)$?

$$P(q = \text{„Gdzie się napić dobrego alkoholu w Poznaniu?”} \\ |d = \text{„świetne miejsce szeroki asortyment piw ...”}) = ?$$

$$P(q = \text{„Gdzie się napić dobrego alkoholu w Poznaniu?”} \\ |d = \text{„dobrych przepisów na piwo domowe jest wiele...”}) = ?$$

Problem

Which of those probabilities should be higher?

How to compute $P(q|d)$?

$$P(q = \text{„Gdzie się napić dobrego alkoholu w Poznaniu?”} \\ | d = \text{„świetne miejsce szeroki asortyment piw ...”}) = ?$$

$$P(q = \text{„Gdzie się napić dobrego alkoholu w Poznaniu?”} \\ | d = \text{„dobrych przepisów na piwo domowe jest wiele...”}) = ?$$

Decomposition step (tokenization):

$$P(q|d) = P(q_1, q_2, q_3, ...|d) = ?$$

Statistical language model: unigram language model

assumption of independence

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2) \cdot P(q_3) \ldots = \prod_{i=1}^{N} P(q_i)$$

Example

Ala ma kota.
Jurek ma kota.
Złego kota ma Zbyszek.

$$P(q_i = \text{Ala}) = \frac{1}{10}$$

$$P(q_i = \text{kota}) = ?$$

$$P(q_i = \text{ma}) = ?$$

$$P(q_i = \text{Zbyszek}) = ?$$

Statistical language model: unigram language model

assumption of independence

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2) \cdot P(q_3) \ldots = \prod_{i=1}^{N} P(q_i)$$

Example

Ala ma kota.
Jurek ma kota.
Złego kota ma Zbyszek.

$$P(q_i = \text{Ala}) = \frac{1}{10}$$

$$P(q_i = \text{kota}) = \frac{3}{10}$$

$$P(q_i = \text{ma}) = ?$$

$$P(q_i = \text{Zbyszek}) = ?$$

Statistical language model: unigram language model

assumption of independence

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2) \cdot P(q_3) \ldots = \prod_{i=1}^{N} P(q_i)$$

Example

Ala ma kota.
Jurek ma kota.
Złego kota ma Zbyszek.

$$P(q_i = \text{Ala}) = \frac{1}{10}$$

$$P(q_i = \text{kota}) = \frac{3}{10}$$

$$P(q_i = \text{ma}) = \frac{3}{10}$$

$$P(q_i = \text{Zbyszek}) = ?$$

Statistical language model: unigram language model

assumption of independence

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2) \cdot P(q_3) \ldots = \prod_{i=1}^{N} P(q_i)$$

Example

Ala ma kota.
Jurek ma kota.
Złego kota ma Zbyszek.

$$P(q_i = \text{Ala}) = \frac{1}{10}$$

$$P(q_i = \text{kota}) = \frac{3}{10}$$

$$P(q_i = \text{ma}) = \frac{3}{10}$$

$$P(q_i = \text{Zbyszek}) = \frac{1}{10}$$

Unigram language model: example cont.

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2) \cdot P(q_3) \ldots = \prod_{i=1}^{N} P(q_i)$$

Example

$$P(q_i = \text{Zbyszek}) = \frac{1}{10} \quad P(q_i = \text{kota}) = \frac{3}{10} \quad P(q_i = \text{ma}) = \frac{3}{10}$$

$$\begin{aligned} P(q = [\text{Zbyszek, ma, kota}]) \\ &= P(q_1 = \text{Zbyszek}) P(q_2 = \text{ ma}) P(q_3 = \text{kota}) \\ &= \frac{1}{10} \cdot \frac{3}{10} \cdot \frac{3}{10} = 0.009 \end{aligned}$$

⇒ generalization to a completely new sentence!

⇒ a naive assumption...

Unigram language model: example cont.

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2) \cdot P(q_3) \ldots = \prod_{i=1}^{N} P(q_i)$$

Example

$$P(q_i = \text{Zbyszek}) = \frac{1}{10} \quad P(q_i = \text{kota}) = \frac{3}{10} \quad P(q_i = \text{ma}) = \frac{3}{10}$$

$$P(q = [\text{Zbyszek, ma, kota}])$$
$$= P(q_1 = \text{Zbyszek}) P(q_2 = \text{ma}) P(q_3 = \text{kota})$$
$$= \frac{1}{10} \cdot \frac{3}{10} \cdot \frac{3}{10} = 0.009$$

$\Rightarrow$ generalization to a completely new sentence!

$\Rightarrow$ a naive assumption...

Unigram language model: example cont.

$$P(q) = P(q_1, q_2, q_3, ...) = P(q_1) \cdot P(q_2) \cdot P(q_3)... = \prod_{i=1}^{N} P(q_i)$$

Example

$$P(q_i = \text{Zbyszek}) = \frac{1}{10} \quad P(q_i = \text{kota}) = \frac{3}{10} \quad P(q_i = \text{ma}) = \frac{3}{10}$$

$$\begin{aligned} & P(q = [\text{Zbyszek, ma, kota}]) \\ & \quad = P(q_1 = \text{Zbyszek}) P(q_2 = \text{ ma}) P(q_3 = \text{kota}) \\ & \quad = \frac{1}{10} \cdot \frac{3}{10} \cdot \frac{3}{10} = 0.009 \end{aligned}$$

$\Rightarrow$ generalization to a completely new sentence!

$\Rightarrow$ a naive assumption...

Bigram language model

Markov property

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2) \ldots = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

Example

Ala ma kota.
Jurek ma kota.
Złego kota ma Zbyszek.

$$P(q_i = \text{Ala}|\varnothing) = \frac{1}{3}$$

$$P(q_i = \text{kota}|\varnothing) = 0$$

$$P(q_i = \text{Ala}|\text{kota}) = ?$$

$$P(q_i = \text{ma}|\text{Ala}) = ?$$

Bigram language model

Markov property

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2) \ldots = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

Example

Ala ma kota.
Jurek ma kota.
Złego kota ma Zbyszek.

$$P(q_i = \text{Ala}|\varnothing) = \frac{1}{3}$$

$$P(q_i = \text{kota}|\varnothing) = 0$$

$$P(q_i = \text{Ala}|\text{kota}) = 0$$

$$P(q_i = \text{ma}|\text{Ala}) = ?$$

Bigram language model

Markov property

$$P(q) = P(q_1, q_2, q_3, ...) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2)... = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

Example

Ala ma kota.
Jurek ma kota.
Złego kota ma Zbyszek.

$$P(q_i = \text{Ala}|\varnothing) = \frac{1}{3}$$

$$P(q_i = \text{kota}|\varnothing) = 0$$

$$P(q_i = \text{Ala}|\text{kota}) = 0$$

$$P(q_i = \text{ma}|\text{Ala}) = 1$$

Bigram language model: example cont.

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2) \ldots = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

Example

$$\begin{aligned} P(q = [\text{Zbyszek}, \text{ma}, \text{kota}]) \\ &= P(\text{Zbyszek}|\varnothing)P(\text{ma}|\text{Zbyszek})P(\text{kota}|\text{ma}) \\ &= 0 \cdot 0 \cdot \frac{2}{3} = 0 \end{aligned}$$

Bigram language model: example cont.

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2) \ldots = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

Example

$$\begin{aligned} P(q = [\text{Zbyszek, ma, kota}]) \\ &= P(\text{Zbyszek}|\varnothing)P(\text{ma}|\text{Zbyszek})P(\text{kota}|\text{ma}) \\ &= 0 \cdot 0 \cdot \frac{2}{3} = 0 \end{aligned}$$

Bigram language model: example cont.

$$P(q) = P(q_1, q_2, q_3, \ldots) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2) \ldots = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

Example

$$\begin{aligned} P(q = [\text{Zbyszek, ma, kota}]) \\ &= P(\text{Zbyszek}|\varnothing)P(\text{ma}|\text{Zbyszek})P(\text{kota}|\text{ma}) \\ &= 0 \cdot 0 \cdot \frac{2}{3} = 0 \end{aligned}$$

Bigram language model: example analysis

- It seems that a smarter model is not working...

- $|V| = 6$
- $P(q_i) \rightarrow 6$ parameters
- $P(q_i|q_{i-1}) \rightarrow 36$ parameters
- $n = 10 \rightarrow ...$

- $|V| = 50\,000$
- $P(q_i) \rightarrow 50\,000$ parameters
- $P(q_i|q_{i-1}) \rightarrow 2\,500\,000\,000$ parameters
- $n = 10 \rightarrow ...$

Bigram language model: example analysis

- It seems that a smarter model is not working...

  - $|V| = 6$
  - $P(q_i) \rightarrow 6$ parameters
  - $P(q_i|q_{i-1}) \rightarrow 36$ parameters
  - $n = 10 \rightarrow \ldots$

  - $|V| = 50\,000$
  - $P(q_i) \rightarrow 50\,000$ parameters
  - $P(q_i|q_{i-1}) \rightarrow 2\,500\,000\,000$ parameters
  - $n = 10 \rightarrow \ldots$

TF-IDF to rescue?

Świetne miejsce, szeroki asortyment piw, świetna atmosfera.
Dobrych przepisów na piwo domowe jest wiele.

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	0	0	0	0	0	0	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	0	0	0	0	0	0	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	1	0	0	0	0	0	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	1	0	0	0	0	1	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	1	0	0	0	0	1	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	1	0	0	0	0	1	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	1	0	1	0	0	1	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	2	0	1	0	0	1	
doc2	0	0	0	0	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	2	0	1	0	0	1	
doc2	0	0	1	1	1	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	2	0	1	0	0	1	
doc2	0	0	1	1	1	0	

gdzie się napić dobry alkohol w poznań

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
query	0	1	0	1	0	0	

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	2	0	1	0	0	1	
doc2	0	0	1	1	1	0	

gdzie się napić dobry alkohol w poznań

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
query	0	1	0	1	0	0	

$$\cos(q, d) = \frac{d^T q}{||d|| ||q||}$$

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	2	0	1	0	0	1	
doc2	0	0	1	1	1	0	

gdzie się napić dobry alkohol w poznań

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
query	0	1	0	1	0	0	

$$\cos(q, d) = \frac{d^T q}{||d||||q||} \qquad \cos(q, d_1) = 0 \qquad \cos(q, d_2) = 0.41$$

TF-IDF to rescue?

świetny miejsce szeroki asortyment piwo świetny atmosfera
dobry przepis na piwo domowy jest wiele

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
doc1	2	0	1	0	0	1	
doc2	0	0	1	1	1	0	

gdzie się napić dobry alkohol w poznań

	świetny	alkohol	piwo	dobry	przepis	miejsce	...
query	0	1	0	1	0	0	

$$\cos(q, d) = \frac{d^T q}{||d||||q||} \qquad \cos(q, d_1) = 0 \qquad \cos(q, d_2) = 0.41$$

TF-IDF $\approx$ Query Likelihood model with smoothed LM

Motivation / agenda

Problem

How can we discover that „piwo” is similar to „alkohol” in a fully automatic way?

1. Query likelihood model
2. Distributional hypothesis
3. Latent Semantic Analysis
4. Word embeddings

Motivation / agenda

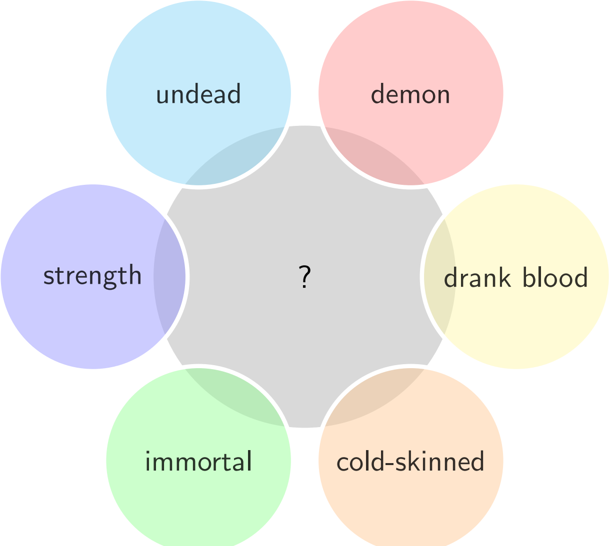
Problem

How can we discover that „piwo” is similar to „alkohol” in a fully automatic way?

1. Query likelihood model
2. Distributional hypothesis
3. Latent Semantic Analysis
4. Word embeddings

Who is Edward?

Who is Edward?



Distributional hypothesis

You shall know a word by the company it keeps

J. R. Firth, 1957.

Distributional hypothesis

You shall know a word by the company it keeps
J. R. Firth, 1957.

Example

czarny ? miałknął
ukochany ? pił mleko

Distributional hypothesis

You shall know a word by the company it keeps
J. R. Firth, 1957.

Example

czarny **kotek** miałknął
ukochany **kotek** pił mleko

Distributional hypothesis

You shall know a word by the company it keeps
J. R. Firth, 1957.

Example

czarny **kotek** miałknął
ukochany **kotek** pił mleko
wyleniały **kot** miałknął
mój czarny **kot** zamruczał z zadowoleniem

Let's start with easy cases

	świetny	dobry	...
doc1	5	5	...
doc2	0	0	...
doc3	15	15	...
doc4	0	0	...
doc5	0	0	...
doc6	0	0	...
doc7	1	1	...

Let's start with easy cases

	świetny	dobry	...
doc1	3	6	...
doc2	0	0	...
doc3	7	14	...
doc4	0	0	...
doc5	0	0	...
doc6	0	0	...
doc7	1	2	...

Let's start with easy cases

	świetny	dobry	...
doc1	3	6	...
doc2	0	0	...
doc3	7	14	...
doc4	0	0	...
doc5	0	0	...
doc6	0	0	...
doc7	1	2	...

$$Cor(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Let's start with easy cases

	świetny	dobry	...
doc1	3	6	...
doc2	0	0	...
doc3	7	14	...
doc4	0	0	...
doc5	0	0	...
doc6	0	0	...
doc7	1	2	...

$$Cor(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Assuming for all columns $\bar{x} = 0$, $s_x = 1$ and omitting a constant $n - 1$

$$Cor(x, y) = \sum_{i=1}^{N} x_i y_i = x^T y$$

Let's start with easy cases

	świetny	dobry	...
doc1	3	6	...
doc2	0	0	...
doc3	7	14	...
doc4	0	0	...
doc5	0	0	...
doc6	0	0	...
doc7	1	2	...

$$Cor(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Assuming for all columns $\bar{x} = 0$, $s_x = 1$ and omitting a constant $n - 1$

$$Cor(x, y) = \sum_{i=1}^{N} x_i y_i = x^T y$$

$$Cor = X^T X$$

Let's start with easy cases

	świetny	dobry	...
świetny	1	0.99	...
dobry	0.99	1	...
...	..	..	...

$$Cor(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Assuming for all columns $\bar{x} = 0$, $s_x = 1$ and omitting a constant $n - 1$

$$Cor(x, y) = \sum_{i=1}^{N} x_i y_i = x^T y$$

$$Cor = X^T X$$

Correlation matrix: example from Psychology[1]

	batting	crossw.	darts	Scrabble	juggling	spelling
batting		0,00	0,91	-0,05	0,96	0,10
crossw.			0,08	0,88	0,02	0,80
darts				-0,01	0,90	0,29
Scrabble					-0,08	0,79
juggling						0,11
spelling						

[1] D. Howitt, D. Cramer: Introduction to Statistics in Psychology, 2005

Correlation matrix: example from Psychology[1]

	batting	crossw.	darts	Scrabble	juggling	spelling
batting		0,00	0,91	-0,05	0,96	0,10
crossw.			0,08	0,88	0,02	0,80
darts				-0,01	0,90	0,29
Scrabble					-0,08	0,79
juggling						0,11
spelling						

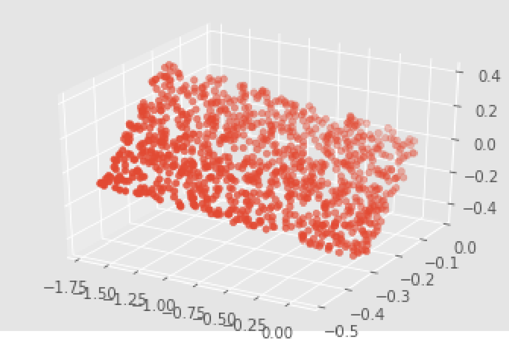
Problem

How to discover the latent variables in data?

Disclaimer: the following part presents oversimplified math

[1] D. Howitt, D. Cramer: Introduction to Statistics in Psychology, 2005

How to discover the *latent* variables in data?



Problem

How the plot of two correlated variables looks like?

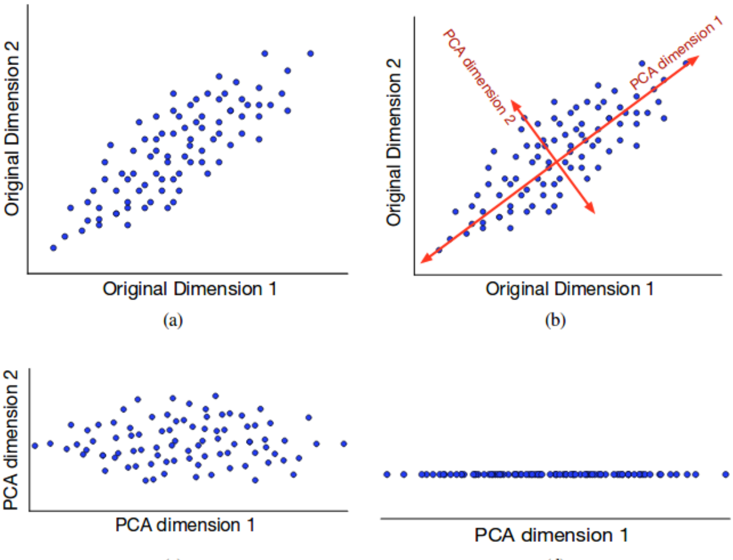
Problem

How to construct the variable representing a new concept?

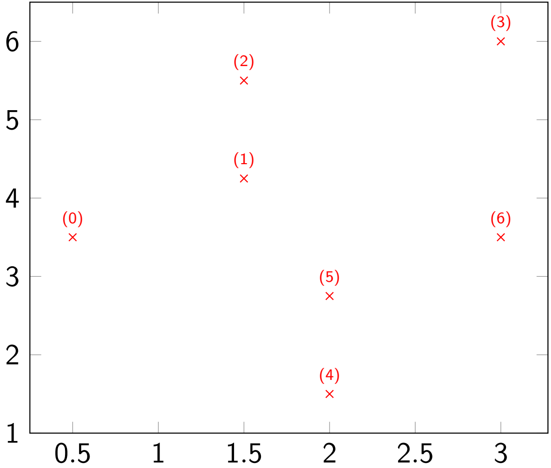
Problem

Assuming two similar words and one which is not similar (i.e. 3rd variable is not correlated) - how the plot looks like?

Latent semantic analysis: main idea

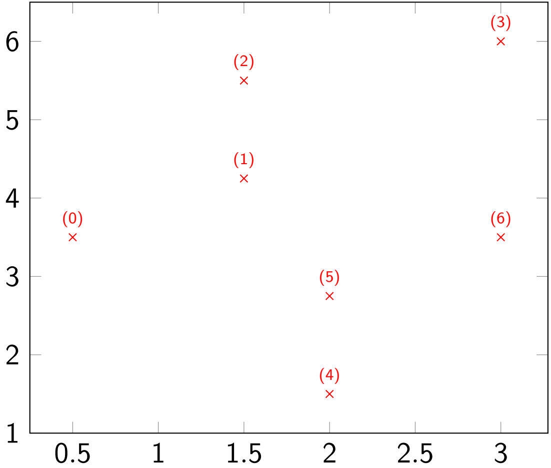


Projecting data onto a line



$$u = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Projecting data onto a line



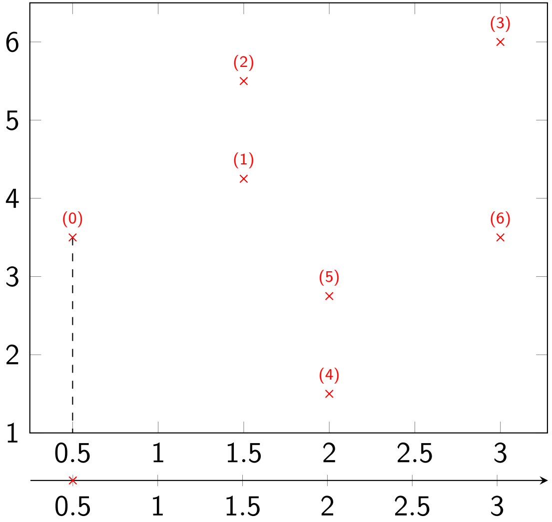
$$u = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_0 = \begin{bmatrix} 0.5 \\ 3.5 \end{bmatrix}$$

$$x_0^T u = \begin{bmatrix} 0.5 & 3.5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_0^T u = 0.5$$

Projecting data onto a line



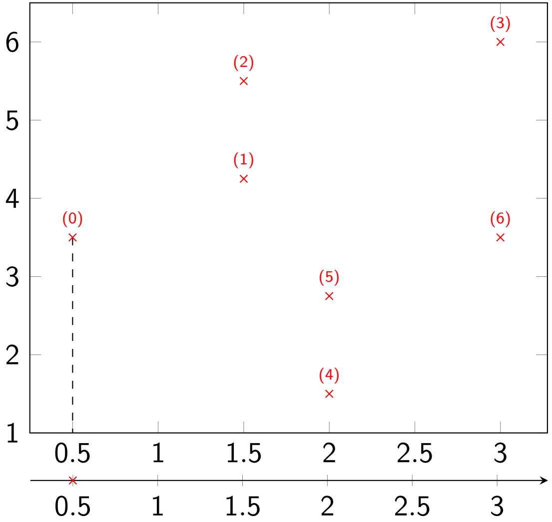
$$u = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_0 = \begin{bmatrix} 0.5 \\ 3.5 \end{bmatrix}$$

$$x_0^T u = \begin{bmatrix} 0.5 & 3.5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_0^T u = 0.5$$

Projecting data onto a line



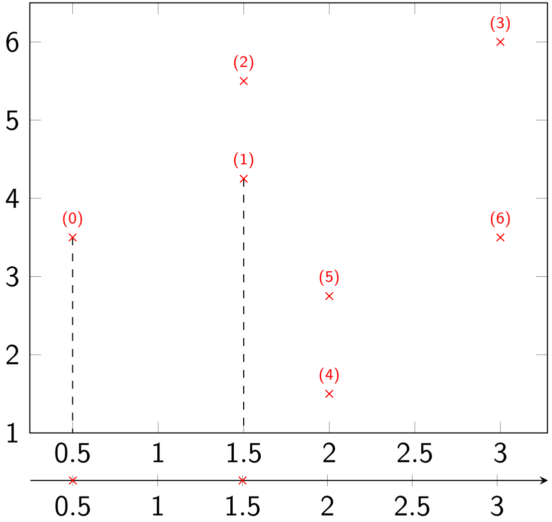
$$u = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_1^T u = \begin{bmatrix} 1.5 & 4.25 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_1^T u = 1.5$$

Projecting data onto a line



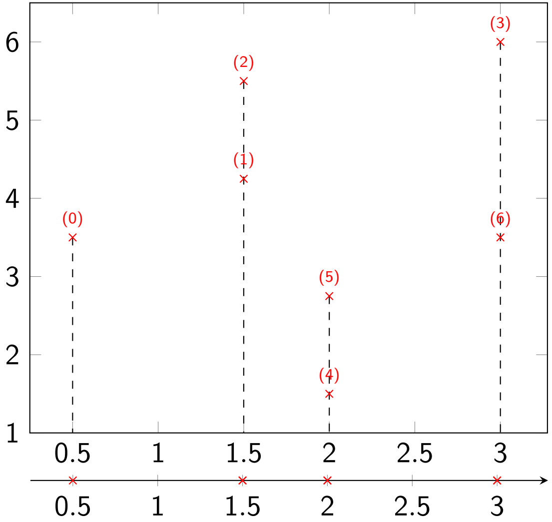
$$u = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_1^T u = \begin{bmatrix} 1.5 & 4.25 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_1^T u = 1.5$$

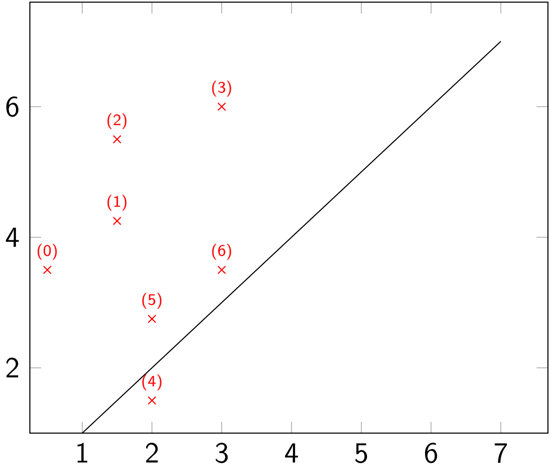
Projecting data onto a line



$$u = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

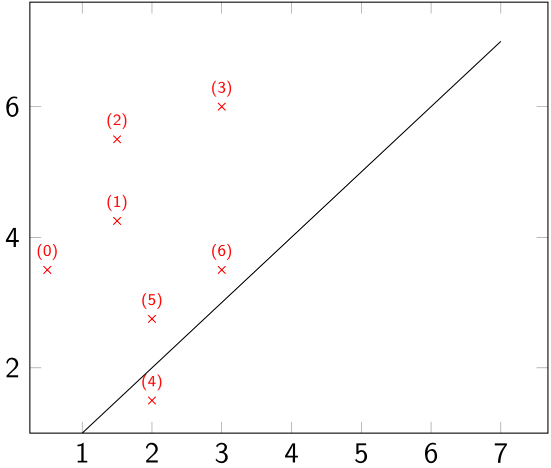
...

Projecting data onto a line



$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Projecting data onto a line



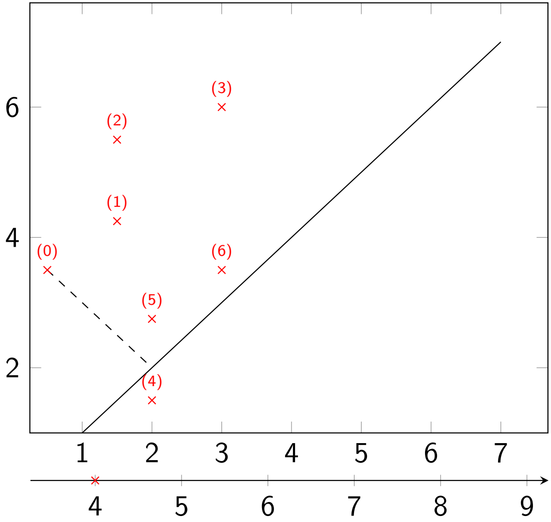
$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_0 = \begin{bmatrix} 0.5 \\ 3.5 \end{bmatrix}$$

$$x_0^T u = \begin{bmatrix} 0.5 & 3.5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_0^T u = 4$$

Projecting data onto a line



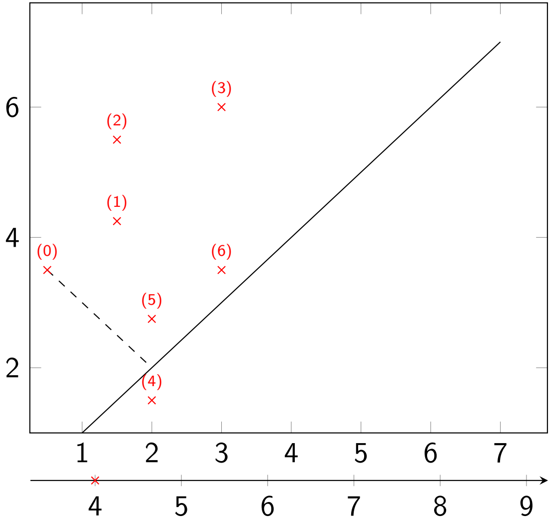
$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_0 = \begin{bmatrix} 0.5 \\ 3.5 \end{bmatrix}$$

$$x_0^T u = \begin{bmatrix} 0.5 & 3.5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_0^T u = 4$$

Projecting data onto a line



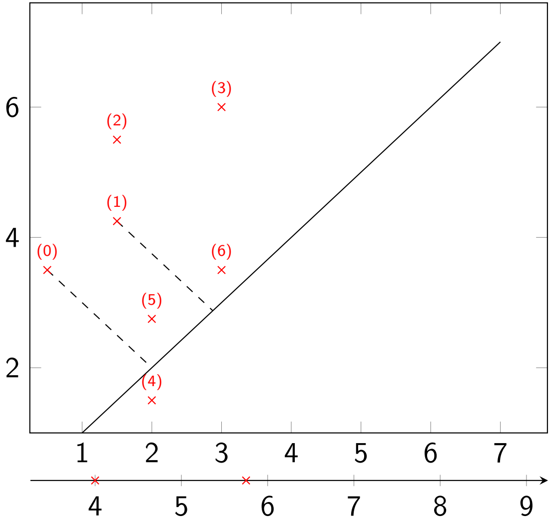
$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_1^T u = \begin{bmatrix} 1.5 & 4.25 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_1^T u = 5.75$$

Projecting data onto a line



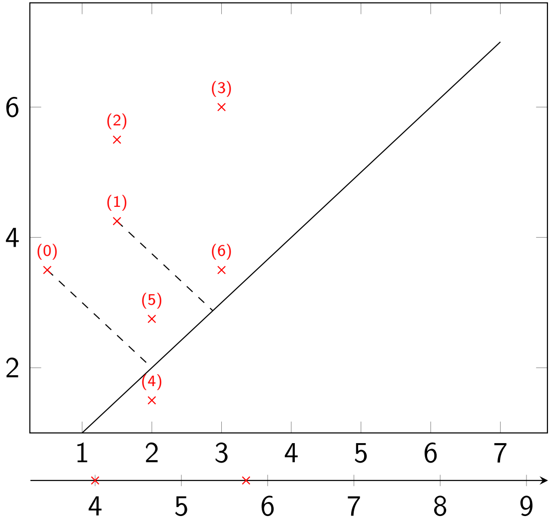
$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_1^T u = \begin{bmatrix} 1.5 & 4.25 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_1^T u = 5.75$$

Projecting data onto a line



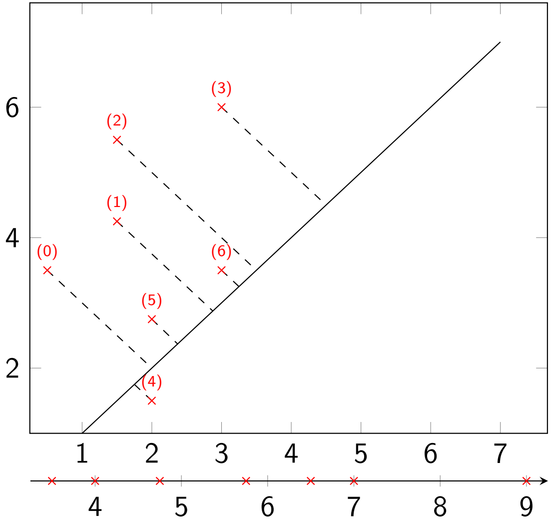
$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_2^T u = \begin{bmatrix} 1.5 & 5.5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_2^T u = ?$$

Projecting data onto a line



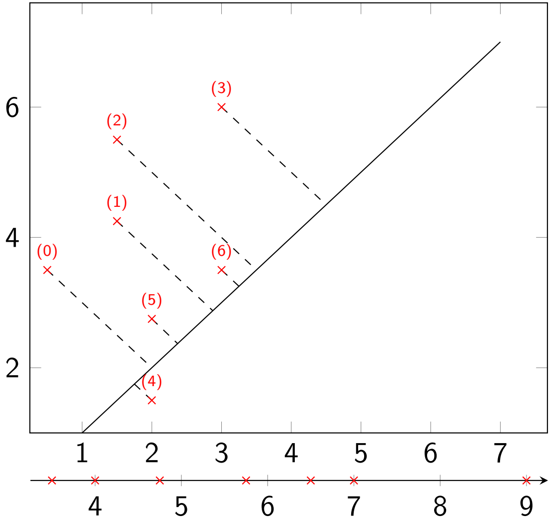
$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_2^T u = \begin{bmatrix} 1.5 & 5.5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x_2^T u = 7$$

Projecting data onto a line

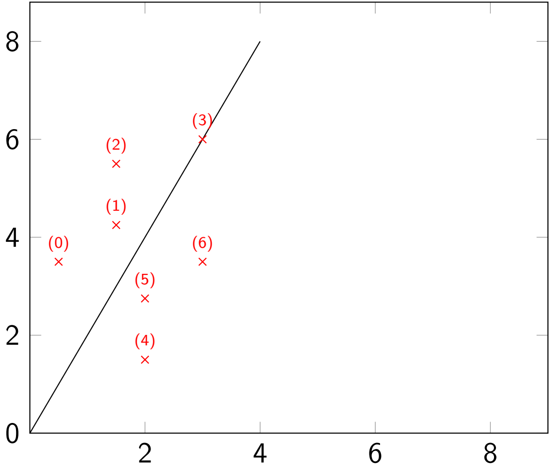


$$u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Problem

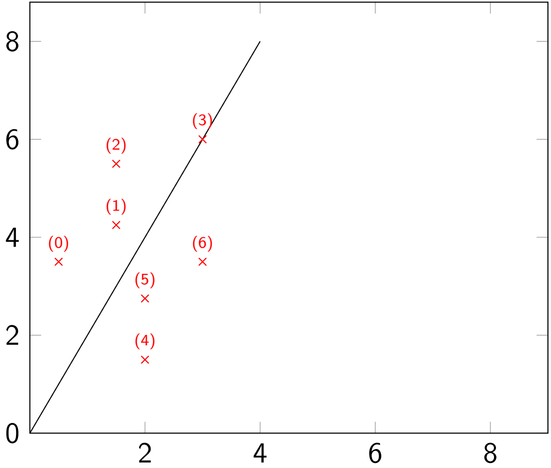
Is it simple linear regression?

Projecting data onto a line



$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Projecting data onto a line



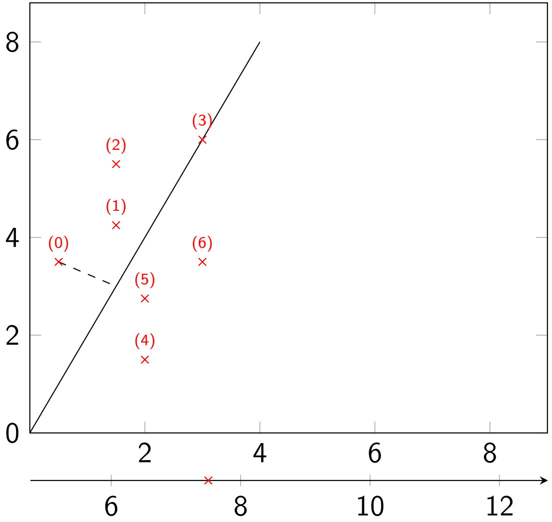
$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_0 = \begin{bmatrix} 0.5 \\ 3.5 \end{bmatrix}$$

$$x_0^T u = \begin{bmatrix} 0.5 & 3.5 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_0^T u = ?$$

Projecting data onto a line



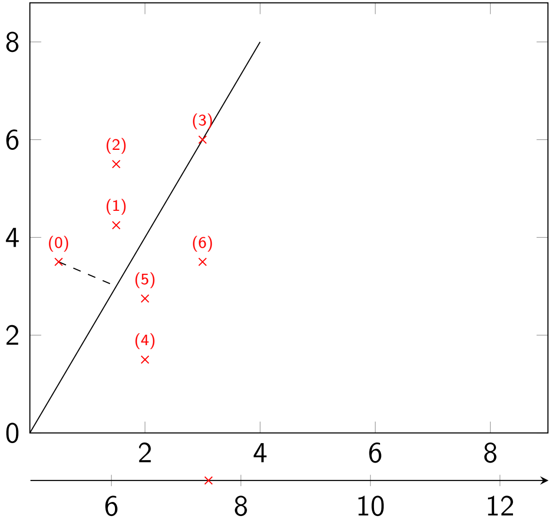
$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_0 = \begin{bmatrix} 0.5 \\ 3.5 \end{bmatrix}$$

$$x_0^T u = \begin{bmatrix} 0.5 & 3.5 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_0^T u = 7.5$$

Projecting data onto a line



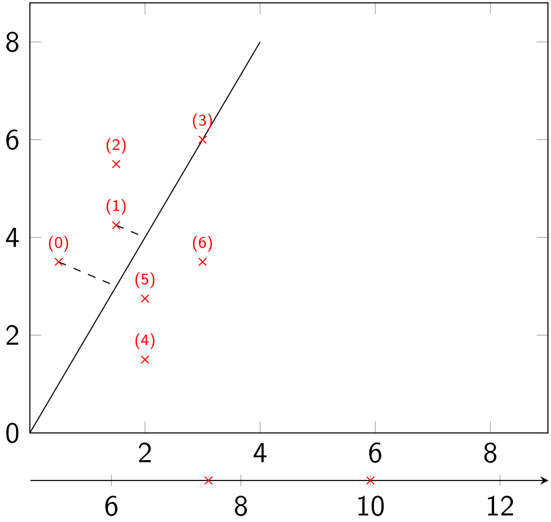
$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_1^T u = \begin{bmatrix} 1.5 & 4.25 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_1^T u = ?$$

Projecting data onto a line



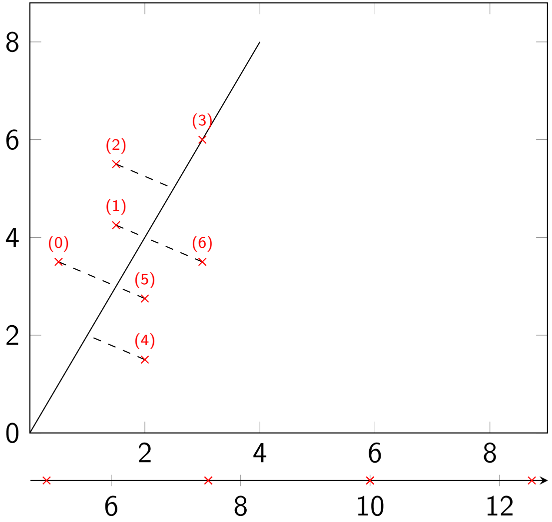
$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

$$x_1^T u = \begin{bmatrix} 1.5 & 4.25 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_1^T u = 10$$

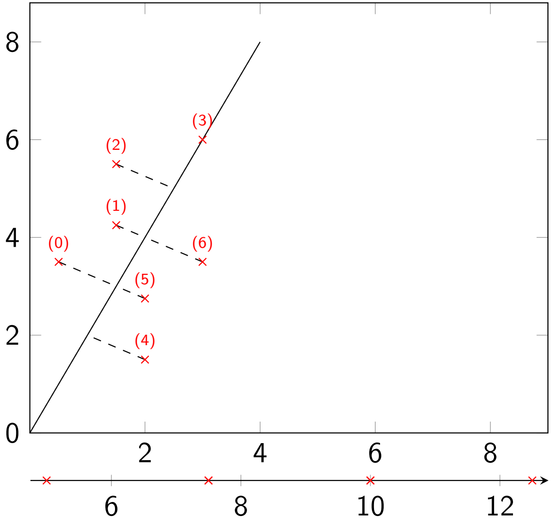
Projecting data onto a line



$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_{proj} = Xu$$

Projecting data onto a line



$$u = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x_{proj} = Xu$$

This works also in more than 2 dimensions

Looking for „best" dimension

Goal: find direction u of a line which used in projection retains maximum amount of information

Looking for „best” dimension

Goal: find direction u of a line which used in projection retains maximum amount of information

variance of data!

Looking for „best" dimension

Goal: find direction u of a line which used in projection retains maximum amount of information

variance of data!

TF-IDF matrix

$$X$$

Variance/covariance matrix

$$X^T X$$

Looking for „best” dimension

Goal: find direction u of a line which used in projection retains maximum amount of information

variance of data!

TF-IDF matrix

$$X$$

Variance/covariance matrix

$$X^T X$$

Projected TF-IDF matrix onto a line

$$x_{proj} = Xu$$

Looking for „best" dimension

Goal: find direction u of a line which used in projection retains maximum amount of information

variance of data!

TF-IDF matrix

$$X$$

Variance/covariance matrix

$$X^T X$$

Projected TF-IDF matrix onto a line

$$x_{proj} = Xu$$

Variance of projected data

$$x_{proj}^T x_{proj} = (Xu)^T (Xu) = u^T X^T X u$$

Looking for „best” dimension

Goal: find direction u of a line which used in projection retains maximum amount of information

variance of data!

TF-IDF matrix

$$X$$

Variance/covariance matrix

$$X^T X$$

Projected TF-IDF matrix onto a line

covariance matrix of the orginal data

Variance of projected data

$$x_{proj}^T x_{proj} = (Xu)^T (Xu) = u^T X^T X u$$

Maximizing variance

Goal: find direction u of a line which used in projection retains maximum amount of information

$$\max_u u^T X^T X u$$

Maximizing variance

Goal: find direction u of a line which used in projection retains maximum amount of information

$$\max_u u^T X^T X u \qquad \text{s.t.} \qquad ||u|| = u^T u = 1$$

Maximizing variance

Goal: find direction u of a line which used in projection retains maximum amount of information

$$\max_u u^T X^T X u \qquad \text{s.t.} \qquad ||u|| = u^T u = 1$$

$$L(u) = u^T X^T X u - \underbrace{\lambda(u^T u - 1)}_{\text{constraint}}$$

Maximizing variance

Goal: find direction u of a line which used in projection retains maximum amount of information

$$\max_u u^T X^T X u \qquad \text{s.t.} \qquad ||u|| = u^T u = 1$$

$$L(u) = u^T X^T X u - \underbrace{\lambda(u^T u - 1)}_{\text{constraint}}$$

$$\nabla L = 2X^T X u - 2\lambda u = 0$$

Maximizing variance

Goal: find direction u of a line which used in projection retains maximum amount of information

$$\max_u u^T X^T X u \qquad \text{s.t.} \qquad ||u|| = u^T u = 1$$

$$L(u) = u^T X^T X u - \underbrace{\lambda(u^T u - 1)}_{\text{constraint}}$$

$$\nabla L = 2X^T X u - 2\lambda u = 0$$

$$X^T X u = \lambda u$$

Maximizing variance

Goal: find direction u of a line which used in projection retains maximum amount of information

$$\max_u u^T X^T X u \qquad \text{s.t.} \qquad ||u|| = u^T u = 1$$

$$L(u) = u^T X^T X u - \underbrace{\lambda(u^T u - 1)}_{\text{constraint}}$$

$$\nabla L = 2X^T X u - 2\lambda u = 0$$

$$X^T X u = \lambda u$$

Conclusion: u is an eigenvector of $X^T X$ with eigenvalue λ

Our example: How this „best dimension” looks like?

Variable	Eigenvector
Skill at batting	0.348
Skill at crosswords	0.003
Skill at darts	0.334
Skill at Scrabble	-0.024
Skill at juggling	0.344
Skill at spelling	0.053

Our example: How this „best dimension" looks like?

supervariable

Variable	Eigenvector
Skill at batting	0.348
Skill at crosswords	0.003
Skill at darts	0.334
Skill at Scrabble	-0.024
Skill at juggling	0.344
Skill at spelling	0.053

Our example: How this „best dimension” looks like?

supervariable

Variable	Eigenvector
Skill at batting	0.348
Skill at crosswords	0.003
Skill at darts	0.334
Skill at Scrabble	-0.024
Skill at juggling	0.344
Skill at spelling	0.053

Problem

What would the second best direction look like?

supervariable

Our example: How this „best dimension" looks like?

Variable	Eigenvector
Skill at batting	0.348
Skill at crosswords	0.003
Skill at darts	0.334
Skill at Scrabble	-0.024
Skill at juggling	0.344
Skill at spelling	0.053

Intuitively, we can compress information about students into one value!

Problem

How to calculate the „supervariable" for a student with scores:

$$x = [8, 4, 6, 3, 7, 10]$$

supervariable

Our example: How this „best dimension" looks like?

Variable	Eigenvector
Skill at batting	0.348
Skill at crosswords	0.003
Skill at darts	0.334
Skill at Scrabble	-0.024
Skill at juggling	0.344
Skill at spelling	0.053

Intuitively, we can compress information about students into one value!

Problem

How to calculate the „supervariable" for a student with scores:

$$x = [8, 4, 6, 3, 7, 10]$$

$$x^T u = 7.60$$

Our example: How this „best dimension" looks like?

supervariable

Variable	Eigenvector
Skill at batting	0.348
Skill at crosswords	0.003
Skill at darts	0.334
Skill at Scrabble	-0.024
Skill at juggling	0.344
Skill at spelling	0.053

Intuitively, we can compress information about students into one value!

Problem

How to calculate the „supervariable" for a student with scores:

$$x = [8, 4, 6, 3, 7, 10]$$

$$x^T u = 7.60$$

Principal Component Analysis (PCA)

1. Normalize data $x_{ij} = \frac{x_{ij} - \mu_j}{s_j}$
2. Calculate covariance matrix $X^T X$
3. Pick the eigenvector u with the highest eigenvalue
4. Project data onto a line Xu

Principal Component Analysis (PCA)

1. Normalize data $x_{ij} = \frac{x_{ij} - \mu_j}{s_j}$
2. Calculate covariance matrix $X^T X$
3. Pick the eigenvector u with the highest eigenvalue
4. Project data onto a line Xu

Problem

How to project data into several dimensions?

Principal Component Analysis (PCA)

1. Normalize data $x_{ij} = \frac{x_{ij} - \mu_j}{s_j}$
2. Calculate covariance matrix $X^T X$
3. Pick several eigenvectors $u_1, u_2, u_3, \ldots$ with highest eigenvalues
4. Project data onto a hyperplane $XU = [Xu_1, Xu_2, Xu_3, \ldots]$

PCA: What about TF-IDF matrix?

d1 : Romeo and Juliet.
d2 : Juliet: O happy dagger!
d3 : Romeo died by dagger.
d4 : "Live free or die", that's the New-Hampshire's motto.
d5 : Did you know, New-Hampshire is in New-England.

	Eig 1	Eig 2
romeo	-0.396	0.280
juliet	-0.314	0.450
happy	-0.178	0.269
dagger	-0.438	0.369
live	-0.264	-0.346
die	-0.524	-0.246
free	-0.264	-0.346
new-hampshire	-0.326	-0.460

Example from Latent Semantic Analysis Tutorial by Alex Thomo

PCA: What about TF-IDF matrix?

d1 : Romeo and Juliet.
d2 : Juliet: O happy dagger!
d3 : Romeo died by dagger.
d4 : "Live free or die", that's the New-Hampshire's motto.
d5 : Did you know, New-Hampshire is in New-England.

	Eig 1	Eig 2
romeo	-0.396	0.280
juliet	-0.314	0.450
happy	-0.178	0.269
dagger	-0.438	0.369
live	-0.264	-0.346
die	-0.524	-0.246
free	-0.264	-0.346
new-hampshire	-0.326	-0.460

Example from Latent Semantic Analysis Tutorial by Alex Thomo

Query „die dagger"

	TF	LSA
d1	0,00	0,77
d2	0,40	0,73
d3	0,81	0,98
d4	0,35	0,61
d5	0,00	0,48

Latent Semantic Analysis

Application of PCA to a term-document matrix is called:

Latent Semantic Analysis

PCA Example - AT&T Facedatabase

Figure from „Compressing arrays of classifiers using Volterra-neural network: Application to face recognition" by M. Rubiolo, G. Stegmayer and D. Milone

PCA Example - Eigenfaces

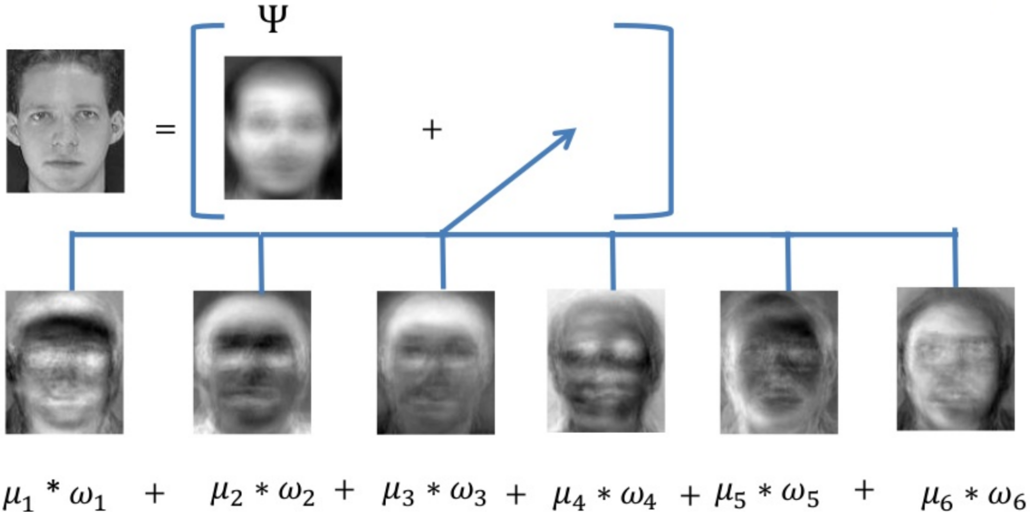
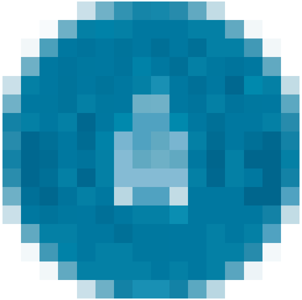


Figure from „EigenFaces For Recognition" by Semih Korkmaz

PCA Example - Eigenfaces

- first image – reconstruction from 10 eigenvectors
- second image – 25 eigenvectors
- third image – 40 eigenvectors
- last image – 300 eigenvectors



Practice: How to get eigenvectors?

A large variety of algorithms: Power Iteration, QR algorithm ...
However, in practice often procedures for SVD decomposition are used.

$$X^T X \overset{SVD}{=} V \Sigma V^T$$

SVD: `numpy.linalg.svd` (Python), `svd()` (R), `svd()` (octave)

Practice: How to get eigenvectors?

A large variety of algorithms: Power Iteration, QR algorithm ...
However, in practice often procedures for SVD decomposition are used.

$$X^T X \overset{SVD}{=} V \Sigma V^T$$

matrix with eigenvectors in columns

SVD: `numpy.linalg.svd` (Python), `svd()` (R), `svd()` (octave)

Practice: How to get eigenvectors?

A large variety of algorithms: Power Iteration, QR algorithm ...
However, in practice often procedures for SVD decomposition are used.

$$X^T X \overset{SVD}{=} V \Sigma V^T$$

matrix with eigenvectors in columns

diagonal matrix with corresponding eigenvalues

SVD: `numpy.linalg.svd` (Python), `svd()` (R), `svd()` (octave)

Singular-value decomposition (SVD)

$$X^T X \stackrel{SVD}{=} V \Sigma V^T$$

$$X^T X_{d\times d} = \begin{bmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{bmatrix}_{d\times d} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}_{d\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{bmatrix}_{d\times d}$$

Singular-value decomposition (SVD)

exact reconstrution

$$X^T X \stackrel{SVD}{=} V \Sigma V^T$$

$$X^T X_{d\times d} = \begin{bmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{bmatrix}_{d\times d} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}_{d\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{bmatrix}_{d\times d}$$

SVD decomposition provides also nice view on the data compression/dimensionality reduction process

Singular-value decomposition (SVD)

approximate reconstrution

$$X^TX \stackrel{SVD}{=} V\Sigma V^T$$

$$X^TX_{d\times d} \approx \begin{bmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{bmatrix}_{d\times d} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{d\times d} \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \\ - & v_3 & - \end{bmatrix}_{d\times d}$$

SVD decomposition provides also nice view on the data compression/dimensionality reduction process

Singular-value decomposition (SVD)

approximate reconstrution

$$X^TX \overset{SVD}{=} V\Sigma V^T$$

$$X^TX_{d\times d} \approx \begin{bmatrix} | & | & 0 \\ v_1 & v_2 & 0 \\ | & | & 0 \end{bmatrix}_{d\times d} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{d\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{bmatrix}_{d\times d}$$

SVD decomposition provides also nice view on the data compression/dimensionality reduction process

Singular-value decomposition (SVD)

approximate reconstrution

$$X^TX \overset{SVD}{=} V\Sigma V^T$$

$$X^TX_{d\times d} \approx \begin{bmatrix} | & | & 0 \\ v_1 & v_2 & 0 \\ | & | & 0 \end{bmatrix}_{d\times d} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{d\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ 0 & 0 & 0 \end{bmatrix}_{d\times d}$$

SVD decomposition provides also nice view on the data compression/dimensionality reduction process

SVD: example

$$X^TX = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{bmatrix}$$

$$X^TX \overset{SVD}{=} V\Sigma V^T$$

$$\begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{bmatrix} = \begin{bmatrix} -0.6 & 0.5 & 0.6 \\ -0.7 & 0.0 & -0.7 \\ -0.3 & -0.8 & 0.3 \end{bmatrix} \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} -0.6 & -0.7 & -0.3 \\ 0.5 & 0.0 & -0.8 \\ 0.6 & -0.7 & 0.3 \end{bmatrix}$$

SVD: example

$$X^TX = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{bmatrix}$$

$$X^TX \stackrel{SVD}{=} V\Sigma V^T$$

$$\begin{bmatrix} 0.84 & 0.67 & -0.09 \\ 0.67 & 0.79 & 0.40 \\ 0 & 0.40 & 0.94 \end{bmatrix} \approx \begin{bmatrix} -0.6 & 0.5 & 0.6 \\ -0.7 & 0.0 & -0.7 \\ -0.3 & -0.8 & 0.3 \end{bmatrix} \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & {\color{red}0} \end{bmatrix} \begin{bmatrix} -0.6 & -0.7 & -0.3 \\ 0.5 & 0.0 & -0.8 \\ 0.6 & -0.7 & 0.3 \end{bmatrix}$$

SVD: example

$$X^T X = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{bmatrix}$$

$$X^T X \overset{SVD}{=} V \Sigma V^T$$

$$\begin{bmatrix} 0.58 & 0.67 & 0.34 \\ 0.67 & 0.79 & 0.40 \\ 0.34 & 0.40 & 0.20 \end{bmatrix} \approx \begin{bmatrix} -0.6 & 0.5 & 0.6 \\ -0.7 & 0.0 & -0.7 \\ -0.3 & -0.8 & 0.3 \end{bmatrix} \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & \color{red}{0} & 0 \\ 0 & 0 & \color{red}{0} \end{bmatrix} \begin{bmatrix} -0.6 & -0.7 & -0.3 \\ 0.5 & 0.0 & -0.8 \\ 0.6 & -0.7 & 0.3 \end{bmatrix}$$

SVD: example

$$X^T X = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.3 \\ 0 & 0.3 & 1 \end{bmatrix}$$

$$X^T X \overset{SVD}{=} V \Sigma V^T$$

$$\begin{bmatrix} 0.58 & 0.67 & 0.34 \\ 0.67 & 0.79 & 0.40 \\ 0.34 & 0.40 & 0.20 \end{bmatrix} \approx \begin{bmatrix} -0.6 & 0.5 & 0.6 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -0.6 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0.6 & 0 & 0 \end{bmatrix}$$

SVD: example 2

$$X^T X = \begin{bmatrix} 35.9 & 28.5 & 64.4 \\ 28.5 & 40.5 & 68.9 \\ 64.4 & 68.9 & 133.3 \end{bmatrix}$$

$$X^T X \overset{SVD}{=} V \Sigma V^T$$

$$\begin{bmatrix} 35.9 & 28.5 & 64.4 \\ 28.5 & 40.5 & 68.9 \\ 64.4 & 68.9 & 133.3 \end{bmatrix} = \begin{bmatrix} -0.39 & -0.42 & -0.81 \\ 0.71 & -0.59 & 0.01 \\ -0.57 & -0.57 & 0.57 \end{bmatrix} \begin{bmatrix} 200 & 0 & 0 \\ 0 & 9.6 & 0 \\ 0 & 0 & 10^{-14} \end{bmatrix} \begin{bmatrix} -0.39 & 0.71 & -0.57 \\ -0.42 & -0.59 & -0.57 \\ -0.81 & 0.01 & 0.57 \end{bmatrix}$$

SVD: example 2

$$X^T X = \begin{bmatrix} 35.9 & 28.5 & 64.4 \\ 28.5 & 40.5 & 68.9 \\ 64.4 & 68.9 & 133.3 \end{bmatrix}$$

$$X^T X \stackrel{SVD}{=} V \Sigma V^T$$

$$\begin{bmatrix} 35.9 & 28.5 & 64.4 \\ 28.5 & 40.5 & 68.9 \\ 64.4 & 68.9 & 133.3 \end{bmatrix} \approx \begin{bmatrix} -0.39 & -0.42 & 0 \\ 0.71 & -0.59 & 0 \\ -0.57 & -0.57 & 0 \end{bmatrix} \begin{bmatrix} 200 & 0 & 0 \\ 0 & 9.6 & 0 \\ 0 & 0 & {\color{red}0} \end{bmatrix} \begin{bmatrix} -0.39 & 0.71 & -0.57 \\ -0.42 & -0.59 & -0.57 \\ 0 & 0 & 0 \end{bmatrix}$$

Practice: How to get eigenvectors? (trick without $X^T X$)

diagonal matrix with square roots of corresponding eigenvalues of $X^T X$

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

matrix with eigenvectors of $X^T X$ in columns

Practice: How to get eigenvectors? (trick without $X^T X$)

diagonal matrix with square roots of corresponding eigenvalues of $X^T X$

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

matrix with eigenvectors of $X^T X$ in columns

Practice: How to get eigenvectors?

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

$$X_{n\times d} = \begin{bmatrix} | & | & | & | \\ u_1 & u_2 & u_3 & u_4 \\ | & | & | & | \\ | & | & | & | \end{bmatrix}_{n\times n} \begin{bmatrix} \sqrt{\sigma_1} & 0 & 0 \\ 0 & \sqrt{\sigma_2} & 0 \\ 0 & 0 & \sqrt{\sigma_3} \\ 0 & 0 & 0 \end{bmatrix}_{n\times d} \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \\ - & v_3 & - \end{bmatrix}_{d\times d}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3$

Practice: How to get eigenvectors?

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

$$X_{n\times d} = \begin{bmatrix} | & | & | & | \\ u_1 & u_2 & u_3 & u_4 \\ | & | & | & | \\ | & | & | & | \end{bmatrix}_{n\times n} \begin{bmatrix} \sqrt{\sigma_1} & 0 & 0 \\ 0 & \sqrt{\sigma_2} & 0 \\ 0 & 0 & \sqrt{\sigma_3} \\ 0 & 0 & 0 \end{bmatrix}_{n\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{bmatrix}_{d\times d}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3$

Practice: How to get eigenvectors?

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

$$X_{n\times d} = \begin{bmatrix} | & | & | & \color{red}{0} \\ u_1 & u_2 & u_3 & \color{red}{0} \\ | & | & | & \color{red}{0} \\ | & | & | & \color{red}{0} \end{bmatrix}_{n\times n} \begin{bmatrix} \sqrt{\sigma_1} & 0 & 0 \\ 0 & \sqrt{\sigma_2} & 0 \\ 0 & 0 & \sqrt{\sigma_3} \\ 0 & 0 & 0 \end{bmatrix}_{n\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{bmatrix}_{d\times d}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3$

Practice: How to get eigenvectors?

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

$$X_{n\times d} \approx \begin{bmatrix} | & | & 0 & 0 \\ u_1 & u_2 & 0 & 0 \\ | & | & 0 & 0 \\ | & | & 0 & 0 \end{bmatrix}_{n\times n} \begin{bmatrix} \sqrt{\sigma_1} & 0 & 0 \\ 0 & \sqrt{\sigma_2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ — & v_3 & — \end{bmatrix}_{d\times d}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3$

Practice: How to get eigenvectors?

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

$$X_{n\times d} \approx \begin{bmatrix} | & | & 0 & 0 \\ u_1 & u_2 & 0 & 0 \\ | & | & 0 & 0 \\ | & | & 0 & 0 \end{bmatrix}_{n\times n} \begin{bmatrix} \sqrt{\sigma_1} & 0 & 0 \\ 0 & \sqrt{\sigma_2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n\times d} \begin{bmatrix} - & v_1 & - \\ - & v_2 & - \\ 0 & 0 & 0 \end{bmatrix}_{d\times d}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3$

Practice: How to get eigenvectors?

$$X \overset{SVD}{=} U\sqrt{\Sigma}V^T$$

$$X_{n\times d} \approx \begin{bmatrix} | & | & 0 & 0 \\ u_1 & u_2 & 0 & 0 \\ | & | & 0 & 0 \\ | & | & 0 & 0 \end{bmatrix}_{n\times n} \begin{bmatrix} \sqrt{\sigma_1} & 0 & 0 \\ 0 & \sqrt{\sigma_2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n\times d} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \\ 0 & 0 & 0 \end{bmatrix}_{d\times d}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3$

Theorem

Resulting matrix is the best approximation of the original matrix by a matrix of rank k in the sense of the difference between the two having the smallest possible Frobenius norm.

$$\|X - \tilde{X}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |x_{ij} - \tilde{x}_{ij}|^2}$$

Practice: How many eigenvectors should be kept?

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$$

1. Decide how much variance you want to retain

$$R^2 = \frac{\sum_{i=1}^{k} \sigma_i}{\sum_{j=1}^{d} \sigma_j} \qquad \text{(Variance retained)}$$

2. Scree Plots
3. Use a fixed k
   - typically from 50 to 1000
   - In LSA usually $k = 300$ (and omit the first vector?)

Practice: Should I start with TF-IDF matrix?

$$X_{i,j} = \underbrace{\log(TF(i,j)+1)}_{\text{local weight}} \cdot \underbrace{\left(1 + \frac{\sum_j p(i,j) \log p(i,j)}{\log D}\right)}_{\text{global weight}}$$

Usually about $1-2\%$ improvement in precision over standard TF-IDF.

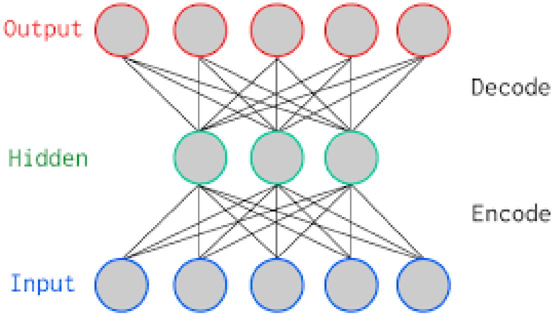
Practice: How to perform search?

We can not calculate cosine similarity because documents are in the concept space (k dimensions) and query is in the original term space (d dimensions)...

LSA: pros and cons

- clean formal framework
- clearly defined optimization criterion (one optimum!)
- more compact representation
- significantly improves recall
- sometimes a decrease in precision
- inverted index cannot be constructed
- normality assumption (reconstruction can have negative counts)
- substantial computational cost

Interesting fact: it can be done by NN



Theorem (Bourlard and Kamp, 1988)

If the hidden units have liner activation functions and quadratic cost is minimized, the network has a unique global minimum. At this minimum the network performs a projection onto the k-dimensional subspace which is spanned by the first k principal components of the data.

LSA is not enough

LSA solves many problems in the task of information retrieval. In other tasks like e.g. text classification it is often not enough.

Example (Sentiment Classification)

- We want to know which words are positive
- We have a handmade list of some positive words ["good", "excellent", "superb",...]
- How to find another positive words?

LSA is not enough

LSA solves many problems in the task of information retrieval. In other tasks like e.g. text classification it is often not enough.

Example (Sentiment Classification)

- We want to know which words are positive
- We have a handmade list of some positive words ["good", "excellent", "superb",...]
- How to find another positive words?

LSA is not enough

LSA solves many problems in the task of information retrieval. In other tasks like e.g. text classification it is often not enough.

Example (Sentiment Classification)

- We want to know which words are positive
- We have a handmade list of some positive words ["good", "excellent", "superb",...]
- How to find another positive words?

Beyond the term-document matrix

Previously we defined a context of the word as a *whole document* in which it appears $\Rightarrow$ **term-document matrix**

In this way we can capture general concept represented by a word (especially in a collection of short documents discussing single topics).

In order to capture the meaning of the word more precisely we should use a shorter context. $\Rightarrow$ **word-context matrix**

Beyond the term-document matrix

Previously we defined a context of the word as a *whole document* in which it appears $\Rightarrow$ **term-document matrix**

In this way we can capture general concept represented by a word (especially in a collection of short documents discussing single topics).

In order to capture the meaning of the word more precisely we should use a shorter context. $\Rightarrow$ **word-context matrix**

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	0	0	0	0	0	0	0
like	0	0	0	0	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	0	0	0	0	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	0	0	0	0	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	0	0	0	0	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	0	0	0	0	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	0	0	0	0	0
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	0	0	0	0	0	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	1	0	0	0	0	0	0
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	0	0	0	0	0	1
like	1	0	0	1	0	0	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	0	0	0	0	0	1
like	2	0	0	1	0	1	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	0	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	0	0	0	0	0	1
like	2	0	0	1	0	1	0	0
enjoy	0	0	0	0	0	0	0	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	1	0	0	0	0	0	0
flying	0	0	0	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Word-context matrix

I like Information Retrieval and I like Statistics.
I enjoy flying.

counts	I	like	enjoy	Info.	Ret.	Stats.	flying	and
I	0	2	1	0	0	0	0	1
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
Info.	0	1	0	0	1	0	0	0
Ret.	0	0	0	1	0	0	0	1
Stats.	0	1	0	0	0	0	0	0
flying	0	0	1	0	0	0	0	0
and	1	0	0	0	1	0	0	0

Problem

How to choose window size?

Can we use Language Model to measure word association?

$$P(q) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2)... = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

Adapted from a presentation by ChengXiang Zhai „Probabilistic Retrieval Model: Statistical Language Model"

Can we use Language Model to measure word association?

$$P(q) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2) \ldots = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

$P(w|\text{computer})$

the 0.032
a 0.019
is 0.014
we 0.008
...
text 0.00018
...
program 0.00013
software 0.0001

Adapted from a presentation by ChengXiang Zhai, Probabilistic Retrieval Model: Statistical Language Model"

Can we use Language Model to measure word association?

$$P(q) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2)... = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

$P(w\|\text{computer})$	$P(w)$
the 0.032	the 0.03
a 0.019	a 0.02
is 0.014	is 0.015
we 0.008	we 0.01
...	..
text 0.00018	text 0.00006
...	...
program 0.00013	program 0.00000125
software 0.0001	software 0.000000667

Adapted from a presentation by ChengXiang Zhai, "Probabilistic Retrieval Model: Statistical Language Model"

Can we use Language Model to measure word association?

$$P(q) = P(q_1) \cdot P(q_2|q_1) \cdot P(q_3|q_2)... = \prod_{i=1}^{N} P(q_i|q_{i-1})$$

$P(w|\text{computer})$

the 0.032
a 0.019
is 0.014
we 0.008
...
text 0.00018
...
program 0.00013
software 0.0001

$P(w)$

the 0.03
a 0.02
is 0.015
we 0.01
..
text 0.00006
...
program 0.00000125
software 0.000000667

$\frac{P(w|\text{computer})}{P(w)}$

software 150
program 104
...
text 3.0
..
the 1.1
a 0.99
is 0.9
we 0.8

Vector Semantics: Pointwise mutual information

Instead of using $P(w_i|w_{i-1})$, we will model a more general $P(c|w)$ basing on a term-context matrix.

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

$$PMI(w, c) = log_2 \frac{P(w, c)}{P(w)P(c)}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right)$$

Vector Semantics: Pointwise mutual information

Instead of using $P(w_i|w_{i-1})$, we will model a more general $P(c|w)$ basing on a term-context matrix.

$$Association(w,c) = \frac{P(c|w)}{P(c)} = \frac{P(w,c)}{P(w)P(c)}$$

$$PMI(w,c) = log_2 \frac{P(w,c)}{P(w)P(c)}$$

$$PPMI(w,c) = \max\left(0, log_2 \frac{P(w,c)}{P(w)P(c)}\right)$$

Vector Semantics: Pointwise mutual information

Instead of using $P(w_i|w_{i-1})$, we will model a more general $P(c|w)$ basing on a term-context matrix.

$$Association(w,c) = \frac{P(c|w)}{P(c)} = \frac{P(w,c)}{P(w)P(c)}$$

$$PMI(w,c) = log_2 \frac{P(w,c)}{P(w)P(c)}$$

$$PPMI(w,c) = \max\left(0, log_2 \frac{P(w,c)}{P(w)P(c)}\right)$$

Vector Semantics: Pointwise mutual information

Instead of using $P(w_i|w_{i-1})$, we will model a more general $P(c|w)$ basing on a term-context matrix.

$$Association(w, c) = \frac{P(c|w)}{P(c)} = \frac{P(w, c)}{P(w)P(c)}$$

$$PMI(w, c) = log_2 \frac{P(w, c)}{P(w)P(c)}$$

$$PPMI(w, c) = \max \left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right)$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = information, c = data) = \frac{6}{19}$$

$$P(w = information) = \frac{11}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) = \log_2 \frac{\frac{6}{19}}{\frac{11}{19} \cdot \frac{7}{19}} = 0.568$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = information, c = data) = \frac{6}{19}$$

$$P(w = information) = \frac{11}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) = \log_2 \frac{\frac{6}{19}}{\frac{11}{19} \cdot \frac{7}{19}} = 0.568$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = information, c = data) = \frac{6}{19}$$

$$P(w = information) = \frac{11}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) = \log_2 \frac{\frac{6}{19}}{\frac{11}{19} \cdot \frac{7}{19}} = 0.568$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = information, c = data) = \frac{6}{19}$$

$$P(w = information) = \frac{11}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) = \log_2 \frac{\frac{6}{19}}{\frac{11}{19} \cdot \frac{7}{19}} = 0.568$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = ?$$

$$P(w = digital) = ?$$

$$P(c = data) = ?$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = ?$$

$$P(c = data) = ?$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = \frac{4}{19}$$

$$P(c = data) = ?$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = \frac{4}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) =$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(w = digital, c = data) = \frac{1}{19}$$

$$P(w = digital) = \frac{4}{19}$$

$$P(c = data) = \frac{7}{19}$$

$$PPMI(w, c) = \max\left(0, log_2 \frac{P(w, c)}{P(w)P(c)}\right) = \log_2 \frac{\frac{1}{19}}{\frac{4}{19} \cdot 719} = \cancel{-0.26} \; 0$$

Vector Semantics: example

	computer	data	pinch	result	sugar
apricot	0	0	2.25	0	2.25
pineapple	0	0	2.25	0	2.25
digital	1.66	0	0	0	0
information	0	0.57	0	0.47	0

Many applications: e.g. extensions of polarity lexicons (often you must set a threshold on PPMI measure)

Other association measures?

Other association measures?

$$H_0: \text{ words occur independently of each other}$$

$$H_1: \text{ words does not occur independently of each other}$$

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

You can choose threshold according to your tolerance for I type error!

PS. In NLP one use smaller α than on standard Statistics courses

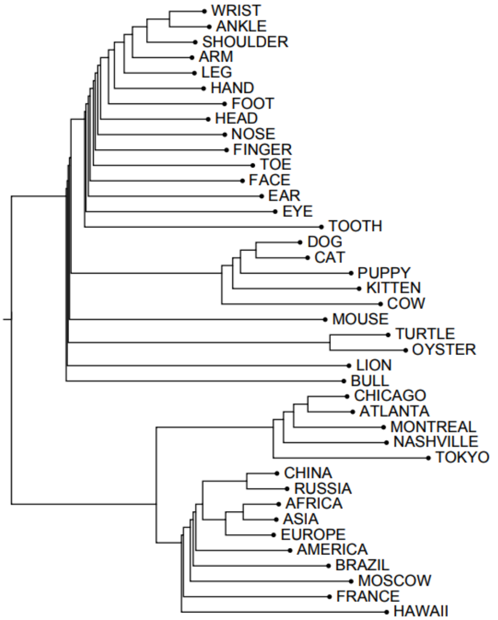
Other association measures?

$$H_0: \text{ words occur independently of each other}$$

$$H_1: \text{ words does not occur independently of each other}$$

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

You can choose threshold according to your tolerance for I type error!
PS. In NLP one use smaller α than on standard Statistics courses

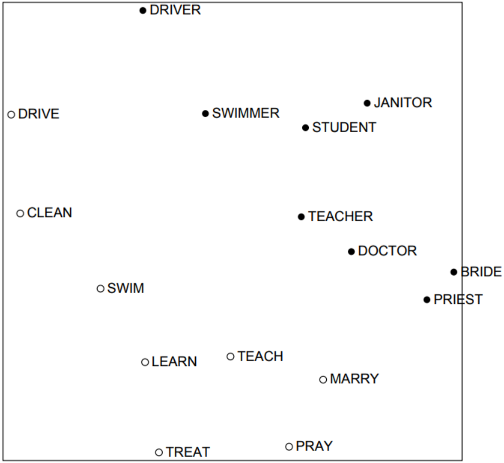

WRIST
ANKLE
SHOULDER
ARM
LEG
HAND
FOOT
HEAD
NOSE
FINGER
TOE
FACE
EAR
EYE
TOOTH
DOG
CAT
PUPPY
KITTEN
COW
MOUSE
TURTLE
OYSTER
LION
BULL
CHICAGO
ATLANTA
MONTREAL
NASHVILLE
TOKYO
CHINA
RUSSIA
AFRICA
ASIA
EUROPE
AMERICA
BRAZIL
MOSCOW
FRANCE
HAWAII

Problems to solve

- Quadratic increase in size with vocabulary
- High dimensional and sparse
- A lot of noise
- Subsequent classification models have sparsity issues (weak generalization)
- Higher order co-occurrence?

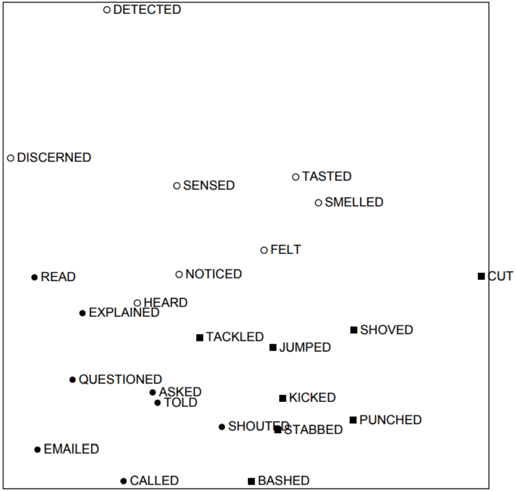
Can we apply PCA to word-context matrix?

We have a sparse matrix... PCA time! $\Rightarrow$ dense representation



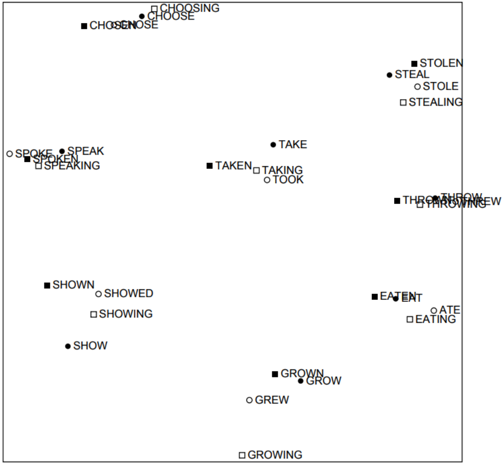
Can we apply PCA to word-context matrix?

We have a sparse matrix... PCA time! $\Rightarrow$ dense representation



Can we apply PCA to word-context matrix?

We have a sparse matrix… PCA time! $\Rightarrow$ dense representation



Problems to solve?

Dense representation: another idea (Skip-gram)

$$Association(w_i, w_j) = c_i^T v_j$$

where c and v are *learned*.

Dense representation: another idea (Skip-gram)

$$Association(w_i, w_j) = e^{c_i^T v_j}$$

where c and v are *learned*.

Dense representation: another idea (Skip-gram)

$$Association(w_i, w_j) = e^{c_i^T v_j}$$

where c and v are *learned*.

$$P(w_i|w_j) = \frac{e^{c_i^T v_j}}{\sum_k e^{c_k^T v_j}}$$

Dense representation: another idea (Skip-gram)

$$Association(w_i, w_j) = e^{c_i^T v_j}$$

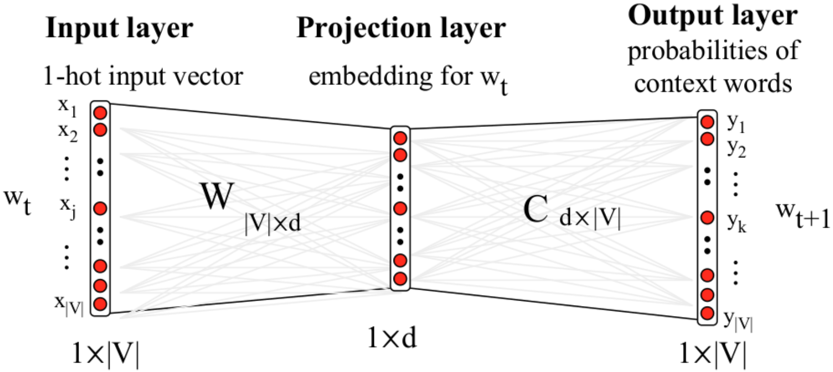
where c and v are *learned*.

$$P(w_i|w_j) = \frac{e^{c_i^T v_j}}{\sum_k e^{c_k^T v_j}}$$

$$L = \sum_{i=1}^{N} \sum_{j \in C(x_i)} \log P(w_j|w_i)$$

Start with random c_i, w_i and... optimize!

Skip-gram model

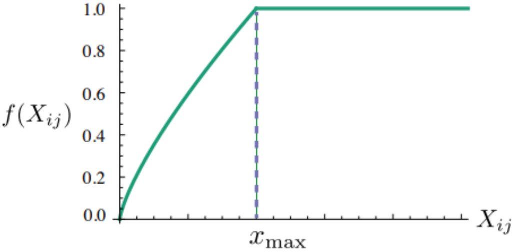


Theorem (Levy& Goldberg, 2014)

Skip-gram model reaches its optimum when $WC^T = X^{PMI} - \log k$

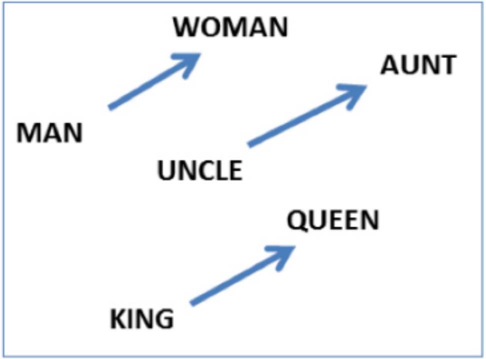
GloVe (Pennington, Socher, & Manning, 2014)

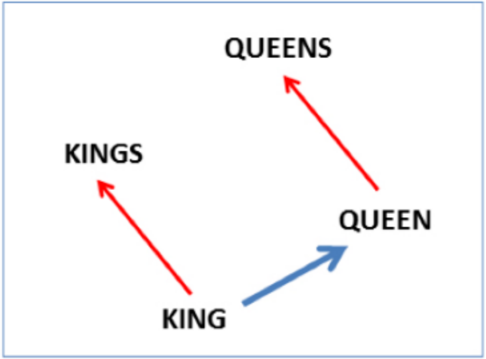
$$L = \sum_{i \in V} \sum_{j \in V} f(n_{i,j}) (c_i^T v_j + b_i + b'_j - \log n_{i,j})^2$$



Evaluation of word embeddings

- word similarity
- word analogy
- task-specific measures





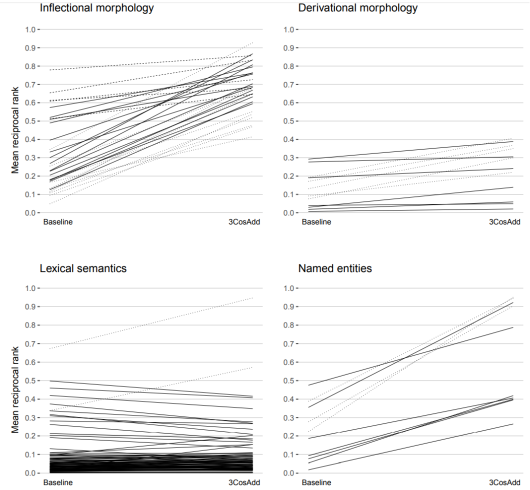
Properties of Word Embeddings

- semantic analogy
  - puppy - dog $\approx$ kitten - cat
- syntactic analogy
  - taller - tall $\approx$ smaller - small
- NN search: find most similar
  - blue: red, black, pink...
  - Japan: Korea, China
  - dance: dancing, singing, dances, music, ...
  - tea: coffee, lemon, sugar
- "words arithmetic": X to Y is as A to ...?
  - king - man + woman = ?
  - Paris - France + Germany = ?
  - Tadeusza - Tadeusz + Marek = ?
  - Shakespeare - English + Polish = ?
  - 0.5 (first + fifth) = ?

SVD vs Word2Vec

win	Method	WordSim Similarity	WordSim Relatedness	Bruni et al. MEN	Radinsky et al. M. Turk	Luong et al. Rare Words	Hill et al. SimLex	Google Add / Mul	MSR Add / Mul
2	PPMI	.732	**.699**	.744	.654	.457	.382	.552 / .677	.306 / .535
	SVD	.772	.671	**.777**	.647	**.508**	.425	.554 / .591	.408 / .468
	SGNS	**.789**	.675	.773	**.661**	.449	**.433**	.676 / **.689**	.617 / **.644**
	GloVe	.720	.605	.728	.606	.389	.388	.649 / .666	.540 / .591
5	PPMI	.732	**.706**	.738	**.668**	.442	.360	.518 / .649	.277 / .467
	SVD	.764	.679	**.776**	.639	**.499**	**.416**	.532 / .569	.369 / .424
	SGNS	**.772**	.690	.772	.663	.454	.403	.692 / **.714**	.605 / **.645**
	GloVe	.745	.617	.746	.631	.416	.389	.700 / .712	.541 / .599
10	PPMI	.735	**.701**	.741	.663	.235	.336	.532 / .605	.249 / .353
	SVD	.766	.681	.770	.628	**.312**	.419	.526 / .562	.356 / .406
	SGNS	**.794**	.700	**.775**	**.678**	.281	**.422**	.694 / .710	.520 / **.557**
	GloVe	.746	.643	.754	.616	.266	.375	.702 / **.712**	.463 / .519

Word2Vec: does „the word arithmetic" really work?



The World of Embeddings: Embed All The Things!

- Word2Vec (Google)
- GloVe (Stanford)
- FastText (Facebook)
- StarSpace a general-purpose neural model for efficient learning of entity embeddings for solving a wide variety of problems
  - Learning word, sentence or document level embeddings.
  - Information retrieval: ranking of sets of entities/documents or objects, e.g. ranking web documents.
  - Text classification, or any other labeling task.
  - Metric/similarity learning, e.g. learning sentence or document similarity.
  - Content-based or Collaborative filtering-based Recommendation
  - Embedding graphs, e.g. multi-relational graphs such as Freebase.
  - Image classification, ranking or retrieval (e.g. by using existing ResNet features).

Summary

- Query likelihood (Unigram& Bigram LM)
- Distributional hypothesis
- Latent Semantic Analysis
- Point-wise Mutual Information
- Skip-gram model

Thank you!

Zapraszam do koła naukowego!

Group of
Horribly
Optimistic
STatisticians

WWW: `ghost.put.poznan.pl`