

Idea Analytics & code along: Intro to different techniques for data mining

Managing Innovation - Master in Innovation Management and Business Development - Spring 2022

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Center for Humanities Computing
Aarhus University, Denmark

April 22, 2022



CENTER FOR HUMANITIES
COMPUTING AARHUS



Outline

① Data*

- Data
- Data science
- Type of Data
- KDD
- Normalization
- Big Data?
- Human-in-the-Loop Models

② Example

- Trend Reservoirs

③ Techniques

- Text analytics
- Word counts
- Word distributions
- Vector space model
- Sentiment analysis
- Non-negative Matrix Factorization
- Classification
- EDA
- Neural embeddings

④ Feedback

- Data feedback
- Collaboration feedback

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models

Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings

Feedback
Data feedback
Collaboration feedback





JSTOR Data For Research, $n = 43,802$, $\sim 400m$ words, Eng*, Fr, Ger.

Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

EDA

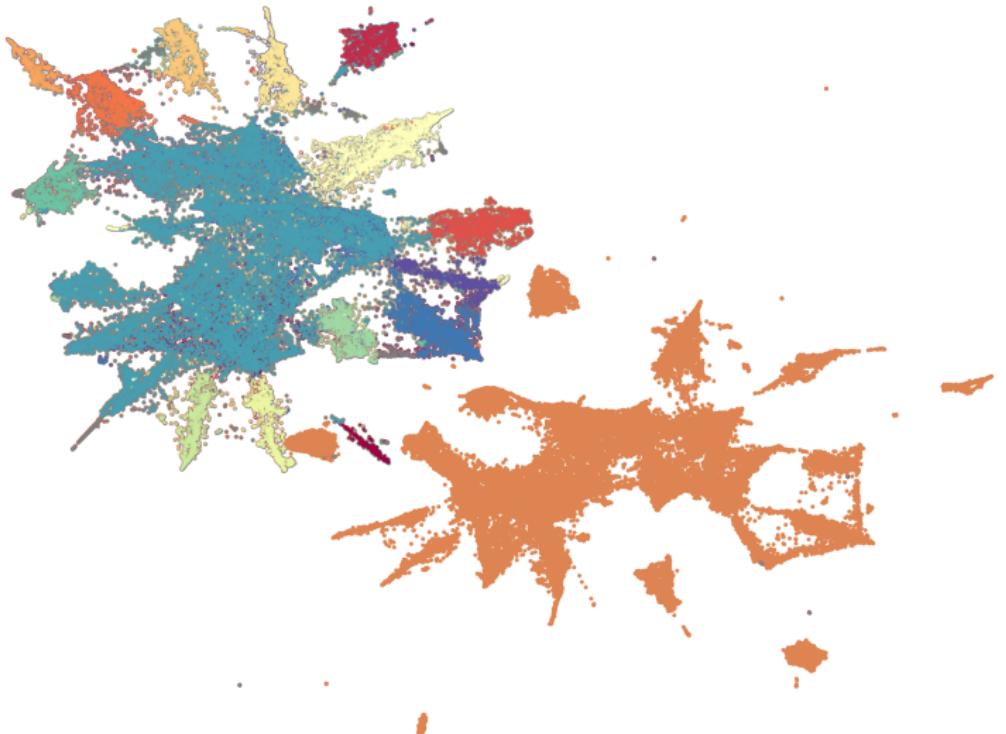
Neural embeddings

Feedback

Data feedback

Collaboration feedback





Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

EDA

Neural embeddings

Feedback

Data feedback

Collaboration feedback

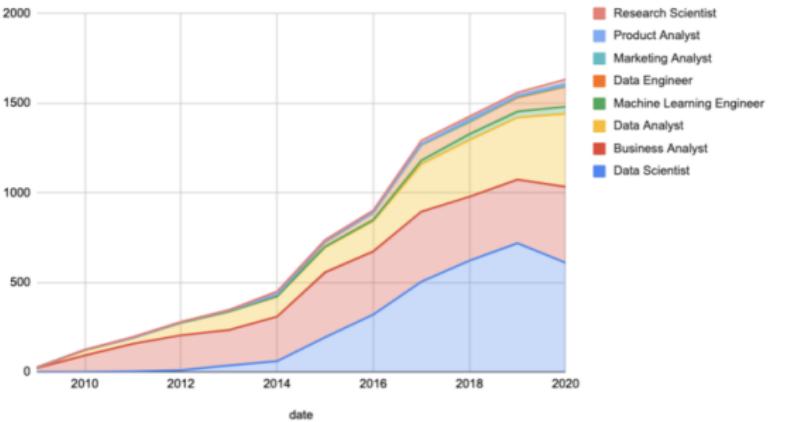


Why does this matter?

Idea Analytics & code along: Intro to different techniques for data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data Science Position Growth (2010-2020)



With the rise of Big Data, we have seen a rapid increase in data-related positions, source: Stephanie Glen (2021) 'Data Science Job Market Shrinking? Not So Fast'

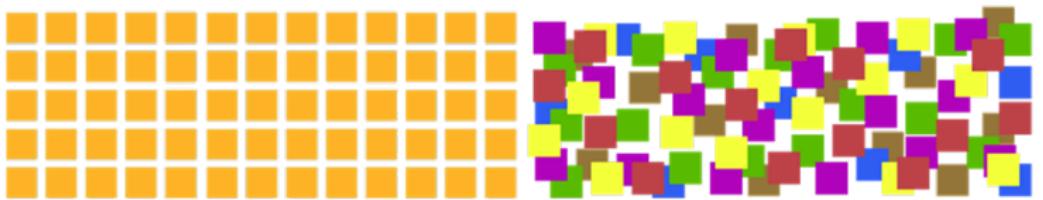
Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback



Type of Data

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io



Structured versus unstructured data.

Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

EDA

Neural embeddings

Feedback

Data feedback

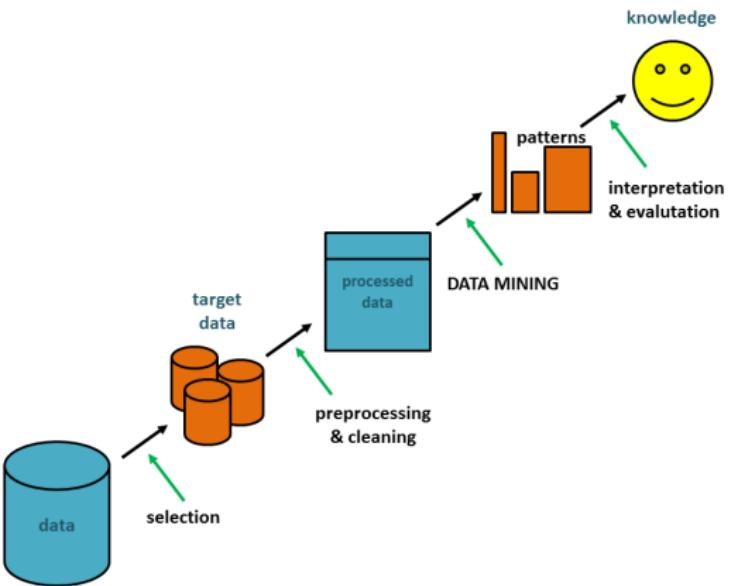
Collaboration feedback



Knowledge Discovery in Data(bases)

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io



Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

EDA

Neural embeddings

Feedback

Data feedback

Collaboration feedback



Data Normalization

We normalize natural language data in order to increase our signal/reduce the randomness

- Removal of duplicate whitespaces and punctuation.
- Casefolding.
- Removal or substitution of special characters/emojis (e.g.: remove hashtags).
- Substitution of contractions (very common in English; e.g.: 'I'm'→'I am').
- Transform word numerals into numbers (eg.: 'twenty three'→'23').
- Substitution of values for their type (e.g.: '50DKK'→'MONEY').
- Acronym normalization (e.g.: 'DK'→'Denmark') and abbreviation normalization (e.g.: 'btw'→'by the way').
- Normalize date formats, social security numbers or other data that have a standard format.
- Spell correction
- Removal of gender/time/grade variation with Stemming or Lemmatization.
- Substitution of rare words for more common synonyms.
- Stop word removal (~ dimensionality reduction technique)

Idea Analytics & code along: Intro to different techniques for data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

EDA

Neural embeddings

Feedback

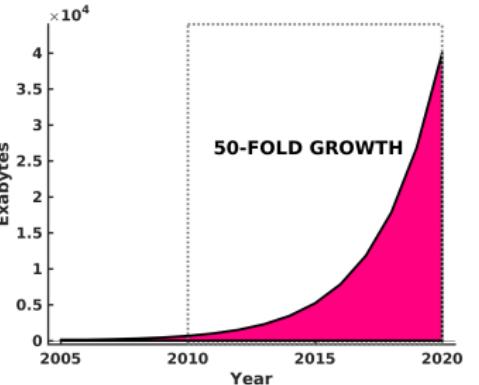
Data feedback

Collaboration feedback



Big Data are not necessarily enough data to solve your problem.

- wrong data
- lack precision/low detection rate
- problem complexity
- (too many data)



revise question OR collect more data OR enrich existing data

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback



Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

EDA

Neural embeddings

Feedback

Data feedback

Collaboration feedback

Human-in-the-Loop Models

as task complexity increases, a need for (operational approaches to) leveraging human intelligence in the development of learning algorithms has become apparent

Type	Human Involvement	Resources	Relevance
Out-of-the-loop	not required	low	low
On-the-loop	checking	medium	medium↓
In-the-loop	required	high	medium↑

WHEN

THEN

algorithms are not understanding the input

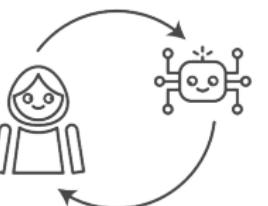
data input is interpreted incorrectly

algorithms do not know how to perform the task

to make models more accurate

cost of errors is too high in development

data is rare or not available



HITL Models



Trend Reservoirs in Social Media

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

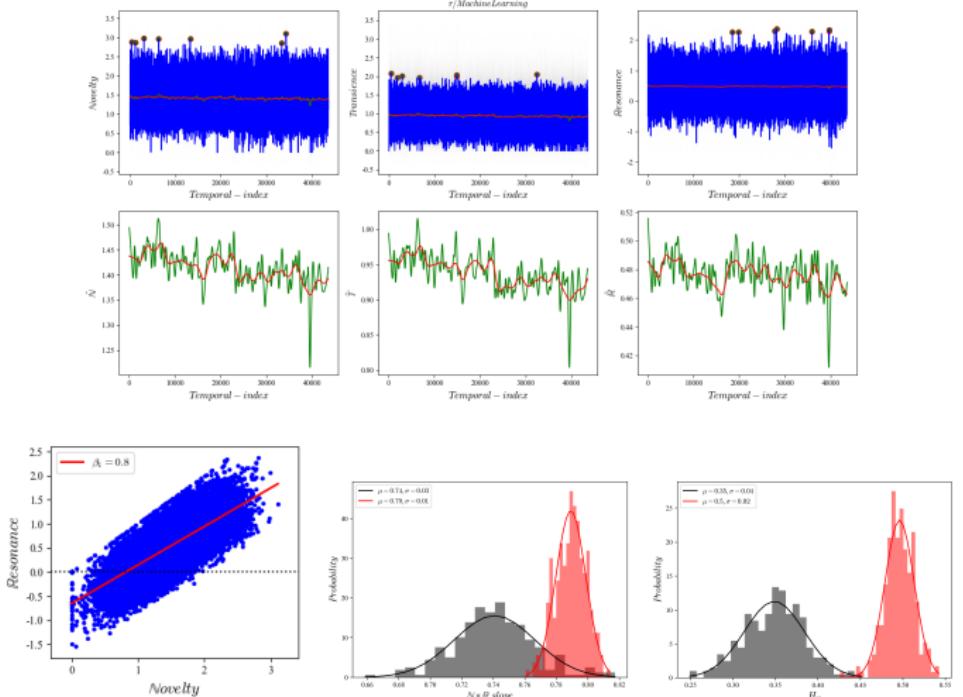
EDA

Neural embeddings

Feedback

Data feedback

Collaboration feedback

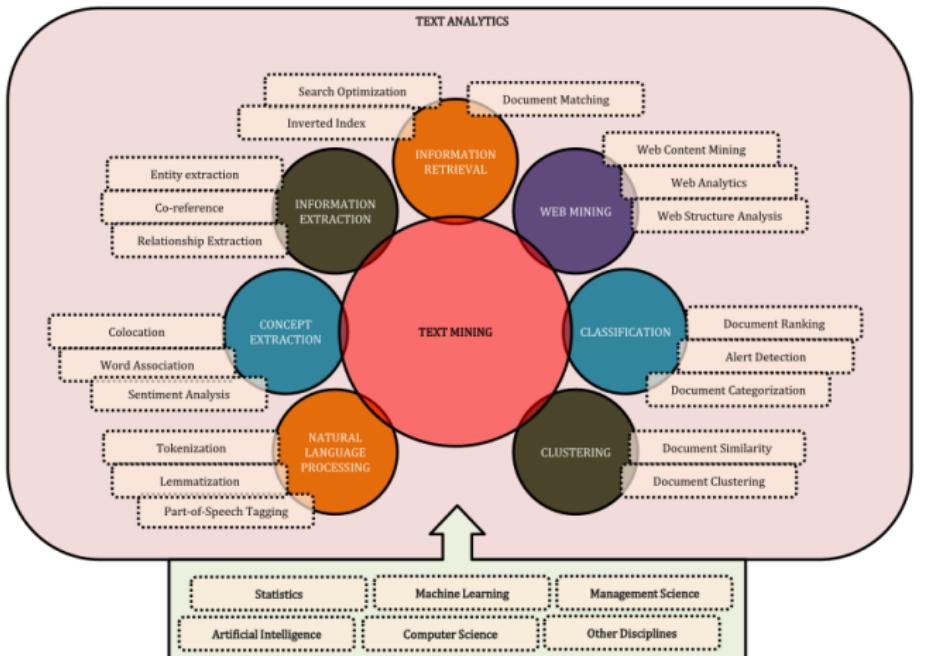


trend reservoirs (i.e., social media signals that display high trend potential) can be identified by their relationship between novel and resonant behavior, and their minimal persistence.

Text Analytics

Idea Analytics & code along: Intro to different techniques for data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io



- Data*
- Data
- Data science
- Type of Data
- KDD
- Normalization
- Big Data?
- Human-in-the-Loop Models
- Example
- Trend Reservoirs
- Techniques
- Text analytics
- Word counts
- Word distributions
- Vector space model
- Sentiment analysis
- Non-negative Matrix Factorization
- Classification
- EDA
- Neural embeddings
- Feedback
- Data feedback
- Collaboration feedback



Word counts

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

words are (one of) the basic units of meaning

most TM techniques rely on word frequencies, that is, we tokenize a text at the word level and count the number of tokens for each type

I am Daniel	
I am Sam	'I' 'am' 'Daniel' 'I'
Sam I am	'am' 'Sam' 'Sam' 'I'
That Sam-I-am	'am' 'That' 'Sam' 'I'
That Sam-I-am!	'am' 'That' 'Sam' 'I'
I do not like	'am' 'I' 'do' 'not' 'like'
that Sam-I-am	'that' 'Sam' 'I' 'am' ...
...	

a	1	59	0.073
am	1	16	0.02
and	1	24	0.03
anywhere	1	1	0.001
anywhere	1	7	0.009
...			
you	1	34	0.042
<i>total</i>	55	804	1.0

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models

Example
Trend Reservoirs

Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings

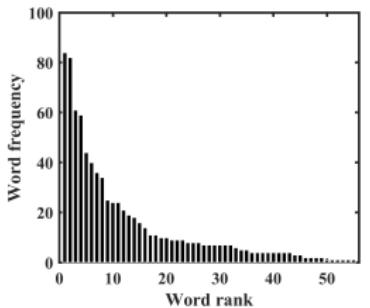
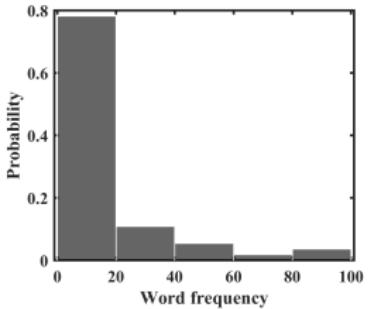
Feedback
Data feedback
Collaboration feedback



Word Distributions

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io



Most words are infrequent, but a few words are very frequent

'i' 'not' 'them' 'a' 'like' 'in' 'do' 'you'
'would'

Model a text as a distribution over words. Some words are more likely than other.

Often times we are looking at the mid-range (not too likely and not too unlikely).

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback



Vector Space Model

any collection of m documents can be represented in the vector space model by a document-term matrix of m documents and n terms

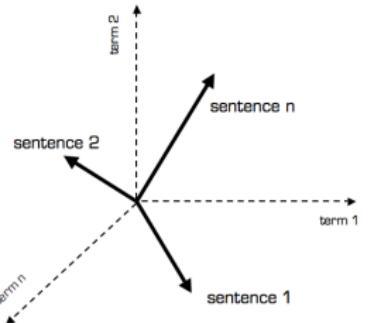
a vector space model is a basic modeling mechanism for a word- or document-space (whether we look at rows or columns)

- a document vector with only one word is collinear to the vocabulary word axis
- a document vector that does not contain a specific word is orthogonal/perpendicular to the word axis
- two documents are identical if they contain the same words in a different order (BOW assumption)

You can use the 'words' as predictors x_1, x_2, \dots, x_m in:

$$Y_i = f(X_i, \beta) + \epsilon_i$$

but it can result in an underdetermined system



Document space	t_1	t_2	t_3	...	t_n	← Term vector space
D_1	a_{11}	a_{12}	a_{13}	...	a_{1n}	
D_2	a_{21}	a_{22}	a_{23}	...	a_{2n}	
D_3	a_{31}	a_{32}	a_{33}	...	a_{3n}	
...						
D_m	a_{m1}	a_{m2}	a_{m3}	...	a_{mn}	
Q	b_1	b_2	b_3	...	b_n	

Idea Analytics & code along: Intro to different techniques for data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models

Example
Trend Reservoirs

Techniques
Text analytics
Word counts
Word distributions

Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings

Feedback
Data feedback
Collaboration feedback



Sentiment Analysis

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

- we can use a dictionary to extract cognitive and affective keywords from a collection of documents and apply a sentiment function

```
1 'Did Crooked Hillary help disgusting (check out sex tape and past) Alicia M become a U.S. citizen  
2 so she could use her in the debate?'  
3  
4 Positive sex, citizen  
5 Negative crooked, hillary, disgusting, out  
6 Sentiment Score (2+1) + (-2-1-3-1) = -4  
7 Sentiment Polarity Negative  
8 Overall Score Sum of all sentence scores
```

- a sentiment vector is a vector of keyword frequencies weighted by sentiment scores

You can avoid the underdetermined system and use the overall 'sentiment' as predictor in: $Y_i = f(X_i, \beta) + \epsilon_i$

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback

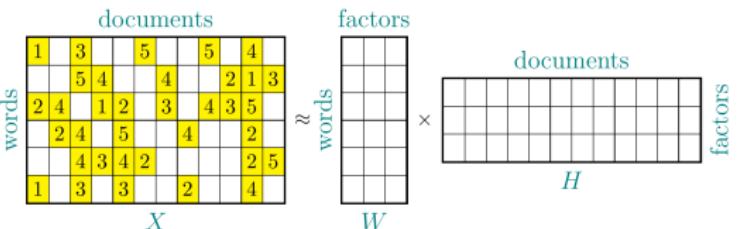


Matrix decomposition of TD matrix

Idea Analytics & code
along: Intro to
different techniques for
data mining

For bag-of-words representation X w. documents in columns and words in rows, such that each entry X_{ij} is the i th word in the j th column, we solve

$$\operatorname{argmin}_{W,H} \| X - WH \|^2_F \quad \text{s.t. } W, H \geq 0$$



Non-negative matrix factorization of TD matrix X

WH is a low-rank approximation of the data (++fewer factors than documents and words) and each document is the weighted sum of columns in W , “document factors” \sim themes or topics

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data*

Data

Data science

Type of Data

KDD

Normalization

Big Data?

Human-in-the-Loop Models

Example

Trend Reservoirs

Techniques

Text analytics

Word counts

Word distributions

Vector space model

Sentiment analysis

Non-negative Matrix Factorization

Classification

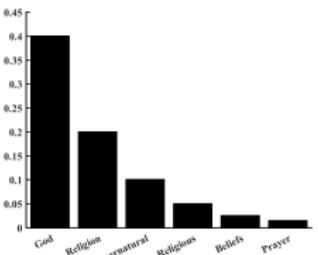
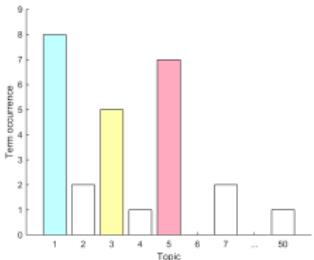
EDA

Neural embeddings

Feedback

Data feedback

Collaboration feedback



Document representation from NMF

Original document

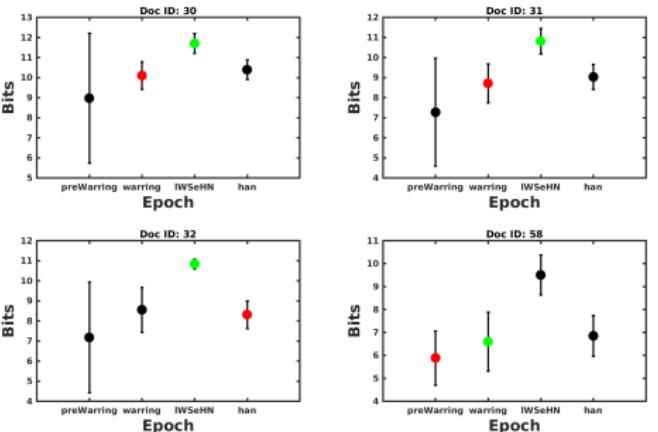
You can use the 'topics' as predictors x_1, x_2, \dots, x_n in $Y_i = f(X_i, \beta) + \epsilon_i$
where $n < m$ for word-based regression



Text Classification

Idea Analytics & code
along: Intro to
different techniques for
data mining

- Given labeled data, a classification algorithm will output a solution that categorizes new examples → associate labels with subsets of the data (c.f., logistic regression)
- data (features) with class values (~ labeled data), excellent opportunity to make use of metadata (e.g., reviews scores, evaluations &c)



Source classification of Chinese historical documents

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback



Naive Bayes

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

A simple and very popular probability learning model that can be implemented very efficiently

The probability of a document d being in class c , $P(c | d)$ is computed as:

$$P(c | d) \propto P(c) \prod_{i=1}^m P(t_i | c) \quad (1)$$

and the class of a document d is then computed as:

$$c_{MAP} = \arg \max_{c \in \{c_1, c_2\}} P(c | d) \quad (2)$$

Naive assumption that the presence/absence of a feature is completely independent of other features.

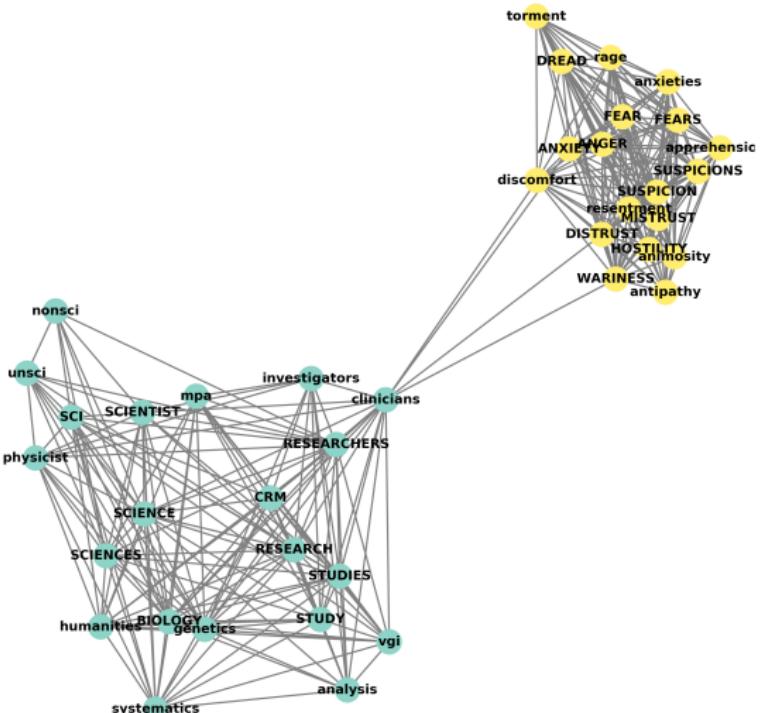
Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback



Exploratory analysis with Deep Neural Nets

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io



Sparse query graph for MISTRUST and SCIENCE

CENTER FOR HUMANITIES
COMPUTING AARHUS



- Data*
- Data
- Data science
- Type of Data
- KDD
- Normalization
- Big Data?
- Human-in-the-Loop Models
- Example
- Trend Reservoirs
- Techniques
- Text analytics
- Word counts
- Word distributions
- Vector space model
- Sentiment analysis
- Non-negative Matrix Factorization
- Classification
- EDA
- Neural embeddings
- Feedback
- Data feedback
- Collaboration feedback

Distributed word representation

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

For a discrete input w_0, w_1, \dots, w_n , we train a simple feedforward network such that

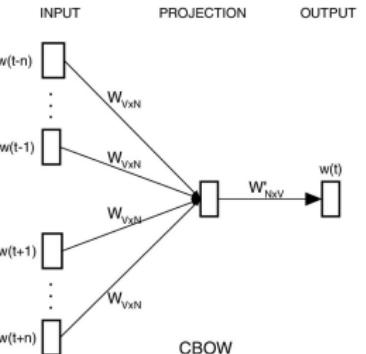


Figure 1: CBOW architecture for learning neural embeddings at the word level

$$\frac{1}{T} \sum_{t=1}^T \sum_{-n \leq j \leq n} \log p(w_t | w_{t+j}), \quad j \neq 0$$

Semantic similarity between any two word embeddings, A and B , can then be measured as their angular similarity

$$1 - \frac{\cos^{-1}\left(\frac{A \cdot B}{\|A\| \|B\|}\right)}{\pi}$$

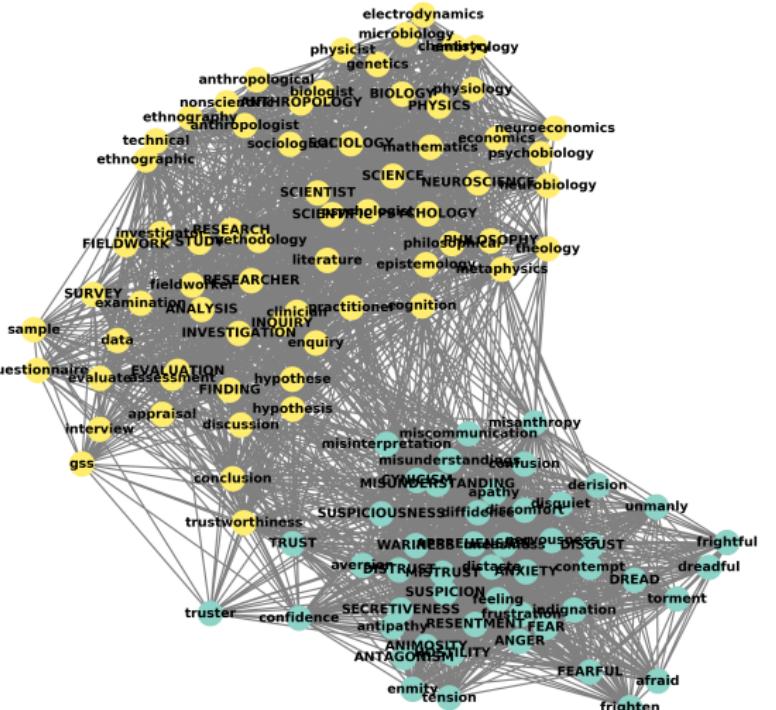


- Data*
- Data
- Data science
- Type of Data
- KDD
- Normalization
- Big Data?
- Human-in-the-Loop Models
- Example
- Trend Reservoirs
- Techniques
- Text analytics
- Word counts
- Word distributions
- Vector space model
- Sentiment analysis
- Non-negative Matrix Factorization
- Classification
- EDA
- Neural embeddings
- Feedback
- Data feedback
- Collaboration feedback

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

- Data*
- Data
- Data science
- Type of Data
- KDD
- Normalization
- Big Data?
- Human-in-the-Loop Models
- Example
- Trend Reservoirs
- Techniques
- Text analytics
- Word counts
- Word distributions
- Vector space model
- Sentiment analysis
- Non-negative Matrix Factorization
- Classification
- EDA
- Neural embeddings
- Feedback
- Data feedback
- Collaboration feedback



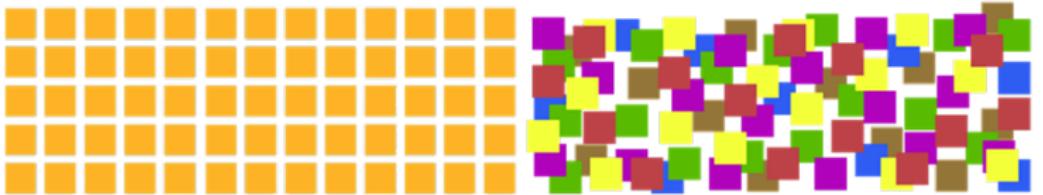
Dense query graph for MISTRUST and SCIENCE



Data

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io



most solutions were focused on structured data - low effort.

- you never have enough data!

no effect, property of the data or 'in the eye of the beholder'

all data come with limitations

understand your data (e.g., visualization: look for unusual patterns)

the data are never as assumed (e.g., latent groups are the rule)

- utilize unstructured data

annotate, enrich and augment...

- dynamics matter!

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models

Example
Trend Reservoirs

Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings

Feedback
Data feedback
Collaboration feedback



Collaboration

Idea Analytics & code
along: Intro to
different techniques for
data mining

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io



- interdisciplinary collaboration is *costly...*
 - requires a common frame of reference
 - psychology and information systems are good matches
- *you will often have to make decisions on weak evidence*
 - you are probably more qualified to make the call
- *predictive accuracy is not the only performance metric...*
 - inspect and test for biases

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models

Example
Trend Reservoirs

Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings

Feedback
Data feedback
Collaboration feedback



Thank you for your attention

chcaa@cas.au.dk
chcaa.io

slides: http://knielbo.github.io/files/kln_mginno.pdf

Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings
Feedback
Data feedback
Collaboration feedback

