

Scientific domains of *mistrust*

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

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
- 2 Model 1
 - Non-negative Matrix Factorization
 - Keyword sets
- 3 Model 2
 - Neural embeddings
 - Sparse graph
 - Dense graph

Scientific domains of
mistrust

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

Dataset

JSTOR digital library

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings

Sparse graph

Dense graph



Dataset

JSTOR digital library

Model 1

Non-negative Matrix
Factorization
Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph



Figure 1: JSTOR Data For Research, $n = 43,802$, ~ 400 m words, Eng*, Fr, Ger.



Scientific domains of
mistrust

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

Dataset

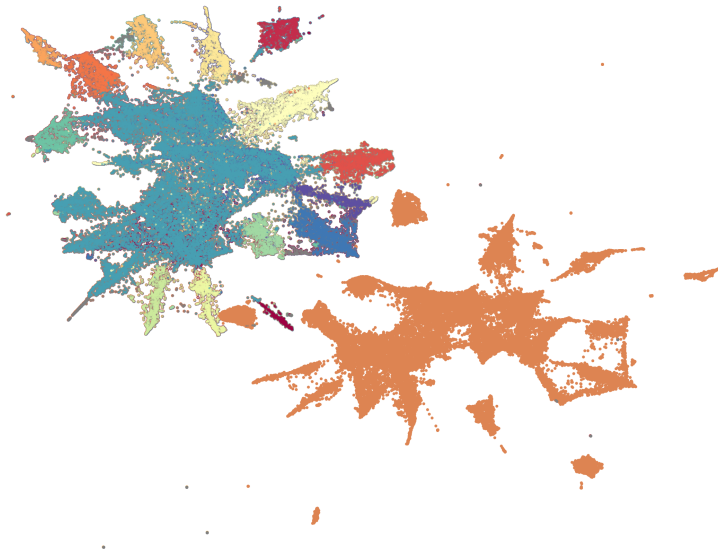
JSTOR digital library

Model 1

Non-negative Matrix
Factorization
Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph



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

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

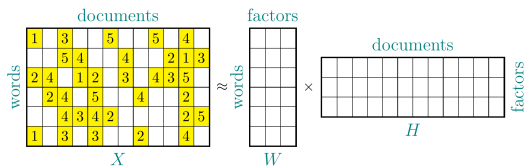


Figure 2: Non-negative matrix factorization of *TD* matrix X

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

[NOTE] because the data set was seeded with *MISTRUST*, the model is de facto seeded

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

```

Dataset

JSTOR digital library

Model 1

Non-negative Matrix
Factorization

Keyword sets

Model 2

Neural embeddings
Sparse graph
Dense graph

*) factors used to subset the jstor data



Distributed word representation

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

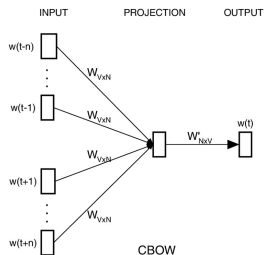


Figure 3: CBOW architecture for learning neural embeddings at the word level

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

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

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



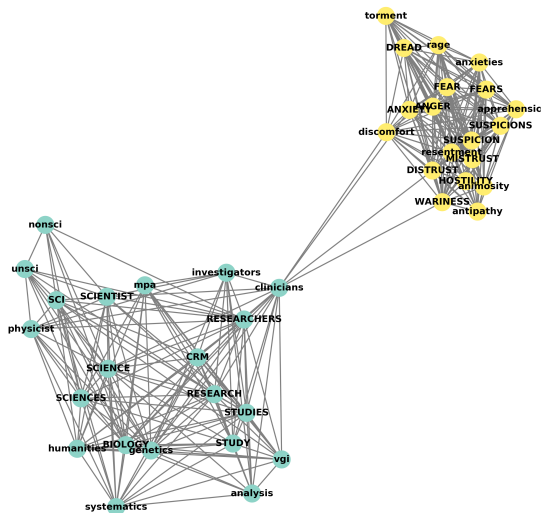


Figure 4: Sparse query graph for MISTRUST and SCIENCE



Dataset

JSTOR digital library

Model 1

Non-negative Matrix

Factorization

Keyword sets

Model 2

Neural embeddings

Sparse graph

Dense graph

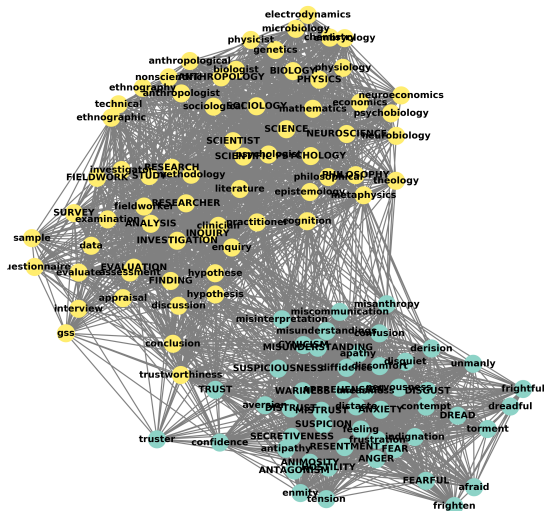


Figure 5: Dense query graph for MISTRUST and SCIENCE



Thank you for your attention

chcaa@cas.au.dk
chcaa.io

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

