

Domains of *mistrust*

Query expansion for *mistrust* using matrix decomposition & neural embeddings in the JSTOR digital library and the Danish net archive

Lasse Hansen & Kristoffer L Nielbo

chcaa.au.dk

chcaa.io

Center for Humanities Computing
Aarhus University, Denmark



CENTER FOR HUMANITIES
COMPUTING AARHUS



Outline

1 Dataset

JSTOR digital library
Danish net archive

2 Model 1

Non-negative Matrix Factorization
Keyword sets

3 Model 2

Neural embeddings
Sparse graph
Dense graph

Domains of *mistrust*

Lasse Hansen &
Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Dataset

JSTOR digital library
Danish net archive

Model 1

Non-negative Matrix
Factorization
Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph



Lasse Hansen &
Kristoffer L. Nielbo
chcaa.au.dk
chcaa.io

Dataset

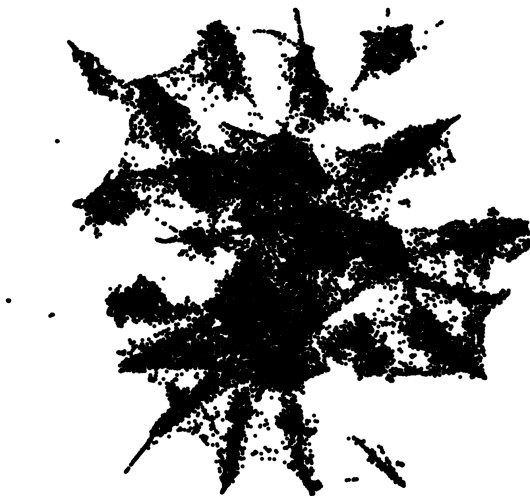
JSTOR digital library
Danish net archive

Model 1

Non-negative Matrix
Factorization
Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph



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



Lasse Hansen &
Kristoffer L. Nielbo
chcaa.au.dk
chcaa.io

Dataset

JSTOR digital library

Danish net archive

Model 1

Non-negative Matrix

Factorization

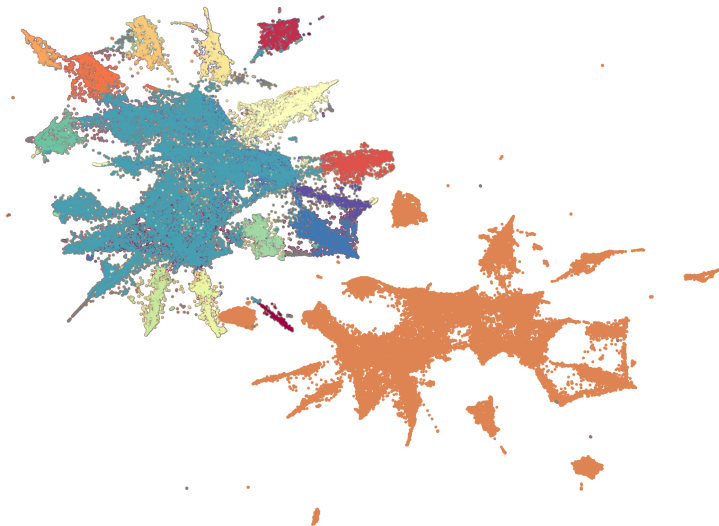
Keyword sets

Model 2

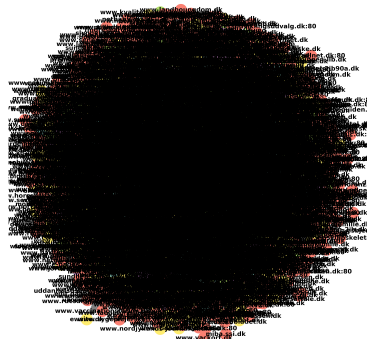
Neural embeddings

Sparse graph

Dense graph



- Queried the Danish net archive [MFR, MMR, klimaforandringer, climate change]
- “experiments in the wild” → real-world data with almost no signal
- problem is ill-defined, for ++noisy data, formulate complex conditional queries
- de-duplication and kwic extraction: 26gp → 35mb
- $S/N = 0.0013$
- for *MFR tilld* (trust) occurs 131 times and *mistillid* (mistrust) 23 times



Semantic graph *MFR* query (top 1000 sites)
in the Danish net archive

Dataset

JSTOR digital library

Danish net archive

Model 1

Non-negative Matrix

Factorization

Keyword sets

Model 2

Neural embeddings

Sparse graph

Dense graph



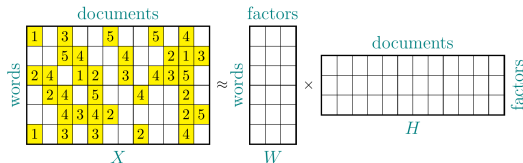
Matrix decomposition of *TD* matrix

Domains of *mistrust*

Lasse Hansen &
Kristoffer L. Nielbo
chcaa.au.dk
chcaa.io

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

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



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

Dataset

JSTOR digital library

Danish net archive

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings

Sparse graph

Dense graph

Keyword sets for *MISTRUST*

```

1 * social research science trust amp theory group model data individual
2 man say life think love thing world god word human
3 * health patient medical care service treatment public mental report dr
4 woman female gender male sexual man sex marriage mother wife
5 court law supra id legal justice lawyer federal note criminal
6 la le les et des que en du french paris
7 student school teacher education educational college teach program university class
8 war military peace united british american international states foreign army
9 soviet russian russia union communist europe policy western foreign west
10 china chinese japan japanese asia asian relation east foreign communist
11 book history pp chapter century author reader text university review
12 black african white africa racial american race south americans ethnic
13 church religious catholic god religion christian faith muslim century england
14 india indian pakistan muslim asia south country asian native region
15 community local land city area environmental water village people forest
16 german und der germany die vo des europe french european
17 israel jewish arab jew muslim peace east al middle conflict
18 political party government democracy election politics democratic social power public
19 child family parent mother father school social amp marriage young
20 film image art narrative director story character media american production
21 poem love literary language line reader write word text read
22 economic market country policy price economy trade bank industry investment
23 music art performance play style voice hear culture movement popular
24 worker union labor employee management labour industrial social company industry
25 nuclear weapon security arm energy pakistan military strategic international treaty

```

*) factors used to subset the jstor data

Dataset

JSTOR digital library
Danish net archive

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph



Keyword sets for MFR

Domains of *mistrust*

Lasse Hansen &
Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Dataset

JSTOR digital library
Danish net archive

Model 1

Non-negative Matrix
Factorization
Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph

```
1 0.045*"indlæg" + 0.021*"mar" + 0.017*"mfr" + 0.015*"første" + 0.014*"juni" + 0.013*"blog"
2 + 0.012*"vaccinen" + 0.010*"danmark" + 0.009*"sygdom" + 0.009*"bruger" + 0.009*"tak"
3 + 0.009*"myths" + 0.009*"holder" + 0.009*"læger"
4
5 0.089*"se" + 0.033*"mfr" + 0.012*"mere" + 0.011*"kl" + 0.010*"vaccinen" + 0.009*"u"
6 + 0.009*"mar" + 0.009*"mæslinger" + 0.009*"nedenfor" + 0.009*"universal"
7 + 0.008*"sundhedsstyrelsen" + 0.008*"marts" + 0.007*"bivirkninger" + 0.006*"dk"
8
9 0.042*"leje" + 0.039*"år" + 0.034*"klinik" + 0.031*"tilslutning" + 0.022*"københavn"
10 + 0.019*"mdr" + 0.016*"vaccination" + 0.015*"serum" + 0.015*"statens" + 0.015*"nyt"
11 + 0.015*"institut" + 0.014*"måneder" + 0.014*"region" + 0.014*"vaccinemøde"
12
13 0.074*"mfr" + 0.014*"universal" + 0.010*"karndybel" + 0.010*"skrue" + 0.010*"sb"
14 + 0.010*"tx" + 0.010*"universalplugs" + 0.010*"hbr" + 0.009*"fzb" + 0.008*"hulstensplugs"
15 + 0.007*"plugs" + 0.007*"uden" + 0.007*"ved" + 0.007*"fzv"
16
17 0.043*"sundhed" + 0.028*"børn" + 0.028*"mere" + 0.027*"dk" + 0.026*"blevet" + 0.025*"igen"
18 + 0.025*"fundet" + 0.025*"kvinder" + 0.025*"sygdomme" + 0.025*"punktet" + 0.025*"sex"
19 + 0.025*"parforhold" + 0.025*"gratis" + 0.025*"brevkasser"
20
21 0.043*"mfr" + 0.036*"vaccinen" + 0.017*"stemplet" + 0.017*"livsstil" + 0.016*"andet"
22 + 0.016*"digitalt" + 0.016*"buler" + 0.016*"nheden" + 0.015*"kokken" + 0.015*"godtfolk"
23 + 0.015*"tjeneren" + 0.015*"movie" + 0.015*"klovn" + 0.015*"nordjylland"
24
25 0.093*"vaccine" + 0.036*"mfr" + 0.011*"danmark" + 0.011*"sygdom" + 0.011*"nej"
26 + 0.009*"mindst" + 0.009*"effektiv" + 0.009*"danskere" + 0.009*"farlig" + 0.009*"bruger"
27 + 0.009*"brugerbetaling" + 0.009*"ramt" + 0.009*"diarré" + 0.009*"kræftbehandling"
28
29 0.083*"mfr" + 0.060*"polio" + 0.059*"mdr" + 0.058*"kighoste" + 0.055*"difteri" + 0.050*"pcv"
30 + 0.047*"tetanus" + 0.042*"hib" + 0.023*"hvp" + 0.022*"piger" + 0.021*"revaccination"
31 + 0.014*"ja" + 0.014*"babylex" + 0.008*"stivkrampe"
32
33 0.044*"mfr" + 0.015*"børn" + 0.013*"nyt" + 0.012*"vaccinen" + 0.011*"voksne"
34 + 0.010*"gratis" + 0.010*"bedste" + 0.010*"koster" + 0.010*"opfylder" + 0.010*"nordisk"
35 + 0.010*"novo" + 0.010*"vaccine" + 0.010*"kbu" + 0.010*"conflict"
36
37 0.046*"mæslinger" + 0.037*"mfr" + 0.037*"gratis" + 0.036*"røde" + 0.036*"hunde"
38 + 0.033*"fåresyge" + 0.022*"vaccinen" + 0.017*"voksne" + 0.016*"april" + 0.015*"unge"
39 + 0.009*"blog" + 0.008*"sundhedsstyrelsen" + 0.008*"alvorlige" + 0.008*"adverse effects"
```

*) factors used to subset the Danish net archive data

Utilization of document representations

Domains of *mistrust*

Lasse Hansen &
Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Dataset

JSTOR digital library

Danish net archive

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings

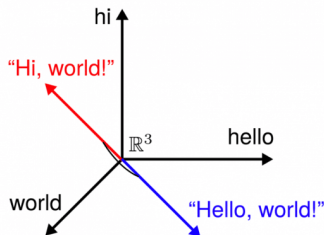
Sparse graph

Dense graph

Using H each document d can be represented as a dense vector of factor weights $[w_1, w_2, \dots, w_n]$ and we can measure the similarity between any pair of documents as

$sim(d_j, d_k) = \frac{d_j \cdot d_k}{|d_j| |d_k|}$, notice that the denominator is impacted by document length. In the study, source documents are normalized to keywords in context *query* ± 20 tokens, e.g.,

$t_{-20}, t_{-19}, \dots, t_{-1}$ MFR $t_1, t_2, \dots, t_{20}$



Geometric interpretation of document similarity in TD matrix



Semantic website comparison #3

Domains of *mistrust*

Lasse Hansen &
Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Dataset

JSTOR digital library

Danish net archive

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings

Sparse graph

Dense graph

image of vaccine HERE

MFR in the context of vaccines

image of turning insert HERE

MFR in the context of (carbide) turning
inserts



Distributed word representation

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

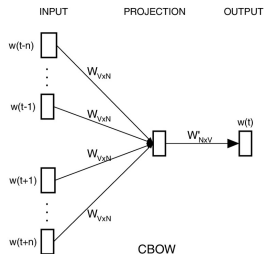


Figure 1: CBOW architecture for learning neural embeddings at the word level

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

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

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



Dataset

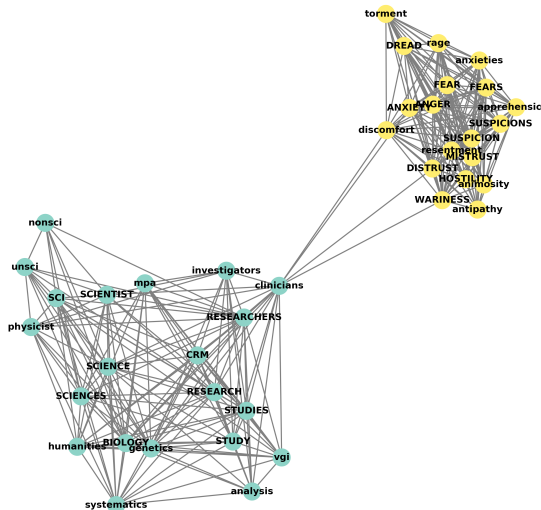
JSTOR digital library
Danish net archive

Model 1

Non-negative Matrix
Factorization
Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph



Sparse query graph for MISTRUST and SCIENCE





Keywords on sparse MFR graph

Domains of *mistrust*

Lasse Hansen &
Kristoffer L Nielbo
chcaa.au.dk
chcaa.io

Dataset

JSTOR digital library

Danish net archive

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings

Sparse graph

Dense graph

TRUST

$Tillid_{MFR}^{c1}$: tiltro (**trust**), pligt (**obligation**), kendskab (**knowledge**), holding (**attitude**)

$Tillid_{MFR}^{c2}$: familielæge (**family physician**), økonomi (**economy**), vaccineindustri (**vaccine industry**), arbejdsskadestyrelse (**national board of industrial injuries**)

MISTRUST

$Mistillid_{MFR}^{c1}$: aflive (**put down**), afvise (**reject**), forbindelse (**connection**), hensyn (**consideration**)

$Mistillid_{MFR}^{c2}$: konklusion (**conclusion**), forsker (**researcher**), årsag (**cause**), nabo (**neighbor**)



Dataset

JSTOR digital library

Danish net archive

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings

Sparse graph

Dense graph

Thank you for your attention

chcaa@cas.au.dk

chcaa.io

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

