

Lesson 14: Idea Analytics

Managing Innovation - Master in Innovation Management and Business
Development - Spring 2021

Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Center for Humanities Computing
Aarhus University, Denmark

April 19, 2021



CENTER FOR HUMANITIES
COMPUTING AARHUS



Outline

1 Data*

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

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

2 Example

- Trend Reservoirs

- Example

Trend Reservoirs

3 Techniques

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

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



Data Access

<https://worker02.chcaa.au.dk/jupyter/hub/login>

Communication

chcaa@cas.au.dk

Add subject: [Group <your group number>] <description>

Example: You want to submit group 5's requirements:

- list the group's requirements, e.g., 'estimate the effect of comment tone on number of votes'
- add '[Group 5] requirements' to the subject line
- submit to chcaa@cas.au.dk

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	





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



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

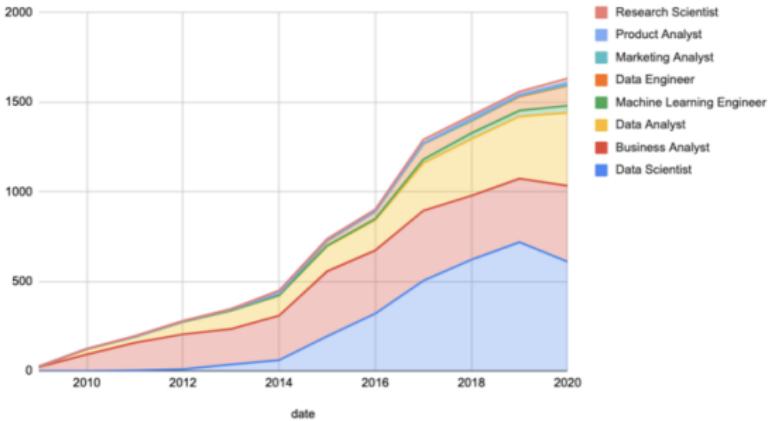
Neural embeddings



Why does this matter?

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'

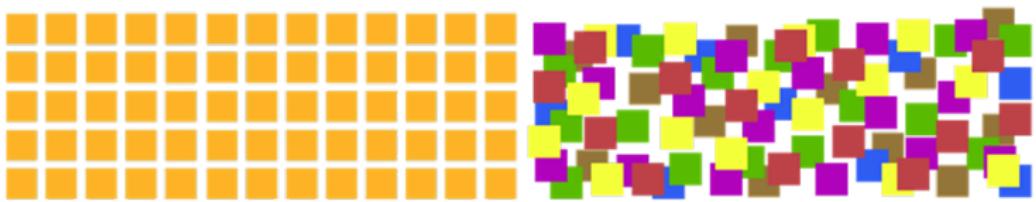


Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models
Example
Trend Reservoirs
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings

Type of Data

Lesson 14: Idea
Analytics

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

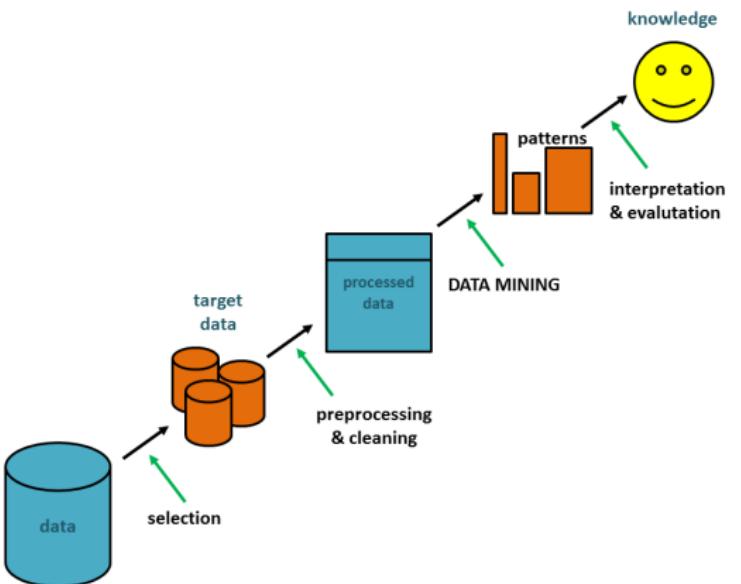
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings



Knowledge Discovery in Data(bases)

Lesson 14: Idea
Analytics

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



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)



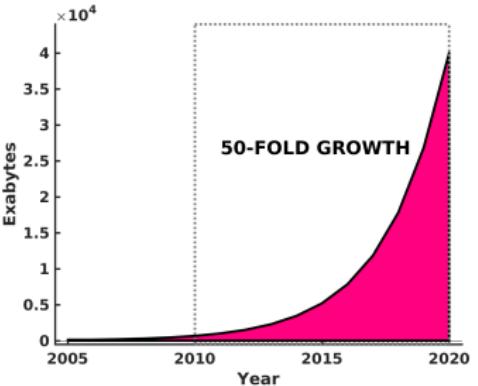
Data*
Data
Data science
Type of Data
KDD
Normalization
Big Data?
Human-in-the-Loop Models

Example
Trend Reservoirs

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

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

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

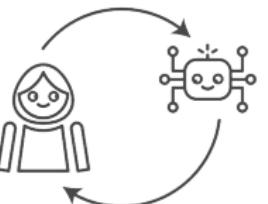
data input is interpreted incorrectly

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



- 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

Trend Reservoirs in Social Media

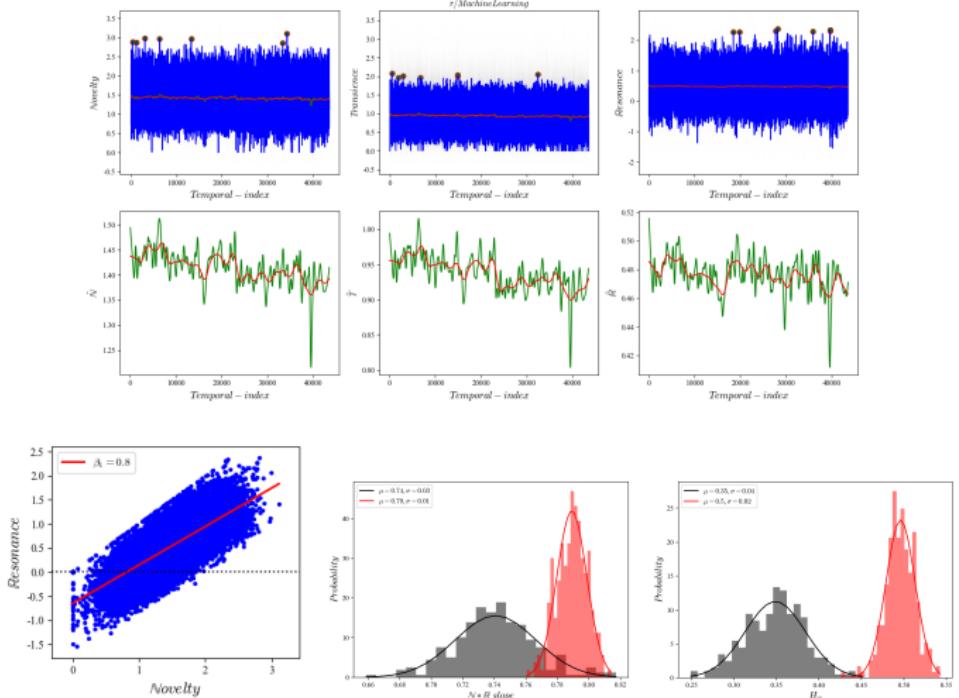
Lesson 14: Idea Analytics

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

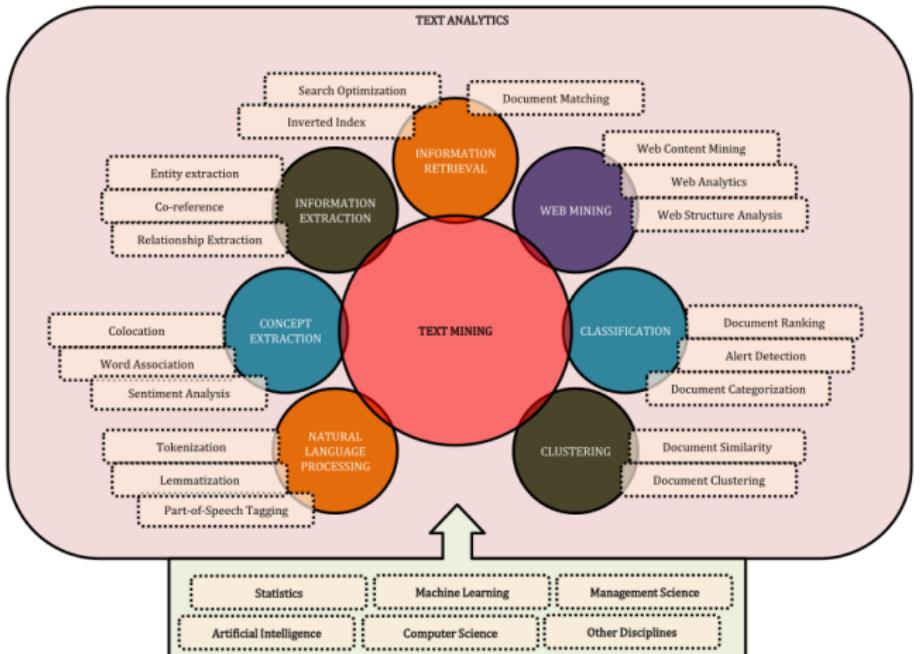


trend reservoirs (i.e., social media signals that display high trend potential) can be identified by their relationship between novel and resonant behavior, and their minimal persistence.

Text Analytics

Lesson 14: Idea
Analytics

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



Word counts

Lesson 14: Idea
Analytics

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'	'Daniel'	'I'	a	1	59	0.073
I am Sam	'am'	'Sam'	'Sam'	'I'	am	1	16	0.02
Sam I am	'am'	'Sam'	'Sam'	'I'	and	1	24	0.03
That Sam-I-am	'am'	'That'	'Sam'	'I'	anywhere	1	1	0.001
That Sam-I-am!	'am'	'That'	'Sam'	'I'	anywhere	1	7	0.009
I do not like	'am'	'I'	'do'	'not'	'like'	...		
that Sam-I-am	'that'	'Sam'	'I'	'am'	you	1	34	0.042
...					total	55	804	1.0

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

Example
Trend Reservoirs

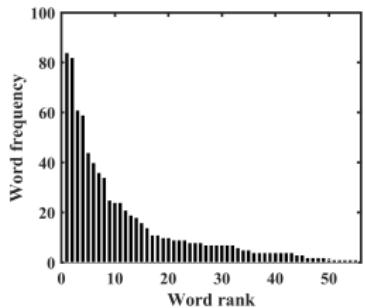
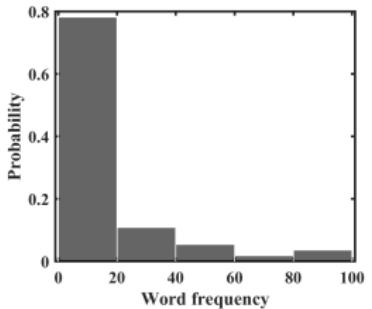
Techniques
Text analytics
Word counts
Word distributions
Vector space model
Sentiment analysis
Non-negative Matrix Factorization
Classification
EDA
Neural embeddings



Word Distributions

Lesson 14: Idea
Analytics

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



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

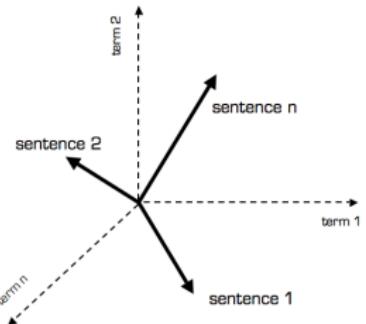
a vector space model is a basic modeling mechanism for a word- or document-space (whether we look at rows or columns)

- 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	



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

Sentiment Analysis

Lesson 14: Idea
Analytics

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

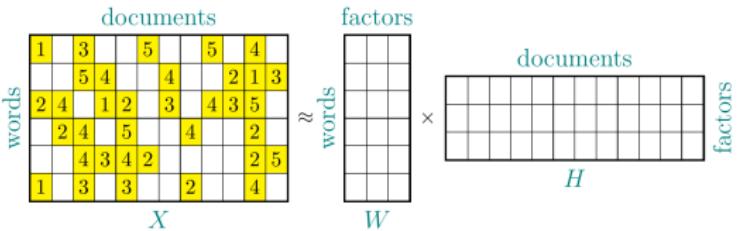
Example
Trend Reservoirs

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

Matrix decomposition of TD matrix

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

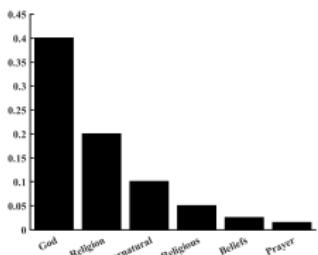
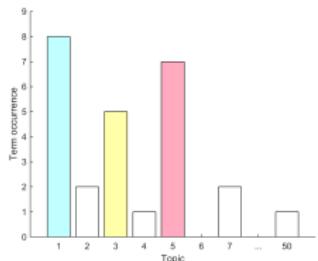
$$\underset{W, H}{\operatorname{argmin}} \| X - WH \|_F^2 \quad \text{s.t. } W, H \geq 0$$



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

- 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



Document representation from NMF

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

ABSTRACT—We present two studies aimed at resolving experimentally whether **religion** increases prosocial behavior in the anonymous dictator game. Subjects allocated more money to anonymous strangers when **God** concepts were implicitly activated than when neutral or no concepts were activated. This effect was at least as large as that obtained when concepts associated with secular moral institutions were primed. A trait measure of self-reported religiosity did not seem to be associated with prosocial behavior. We discuss different possible mechanisms that may underlie this effect, focusing on the hypotheses that the religious prime had an **ideomotor effect** on generosity or that it activated a **felt presence** of supernatural watchers. We then discuss implications for theories positing religion as a facilitator of the emergence of early large-scale societies of cooperators.

Many theorists have suggested that the cognitive availability of omniscient and omnipresent **supernatural** agents has had a dramatic impact on the development of large-scale human societies. The imagined presence of such agents, along with emotional ritual and costly commitment to the social group they govern, may have been the major development that allowed genetically unrelated individuals to interact in cooperative ways (e.g., Atman & Neocleous, 2004; Irons, 1991; Sosis & Ruffle, 2004). The research reported in this article experimentally investigated this link between two broad classes of culturally widespread phenomena of interest to social science—**religious** beliefs and cooperative behavior among unrelated strangers.

Although anecdotes documenting religion's prosocial and antisocial effects abound, the empirical literature has produced mixed results regarding religion's role in prosocial behavior. In the research reported here, we examined the effect of **God** concepts specifically on selfish and prosocial behavior. Our research design was novel in two ways. First, we introduced an

Sosis and Ruffle (2004) examined levels of generosity in an experimental cooperative pool game in religious and secular kibbutzim in Israel and found higher levels of cooperation in the religious ones, and the highest levels among religious men who engaged in daily communal **prayer**. Batson and his colleagues (Batson et al., 1999; Batson, Schoenrade, & Ventis, 1993) have shown that although religious people report more explicit willingness to care for others than do nonreligious people, controlled laboratory measures of altruistic behavior often fail to corroborate this difference. Furthermore, when studies demonstrate that helpfulness is higher among more devoted people, this finding is typically better explained by egoistic motives such as seeking praise or avoiding guilt, rather than by higher levels of compassion or by a stronger motivation to benefit other people.

However, insightful findings are, research on religion and prosocial behavior has been limited by its overreliance on **conditional** designs. If religiosity and prosocial behavior are found to be correlated, it is just as likely that having a prosocial disposition causes one to be religious, or that some third variable such as guilt proneness or dispositional empathy causes both cooperative behavior and religiosity, or that religious beliefs somehow cause prosocial behavior. Only rarely have studies induced **supernatural** beliefs to examine whether a causal factor. Bering (2003, 2006) inhibited 3- to 6-year-old children's tendencies to cheat (i.e., open a "forbidden box") by telling them that an invisible agent ("Princess Alice") was in the room with them. In a different study, college students who were casually told that the ghost of a dead graduate student had been spotted in their private living room were less willing to cheat on a computerized **financial reasoning** task than were those told nothing (Bering, McLeod, & Shackelford, 2009). These studies suggest that explicit thoughts of supernatural agents curb cheating behavior.

In the research reported here, we examined the effect of **God** concepts specifically on selfish and prosocial behavior. Our research design was novel in two ways. First, we introduced an

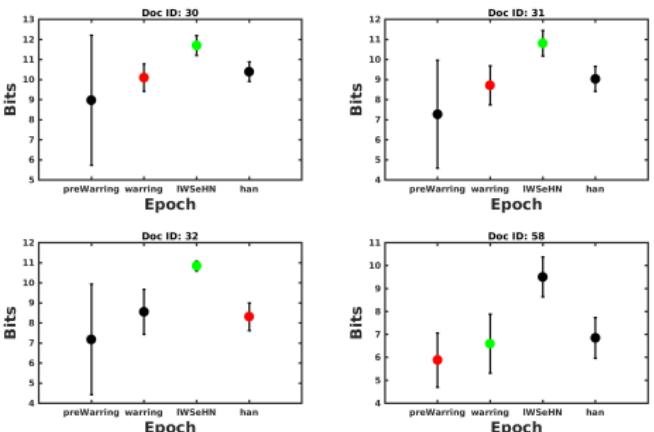
Original document

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	



Text Classification

- 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



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

Naive Bayes

Lesson 14: Idea
Analytics

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

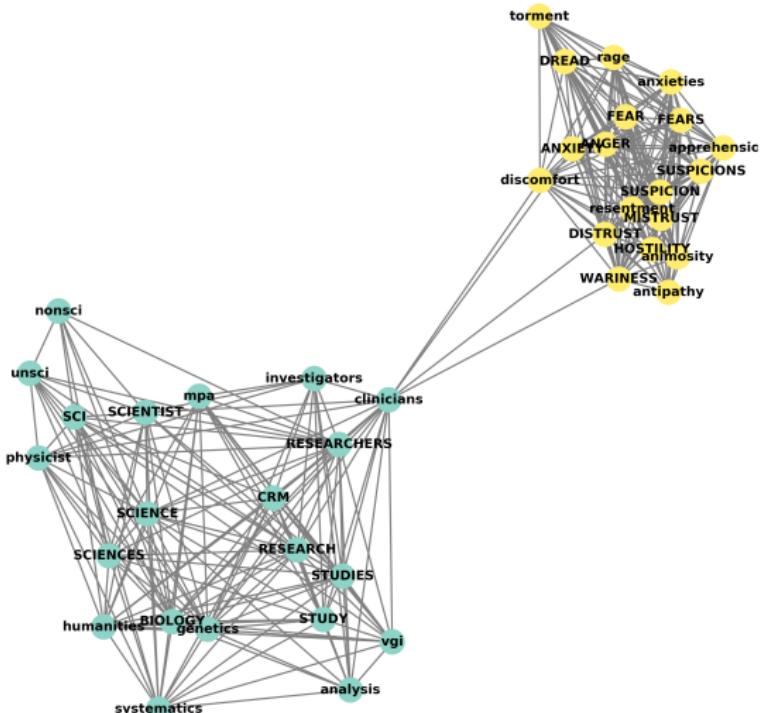
Classification
EDA
Neural embeddings



Exploratory analysis with Deep Neural Nets

Lesson 14: Idea
Analytics

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

Distributed word representation

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

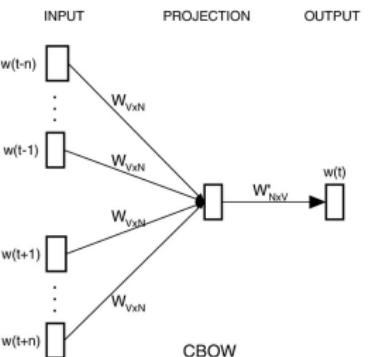


Figure 1: CBOW architecture for learning neural embeddings at the word level

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

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

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

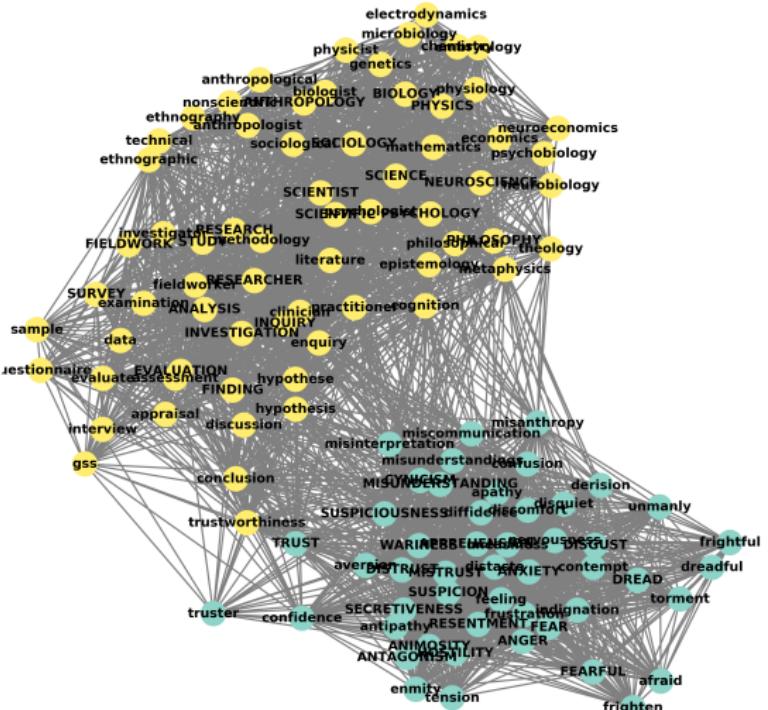


- Data*
- Data
- Data science
- Type of Data
- KDD
- Normalization
- Big Data?
- Human-in-the-Loop Models
- Example
- Trend Reservoirs
- Techniques
- Text analytics
- Word counts
- Word distributions
- Vector space model
- Sentiment analysis
- Non-negative Matrix Factorization
- Classification
- EDA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

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



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

Thank you for your attention

chcaa@cas.au.dk
chcaa.io

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

