Topic Modelling for Search

Data Science Lab
Fall 2020 edition

By Andreas Opedal, Daniel Garellick and Giulia Lanzillotta

Advisors: Paul Cross, Tarun Chadha and Christine Khammash, Ce Zhang

The context

We took part in the **ETH Search next generation** project

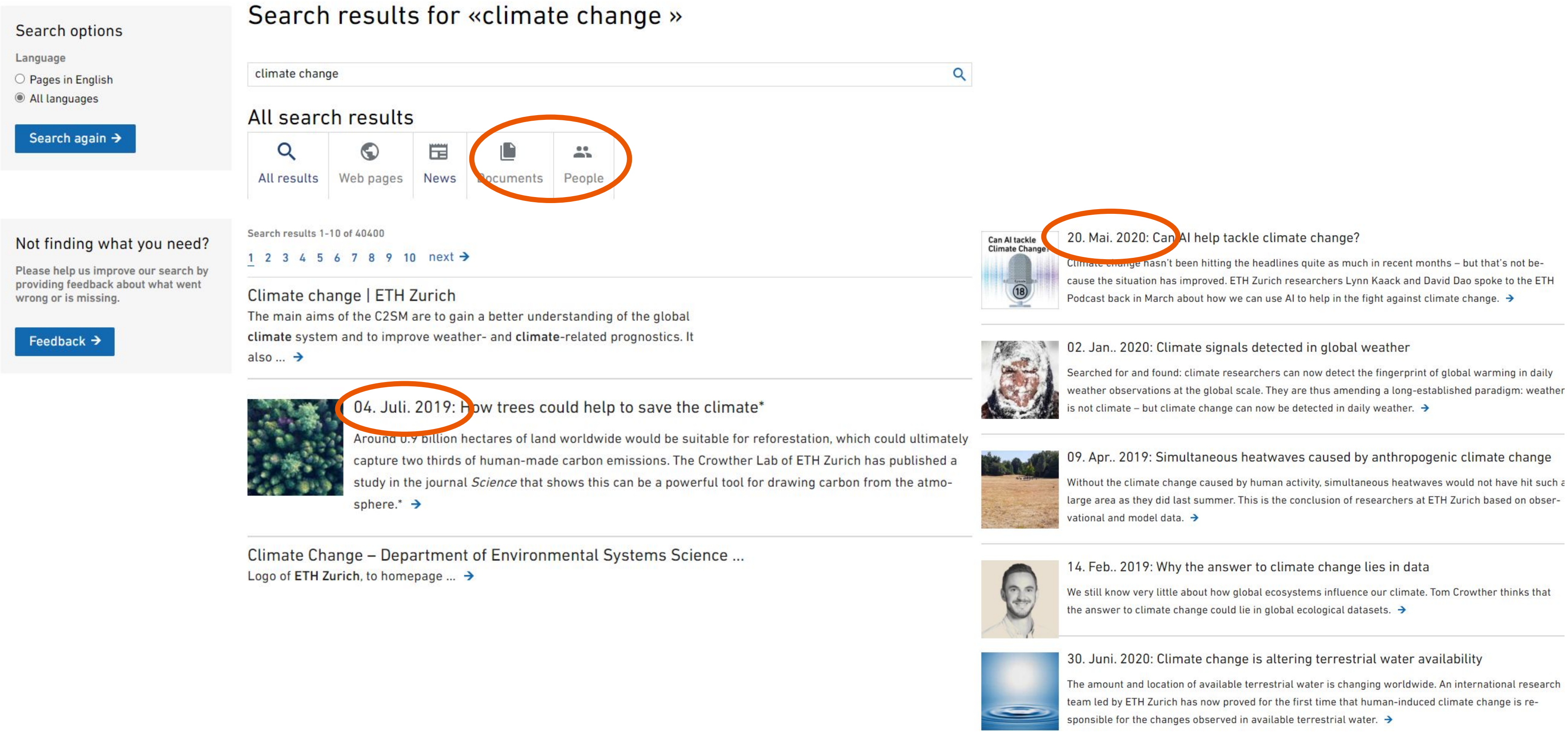
Final goal: improve the *ethz.ch* search experience

Our project targets:

- Unlock information present in the Research Collection
- Identify and link publications, researchers and research areas
- Generate valuable insights for the ETH research community

Results from the current search engine

A focus on what we believe is missing today



Snapshots from [Search | ETH Zurich]

Results from our system

Our system is meant to be a component of the search engine

Focused on improving the quality of search results for a given query

Query = *"Climate change"*

I lead the climate physics group and do research and teaching on many topics related to climate change.

These include long term projections, scenarios, the 2°C target, uncertainties in projections, climate model evaluation, model weighting, natural climate variab-

Most relevant publications

"The sensitivity of the modeled energy budget and hydrological cycle to CO_2 and solar forcing"

"Uncertainty partition challenges the predictability of vital details of climate change"

ETH Experts in the field

1. Lohmann, Ulrike
2. Knutti, Reto
3. Gruber, Nicolas
4. Peter, Thomas
5. Buchmann, Nina

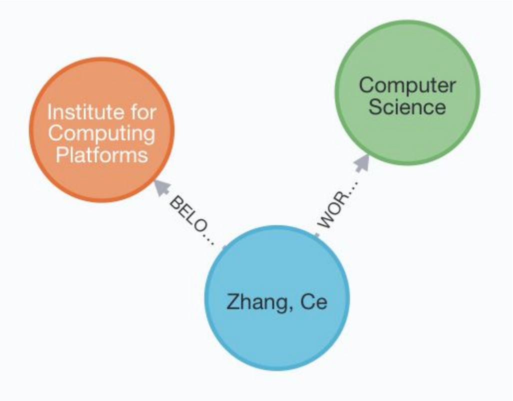
Related research areas

1. ice climate
2. climate precipitation
3. climate change cost
4. species biodiversity
5. precipitation water
6. air climate seismic

*The above is a reconstruction of the results obtained from different components of our system

Query = *"Ce Zhang"*

Department & Organisation



Recent Publications

"CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks"

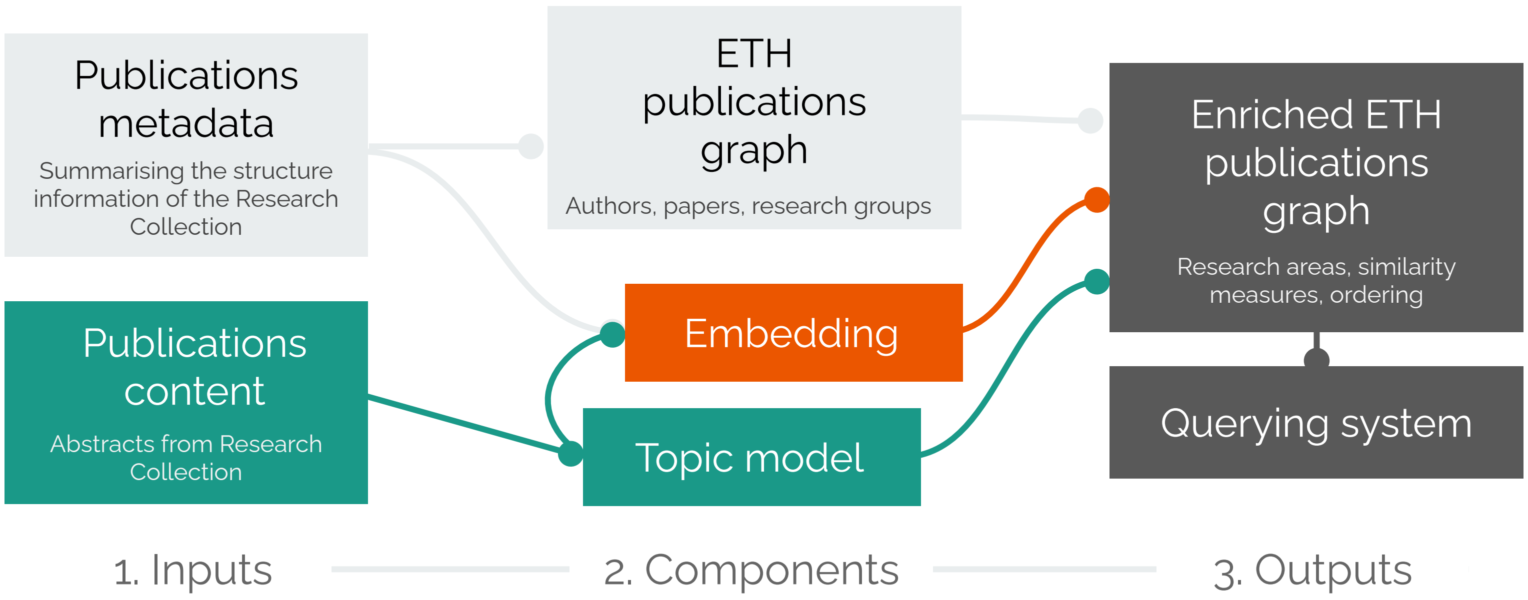
"Distributed Learning Systems with First-Order Methods"

"ColumnSGD: A Column-oriented Framework for Distributed Stochastic Gradient Descent"

Common Collaborators

1. Wu, Wentao - 10
2. Liu, Ji - 10
3. Schawinski, Kevin - 9
4. Cui, Bin - 8

Summary of technical approach



Data Pipeline

Initial data

Data sources:

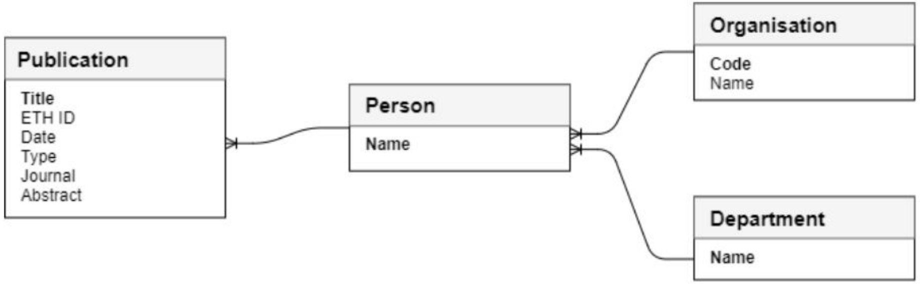
- Research collection 1930-2020
- Professor list
- Organisations mapping
- ETH website content (web crawl result)
- ETH search logs

113 initial attributes only from the first source
124 MB in zipped version

Multiple restructurings over the years and *data entry not performed centrally* caused:

- Half-empty fields
- Duplicated attributes
- Multiple non-unique identifiers
- Intrinsic variability in the attribute values

Filtering



Information extracted:

267877 authors
170284 publications
17 departments
383 organisations

21421 publications with abstracts in english (Conference papers, journal papers, book chapters, ...)

Cleaning

Main goal of cleaning operation: unify the publication date and author fields to a single format.

Fundamental for successful data integration step.

Date: reduced to *YYYY* format
Name: reduced to *Name, First Name*

Starting point: **Coverage = 49086**

```
def separate_names(names):
    """ Separes a string of names of the form name1||name2||name3||... into a list of names."""
```

```
org_df["Professor"] = org_df["Name"] + ", " + org_df["First name"]
```

End point: **Coverage = 68999**

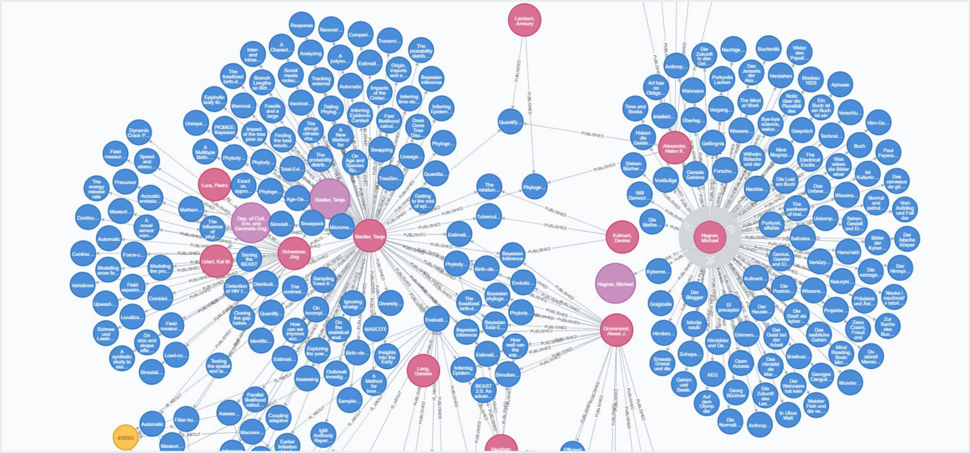
Integrate different sources of organisational data

```
def delete_parenthesis(name):
    if isinstance(name, str):
        return re.split('(\s\([a-zA-Z.]+\))', name)[0]
```

```
delete_parenthesis('Baccini, Peter (em.)')
```

```
Baccini, Peter
```

The ETH publications graph



4 + 2 nodes

3 + 3 relationships

11 + 7 attributes

Graph data model

Modelling

Topic Model
Requirements

1. High performance
2. Lifelong/online learning
3. Not fixed # of topics
4. Automatic topic name inference
5. Topic correlation & hierarchy

Models implemented:

- LDA (Latent Dirichlet Allocation)
- CTM (Correlated Topic Model)
- PAM (Pachinko Allocation Model)
- HDP (Hierarchical Dirichlet Process)
- online-HDP
- ETM (Embedded Topic Model)

Preprocessing Text data

Always applied:

- Tokenisation
- Lower-casing
- Accent marks removal
- Stop-words removal

Varying depending on the model:

- Stemming
- Lemmatisation

Additionally: bigrams/trigrams

```
gensim.utils.simple_preprocess(doc, deacc=False, min_len=2, max_len=15)
```

Convert a document into a list of lowercase tokens, ignoring tokens that are too short or too long.

Uses `tokenize()` internally.

NLTK's list of english stopwords

```
i
me
my
myself
we
```

```
>>> plurals = ['caresses', 'flies', 'dies', 'mules', 'denied',
...            'died', 'agreed', 'owned', 'humbled', 'sized',
>>> singles = [stemmer.stem(plural) for plural in plurals]
>>> print(' '.join(singles))  # doctest: +NORMALIZE_WHITESPACE
caress fli die mule deni die agre own humbl size meet
```

```
>>> wnl = WordNetLemmatizer()
>>> print(wnl.lemmatize('dogs'))
dog
>>> print(wnl.lemmatize('churches'))
church
```

Metrics

- Log-likelihood
  - Quantity maximised during training
  - Not well-correlated with human judgement
- Perplexity
  - Proportional to log-likelihood
  - Easily interpretable

$$PP(p) := 2^{H(p)} = 2^{-\sum_x p(x)\log_2 p(x)}$$

- Coherence
  - Computationally intensive
  - Measures degree of semantic similarity between high scoring words in each topic
- Eyeballing topics
  - Infeasible
  - Optimal

Metrics

Final model selection and assessment process:

1. Grid search over hyperparameters
2. Top 3 models selected based on perplexity on test set
3. Compare best models by human judgment
4. Assess final performance with coherence

Baselines
Code base

All the code for the baseline models is based on the tomotopy APIs

Features:

- Slower convergence of the algorithm but faster iterations (iterates 20 times more than `gensim` but overall running time is still 5-10 times faster)
- Automatic parallelisation when run on multi-core CPUs
- Easy to use
- Qualitatively and quantitatively better results on LDA compared to `gensim`

tomotopy

`tomotopy` is a Python extension of `tomoto` (Topic Modeling Tool) which is a Gibbs-sampling based topic model library written in C++. It utilizes a vectorization of modern CPUs for maximizing speed. T

- Latent Dirichlet Allocation (`LDAModel`)
- Labeled LDA (`LLDAModel`)
- Partially Labeled LDA (`PLDAModel`)
- Supervised LDA (`SLDAModel`)
- Dirichlet Multinomial Regression (`DMRModel`)
- Generalized Dirichlet Multinomial Regression (`GDMRModel`)
- Hierarchical Dirichlet Process (`HDPModel`)
- Hierarchical LDA (`HLDAModel`)
- Multi Grain LDA (`MGLDAModel`)
- Pachinko Allocation (`PAModel`)
- Hierarchical PA (`HPAModel`)
- Correlated Topic Model (`CTModel`)
- Dynamic Topic Model (`DTModel`).

Baselines
LDA

Number of topics

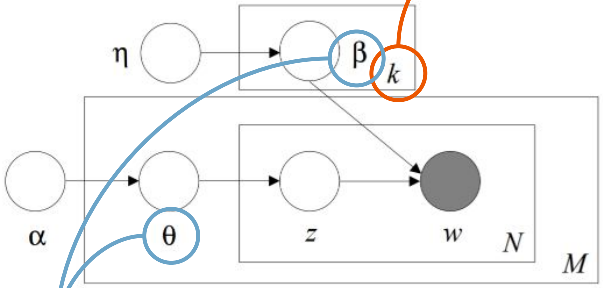


Illustration of the generative process.

Learnable
parameters

Requirements satisfied:

1. High performance
2. Lifelong/online learning
3. Not fixed # of topics
4. Automatic topic name inference
5. Topic correlation & hierarchy

Parameters to be tuned:

- **k** ∈ {50, 100, 150, 200, 300, 350, 450}
- ***α*** ∈ {10/k, 1/k, 0.1/k}
- **η** ∈ {10/w, 1/w, 0.1/w}

Sparsity assumption

Baselines
CTM

Correlated Topic model:

- Change in choice of prior distribution
- Allows discovery of correlations between topics
- Higher computational costs

Requirements satisfied:

1. High performance
2. Lifelong/online learning
3. Not fixed # of topics
4. Automatic topic name inference
5. Topic correlation & hierarchy

Parameters to be tuned:

- **k** $\in$ {50, 100, 150, 200, 300, 350, 450}
- **η** $\in$ {10/w, 1/w, 0.1/w}

Baselines
PA

Pachinko allocation model:

- Modelling a DAG of topics
- Words are leaf nodes, higher level topics as mixtures over topics
- Allows discovery of arbitrary hierarchy
- Explosion in number of parameters

Requirements satisfied:

1. High performance
2. Lifelong/online learning
3. Not fixed # of topics
4. Automatic topic name inference
5. Topic correlation & hierarchy

Parameters to be tuned:

- $\mathbf{k_1} \in \{0.2\,\mathbf{k_2}, 0.1\,\mathbf{k_2}, 0.05\,\mathbf{k_2}\}$
- $\mathbf{k_2} \in \{100, 150, 200, 300, 350\}$
- $\boldsymbol{\alpha} \in \{1/k_1, 0.1/k_1, 0.01/k_1\}$
- $\boldsymbol{\eta} \in \{10/w, 1/w, 0.1/w\}$

Baselines
Results

Error in CTM #81

Closed danielgarel opened this issue on 2 Nov 2020 · 3 comments

Solved in version 0.10.2 - released 6 days ago

Baselines
Results

Table 1: Baseline Grid Search results

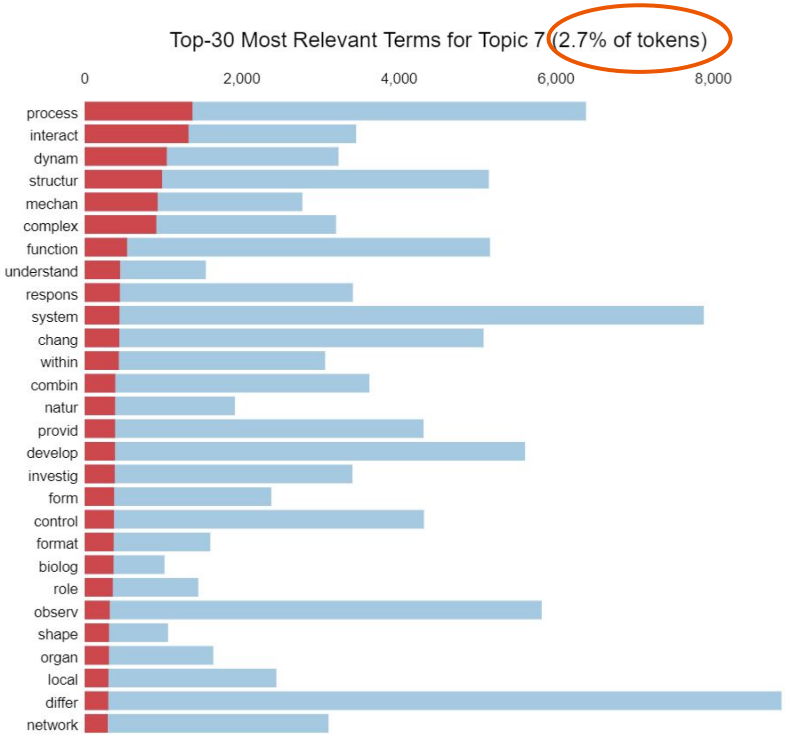
Model	Description	PP	Training time
LDA 1	$k = 450, \alpha = 1/k, \eta = 10/w$	1.000023416	461 s
LDA 2	$k = 450, \alpha = 10/k, \eta = 10/w$	1.000023435	461 s
LDA 3	$k = 350, \alpha = 0.1/k, \eta = 10/w$	1.000023437	361 s
PAM 1	$k_2 = 300, k_1 = k_2/20, \alpha = 0.01/k_1, \eta = 0.1/w$	1.00156366	4,975 s
PAM 2	$k_2 = 350, k_1 = k_2/5, \alpha = 0.01/k_1, \eta = 0.1/w$	1.00156373	34,144 s
PAM 3	$k_2 = 300, k_1 = k_2/5, \alpha = 0.01/k_1, \eta = 0.1/w$	1.00156426	23,042 s

LDA1

pyLDAvis tool for visualisation
(only for LDA)

High coherence

Small set of general topics and a constellation of narrow but sensible topics

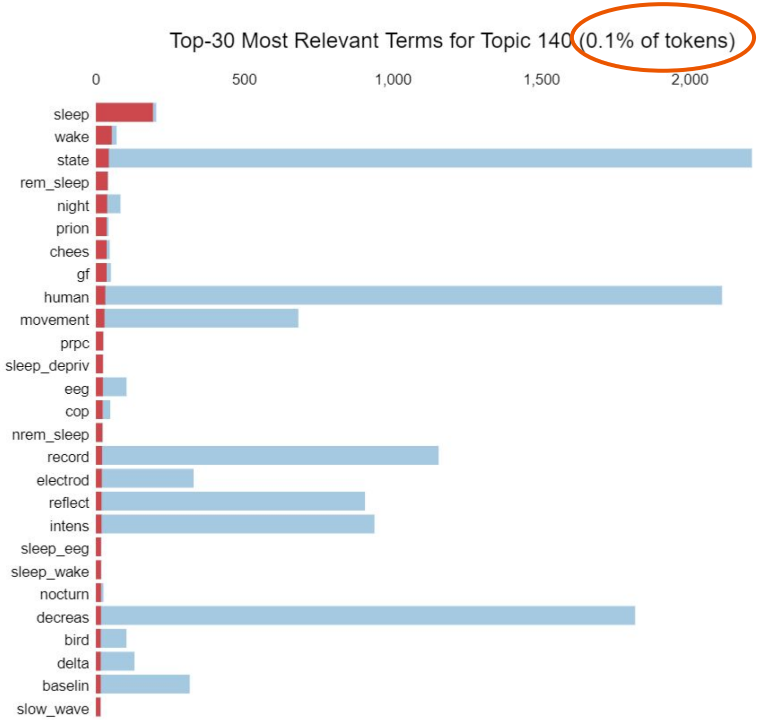


LDA1

pyLDAvis tool for visualisation
(only for LDA)

High coherence

Small set of general topics and a constellation of narrow but sensible topics



Experiments I
HDP and O-HDP

Target of experiments I: continual learning and flexibility in the number of topics

HDP

- HDP infers the number of topics from the data
- Highly sensible to the hyperparameters values

O-HDP

- Extension of HDP to online inference
- Better for larger datasets
- No flexibility in the vocabulary (adapted library code)

Experiments I
Results HDP

Low coherence

Noise

Estimate of the optimal number of topics 133/150 topics used

```
<Topics>
| #0 (2304) : differ stimul increas tremor impuls
| #1 (1129) : cr water bell age self_injuri
| #2 (2838) : particip vr two group compar
| #3 (19549) : treatment effect differ data infect
| #4 (1453) : agnp teamwork se electrod particl
| #5 (8421) : data map design research project
| #6 (1914) : chemic concentr effect system measur
| #7 (1357) : h differ explos crisi_map detect
| #8 (2735) : child rotor peak increas patient
| #9 (2831) : cell activ human effect product
| #10 (3096) : water alloy age increas coupl
| #11 (4133) : biofilm diffus snow water system
| #12 (2694) : cassava show transform product r
| #13 (25139) : gene plant protein genom differ
| #14 (1989) : fossil sampl estim speci differ
| #15 (1299) : particl tqi simul land_use trial
| #16 (3816) : effect snail temperatur individu trait
| #17 (16752) : neuron function cell network activ
| #18 (2418) : exposur ec insect trend differ
| #19 (1893) : read develop lago allerg differ
| #20 (1484) : ratt potenti correl migrant hf
| #21 (2095) : measur patient paint sampl date
| #22 (3908) : ferment yeast wine associ product
```

Experiments I
Results OHDP

Topics from 0 to 6 hardly distinguishable and generic

Most of the documents in the collection strongly or exclusively associated with one topic in the first 8

Low coherence and noise like in HDP for the remaining topics

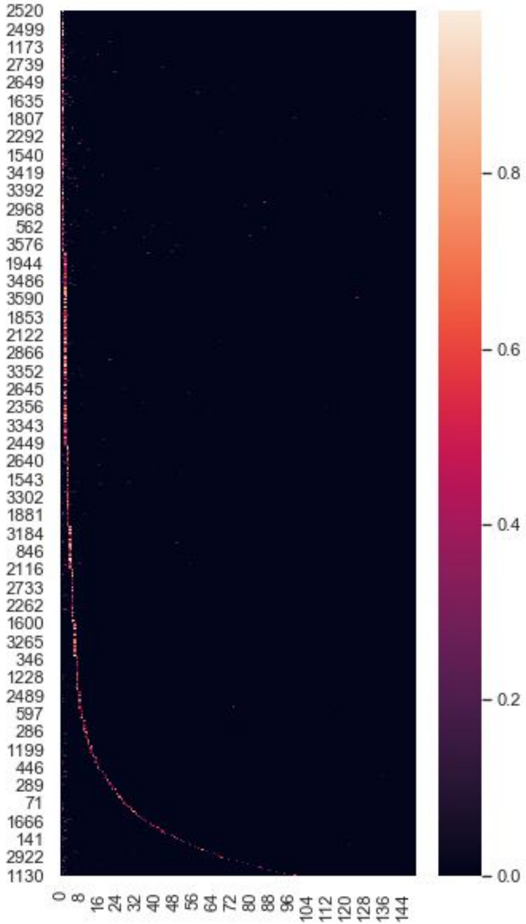
Experiment setup

1. Same hyperparameters used for HDP
2. Test both inference and update steps

```
Topic 0 -------
use - model - result - measur - system - base - data - studi - develop - differ - effect - method - user - test - design - incr
eas - process - ass - approach - show

Topic 1 -------
use - model - studi - data - measur - result - base - differ - effect - show - system - increas - process - develop - st - obse
rv - gener - also - compar - perform

Topic 2 -------
use - model - result - cpc - studi - concentr - chang - differ - activ - measur - base - effect - system - show - data - proces
s - ec - howev - present - method

Topic 38 -------
urban - analys - park - indic - studi - photograph - result - map - use - cdad - differ - measur - time - two - locat - pmmo -
definit - charg - pion - cancer - tumour

Topic 39 -------
product - qubit - effect - result - use - logic - input - voltag - growth - process - countri - show - sophist - ass - anomal -
isotop - taxat - overgraz - produc - orthorhomb - map

Topic 40 -------
base - high - iron - decomposit - neuropath - pain - algorithm - nerv - injuri - distribut - hemogen - object - three - transmi
ss - optim - iq - ferment - mouse - studi - age - mouse - progranulin - problem - price
```

Inference step results - performed on 70% of the training data

Experiments I
Results OHDP



Document to topics distribution heatmap

Inference step results performed on 70% of the data

Experiments I
Results OHDP

Topics from 0 to 6 stay the same

Substantial change at the semantic level for the rest of the topics

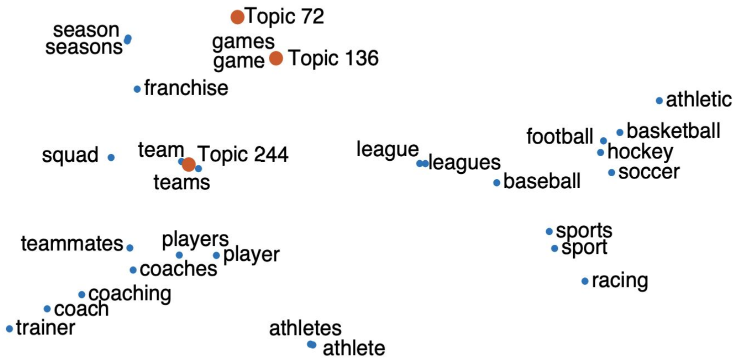
```
Topic 0 -------
use - model - result - measur - studi - data - differ - effect - system - base

Topic 1 -------
use - model - data - system - base - result - studi - differ - measur - effect

Topic 2 -------
use - model - base - data - measur - result - studi - differ - system - effect

Topic 38 -------
np - energi - system - chang - ion - cost - product - process - household - use

Topic 39 -------
hon - use - approxim - observ - format - heartwood - model - glacier - charg - particl - possibl

Topic 40 -------
use - protein - express - scatter - strain - glacier - shop - trip - control - ascent - region
```

Update step results - performed on 30% of the data
(same topics as in inference step shown above)

Experiments II
Embedding Approach

Target of experiments II: continual learning and flexibility in the number of topics

Embedding-based topic models incorporate topics in the vocabulary vector space



Dieng et al, 2019

ETM
- Generative process similar to LDA
- Words sampled according to proximity to topic vectors
- Deals with “heavy-tailed” vocabularies

Idea: run topic model inference on every streaming batch and integrate the results in the embedding space

Also obtain automatic topic-name inference

Experiments II

Word Embeddings

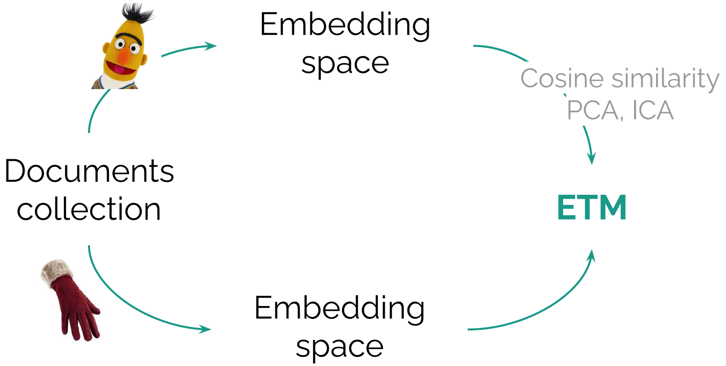
Learn with ETM or use pre-trained?
- Pre-trained embedding space for global use across batches

2 ways to feed (fixed) embeddings into ETM:

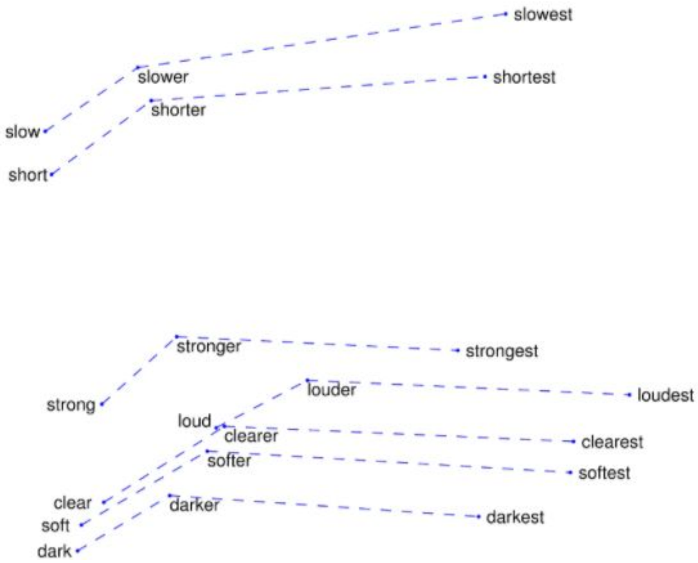
- Leverage contextualized embeddings (BERT)
- Static word embeddings (GloVe)

Experiments:

1. DistilBERT + cosine similarity filtering
2. DistilBERT + PCA reduction
3. GloVe



Expectations



Reality

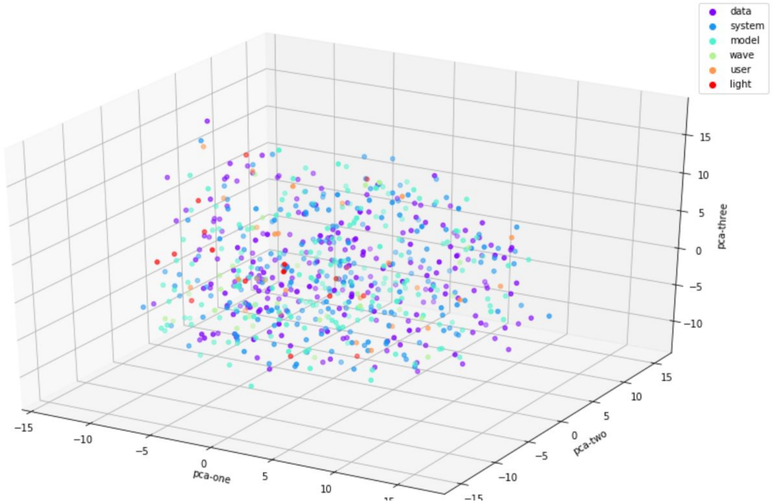


Image on the left taken from [GloVe: Global Vectors for Word Representation (stanford.edu)]

Experiments II

Results - word similarities

```
Visualize word embeddings by using output embedding matrix
word: insurance .. neighbors: ['mimic', 'daily', 'trace', 'cost', 'n', '(', 'determined', 'data', 'glacier', 'even']
word: weather .. neighbors: ['mimic', 'cost', 'n', 'daily', 'trace', 'progress', 'biology', 'negotiation', 'determined', 'even']
word: particles .. neighbors: ['daily', 'mimic', 'arab', 'n', 'trace', '(', 'runoff', '.', 'glacier', '.']
word: religion .. neighbors: ['mimic', 'question', 'interference', 'progress', 'trace', 'daily', '(', 'n', 'cost', 'negotiation']
word: man .. neighbors: ['mimic', 'n', 'cost', '(', 'negotiation', 'daily', 'energies', 'models', 'determined', 'even']
word: love .. neighbors: ['mimic', 'n', 'daily', 'cost', 'progress', 'negotiation', 'trace', '(', 'energies', 'determined']
word: intelligence .. neighbors: ['mimic', 'daily', 'trace', 'energies', 'n', 'arab', 'data', 'pathways', 'question', 'cost']
word: money .. neighbors: ['mimic', 'question', 'daily', 'n', 'cost', 'energies', '(', 'trace', 'interference', 'negotiation']
word: politics .. neighbors: [',', 'question', 'interference', '-', 'mimic', 'although', 'evaluate', 'daily', '-', 'n']
word: health .. neighbors: ['question', 'interference', 'flux', 'evaluate', ')', 'train', 'van', ',', 'chi', 'data']
word: people .. neighbors: ['mimic', 'n', 'daily', 'trace', '(', 'energies', 'negotiation', 'cost', 'determined', 'runoff']
word: family .. neighbors: ['mimic', 'n', '(', 'daily', 'cost', 'negotiation', 'determined', 'energies', 'trace', 'even']
```

Nearest neighbor visualisation of the embedding space for DistilBERT+PCA embedding

Experiments II

Results - word similarities

```
Visualize word embeddings by using output embedding matrix
word: insurance .. neighbors: ['insurance', 'insurers', 'premiums', 'insurer', 'pension', 'insured', 'care', 'savings', 'benefits', 'liability']
word: weather .. neighbors: ['weather', 'inclement', 'rain', 'temperatures', 'rainy', 'conditions', 'storms', 'winter', 'winds', 'rains']
word: particles .. neighbors: ['particles', 'particle', 'molecules', 'electrons', 'photons', 'subatomic', 'atoms', 'protons', 'droplets', 'microscopic']
word: religion .. neighbors: ['religion', 'religions', 'religious', 'christianity', 'beliefs', 'faith', 'belief', 'spirituality', 'catholicism', 'islam']
word: man .. neighbors: ['man', 'woman', 'person', 'boy', 'he', 'men', 'himself', 'one', 'another', 'who']
word: love .. neighbors: ['love', 'loves', 'passion', 'loved', 'romantic', 'lovers', 'lover', 'you', 'me', 'affection']
word: intelligence .. neighbors: ['intelligence', 'cia', 'information', 'security', 'counterterrorism', 'operatives', 'fbi', 'military', 'secret', 'spy']
word: money .. neighbors: ['money', 'funds', 'cash', 'fund', 'donations', 'pay', 'amount', 'paying', 'paid', 'millions']
word: politics .. neighbors: ['politics', 'political', 'politicians', 'religion', 'culture', 'ideology', 'partisan', 'liberal', 'debate', 'social']
word: health .. neighbors: ['health', 'care', 'healthcare', 'education', 'medical', 'hospitals', 'welfare', 'nutrition', 'benefits', 'social']
word: people .. neighbors: ['people', 'others', 'those', 'least', 'many', 'some', 'all', 'them', 'thousands', 'hundreds']
word: family .. neighbors: ['family', 'families', 'relatives', 'father', 'parents', 'mother', 'friends', 'daughter', 'son', 'wife']
```

Nearest neighbor visualisation of the embedding space for GloVe embedding

Experiments II
Results - topics

Two evolution patterns:

Collapse to single topic
Overlapping & noisy topics

```
Visualize topics...
Topic 0: ['proteins', 'protein', 'molecules', 'cells', 'receptor', 'membrane', 'receptors', 'enzyme', 'genes']
Topic 1: ['proteins', 'cells', 'protein', 'membrane', 'data', 'human', 'molecules', 'density', 'systems']
Topic 2: ['protein', 'proteins', 'membrane', 'molecules', 'receptor', 'density', 'cells', 'particles', 'acids']
Topic 3: ['proteins', 'protein', 'membrane', 'receptor', 'cells', 'molecules', 'receptors', 'particles', 'acids']
Topic 4: ['protein', 'proteins', 'membrane', 'cells', 'molecules', 'function', 'receptor', 'particles', 'electrons']
Topic 5: ['protein', 'proteins', 'molecules', 'receptor', 'cells', 'membrane', 'particles', 'enzyme', 'molecular']
Topic 6: ['proteins', 'protein', 'cells', 'particles', 'molecules', 'membrane', 'electrons', 'data', 'electron']
Topic 7: ['protein', 'proteins', 'receptor', 'membrane', 'molecules', 'cells', 'enzyme', 'function', 'rna']
Topic 8: ['proteins', 'protein', 'cells', 'molecules', 'membrane', 'acids', 'receptor', 'particles', 'molecular']
Topic 9: ['protein', 'proteins', 'membrane', 'molecules', 'receptor', 'cells', 'particles', 'density', 'molecular']
Topic 10: ['proteins', 'protein', 'membrane', 'cells', 'receptor', 'molecules', 'density', 'systems', 'system']
Topic 11: ['protein', 'proteins', 'receptor', 'membrane', 'molecules', 'cells', 'particles', 'acids', 'density']
Topic 12: ['protein', 'proteins', 'receptor', 'membrane', 'cells', 'molecules', 'enzyme', 'acids', 'genes']
Topic 13: ['proteins', 'protein', 'membrane', 'cells', 'molecules', 'receptor', 'enzyme', 'molecular', 'electrons']
Topic 14: ['proteins', 'protein', 'cells', 'molecules', 'membrane', 'receptor', 'systems', 'system', 'function']
Topic 15: ['protein', 'proteins', 'cells', 'molecules', 'membrane', 'density', 'receptor', 'system', 'particles']
Topic 16: ['protein', 'proteins', 'molecules', 'receptor', 'cells', 'membrane', 'data', 'enzyme', 'acids']
Topic 17: ['protein', 'proteins', 'cells', 'receptor', 'membrane', 'molecules', 'function', 'particles', 'systems']
Topic 18: ['proteins', 'protein', 'cells', 'membrane', 'molecules', 'receptor', 'function', 'tissue', 'particles']
Topic 19: ['proteins', 'protein', 'receptor', 'molecules', 'cells', 'membrane', 'systems', 'molecular', 'genes']
Topic 20: ['proteins', 'protein', 'cells', 'system', 'membrane', 'molecules', 'particles', 'density', 'temperature']
Topic 21: ['proteins', 'protein', 'molecules', 'membrane', 'cells', 'particles', 'density', 'receptor', 'electrons']
Topic 22: ['protein', 'proteins', 'cells', 'membrane', 'molecules', 'receptor', 'density', 'particles', 'molecular']
Topic 23: ['protein', 'proteins', 'cells', 'system', 'systems', 'membrane', 'molecules', 'particles', 'density']
Topic 24: ['proteins', 'protein', 'membrane', 'molecules', 'cells', 'receptor', 'electrons', 'electron', 'particles']
```

Experiments II
Results - topics

Two evolution patterns:

Collapse to single topic
Overlapping & noisy topics

```
Topic 0: ['membrane', 'proteins', 'molecular', 'species', 'protein', 'electron', 'equilibrium', 'organisms', 'molecules']
Topic 1: ['impedance', 'forewings', 'cauchy', 'polynomial', 'nucleotide', 'eukaryotic', 'lagrangian', 'density', 'capacitance']
Topic 2: ['receptor', 'protein', 'proteins', 'rna', 'extracellular', 'membrane', 'chromosome', 'molecules', 'mutations']
Topic 3: ['system', 'water', "n't", '-', 'surface', 'level', 'open', 'china', 'air']
Topic 4: ['endothelial', 'phenotype', 'synaptic', 'neuronal', 'capacitance', 'polynomial', 'metabolic', 'proteins', 'triglycerides']
Topic 5: ['-', 'system', 'foreign', 'country', "n't", 'level', 'low', 'high', 'economic']
Topic 6: ['protein', 'proteins', 'acids', 'tissue', 'molecules', 'membrane', 'layer', 'calcium', 'cells']
Topic 7: ['receptor', 'eukaryotic', 'neural', 'metabolic', 'density', 'protein', 'molecular', 'neuronal', 'trophic']
Topic 8: ['security', 'government', 'countries', 'iraq', 'weapons', "n't", 'people', 'measures', 'china']
Topic 9: ['surface', 'systems', 'temperature', 'electron', 'electrons', 'particle', 'particles', '-', 'system']
Topic 10: ['polynomial', 'paginated', 'coefficients', 'eukaryotic', 'non-linear', 'receptor', 'vowel', 'subunits', 'impedance']
Topic 11: ['countries', '-', "n't", 'human', 'make', 'china', 'people', 'system', '``']
Topic 12: ['tensor', 'membrane', 'bushel', 'bacterial', 'protein', 'proteins', 'necrosis', 'neuronal', 'isomorphic']
Topic 13: ['weapons', 'cells', 'countries', 'system', 'systems', '-', 'products', 'human', 'electron']
Topic 14: ['polynomial', 'protein', 'proteins', 'extracellular', 'sedimentary', 'eukaryotic', 'non-linear', '%', 'nonlinear']
Topic 15: ['system', 'nuclear', 'data', 'information', 'weapons', 'security', 'military', 'systems', 'nato']
Topic 16: ['transmembrane', 'polynomial', 'receptor', 'membrane', 'proteins', 'tensor', 'hamiltonian', 'amino', 'receptors']
Topic 17: ['good', 'countries', 'make', "n't", 'system', 'air', 'level', '-', "'ve"]
Topic 18: ['protein', 'membrane', 'particle', 'proteins', 'electrons', 'particles', 'electron', 'neutron', 'molecules']
Topic 19: ['-', 'information', 'system', 'high', 'data', 'systems', 'human', '``', 'energy']
Topic 20: ['coxeter', 'polynomial', 'extracellular', 'baronetcies', 'eigenvalues', 'formula_15', 'formula_2', 'transmembrane', 'subunit']
Topic 21: ['density', 'diameter', 'equations', 'protein', 'gravitational', 'membrane', 'polynomial', 'taxonomic', 'extracellular']
Topic 22: ['forewings', 'morphological', 'paginated', 'impedance', 'gradient', 'equations', 'nonlinear', 'nucleotide', 'phylogenetic']
Topic 23: ['polynomial', 'extracellular', 'intracellular', 'phenotype', 'receptor', 'membrane', 'tensor', 'proteins', 'necrosis']
Topic 24: ['non-linear', 'paginated', 'extracellular', 'necrosis', 'polynomial', 'sedimentary', 'equations', 'ecoregions', 'taxonomic']
```

Final model and querying

How to make use of our model to get better search results

Final Model: Streaming LDA

Beyond Topic Modelling

Tested topic models had limitations:
→ Augment using data already in graph!

LDA outperforms other models
→ Train LDA on batches of ~5000 publications
→ Use **graph and embeddings** to link topics across batches

Final Model: Streaming LDA -

LDA outperforms other models

→ Train LDA on batches of ~5000 publications

Only assumption:
Fixed number of topics per batch

- High performance
- Lifelong/online learning
- Not fixed # topics
- Automatic topic name inference
- Topic correlation & hierarchy

Component 1 - Graph

Move topics and words into graph

Matching scores on the edges (experts)

Can use least-cost-path algorithms to match complicated relational patterns

Component 2 - Embedding Space

Move graph into embedding space

Leverage embedding vocabulary

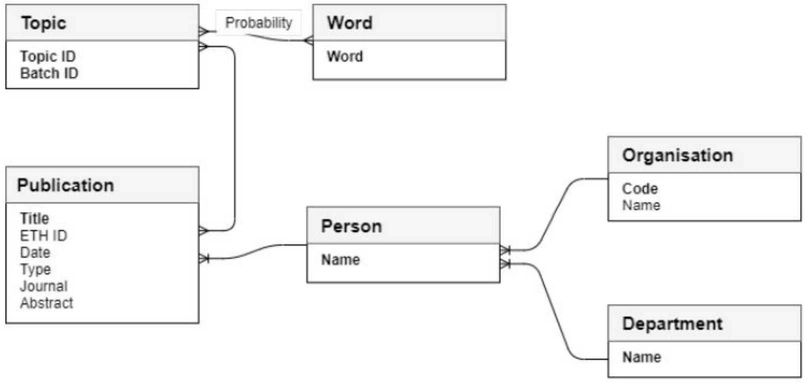
Enjoy flexibility in term-matching and similarity measures

Enriched graph

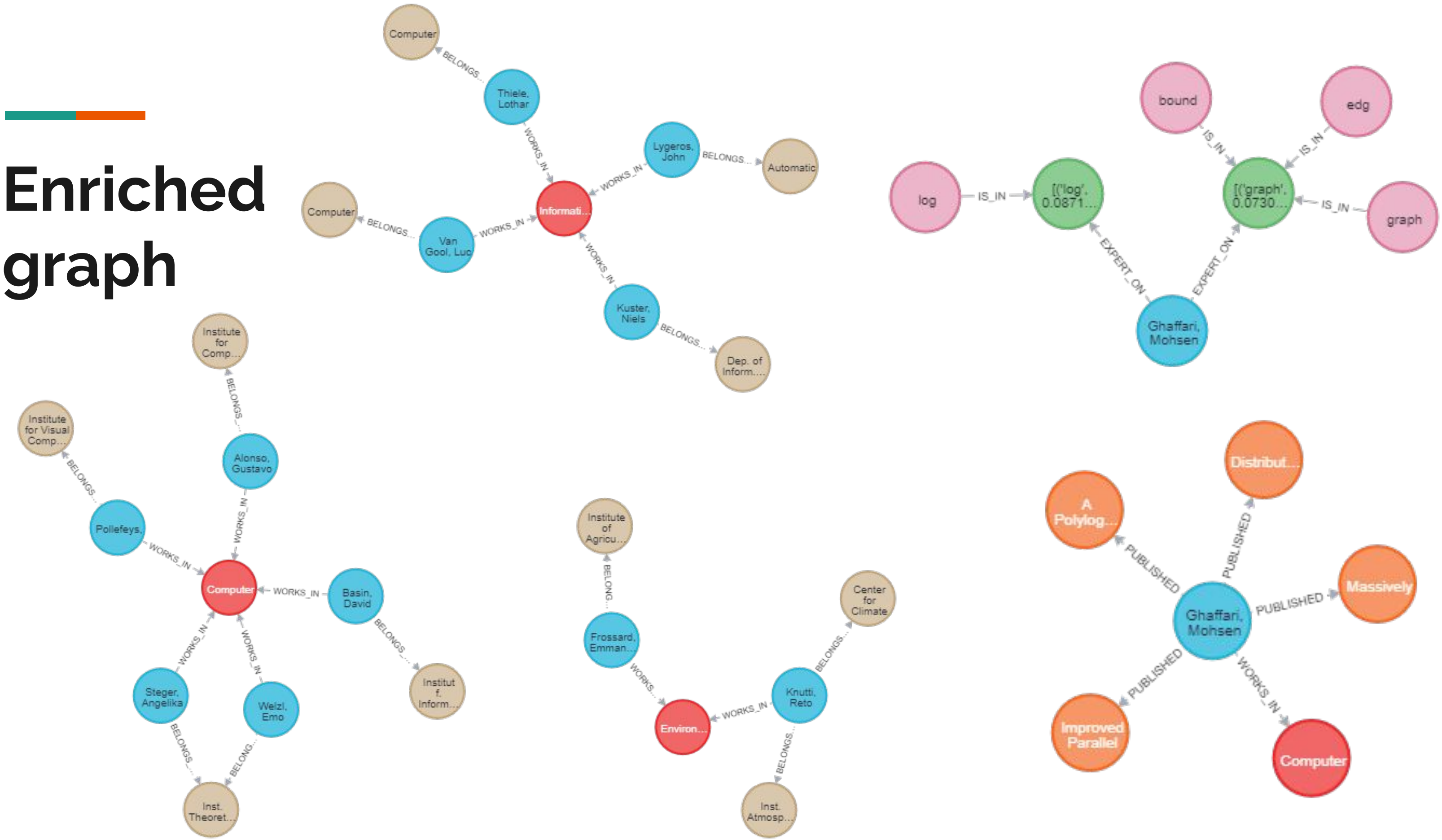
267,877 Person nodes
170,284 Publication nodes
17 Department nodes
383 Organisation nodes
500 Topic nodes
5,057 Word nodes

1,100,000 Relationships

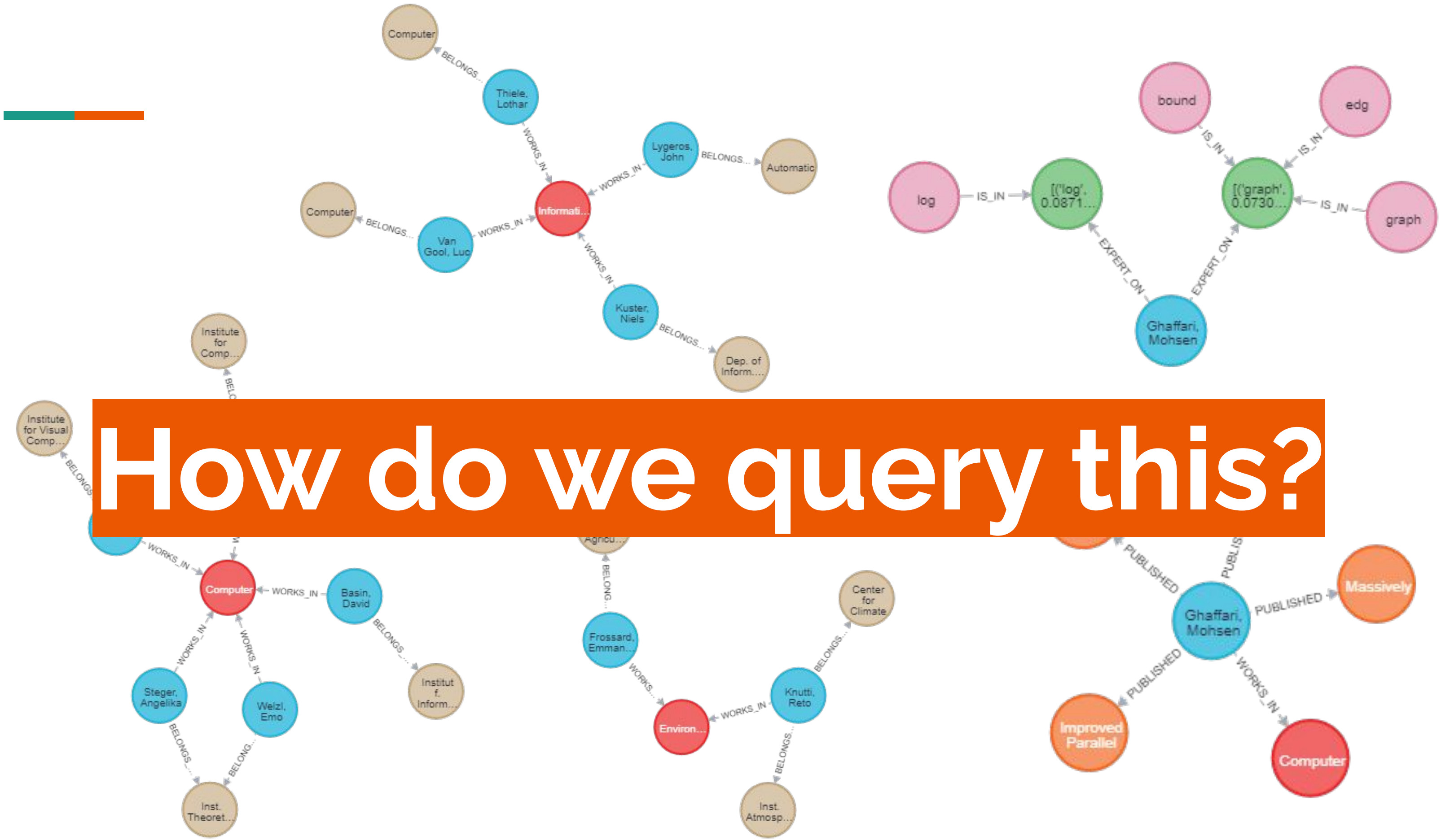
⟶ 285MB



Enriched graph

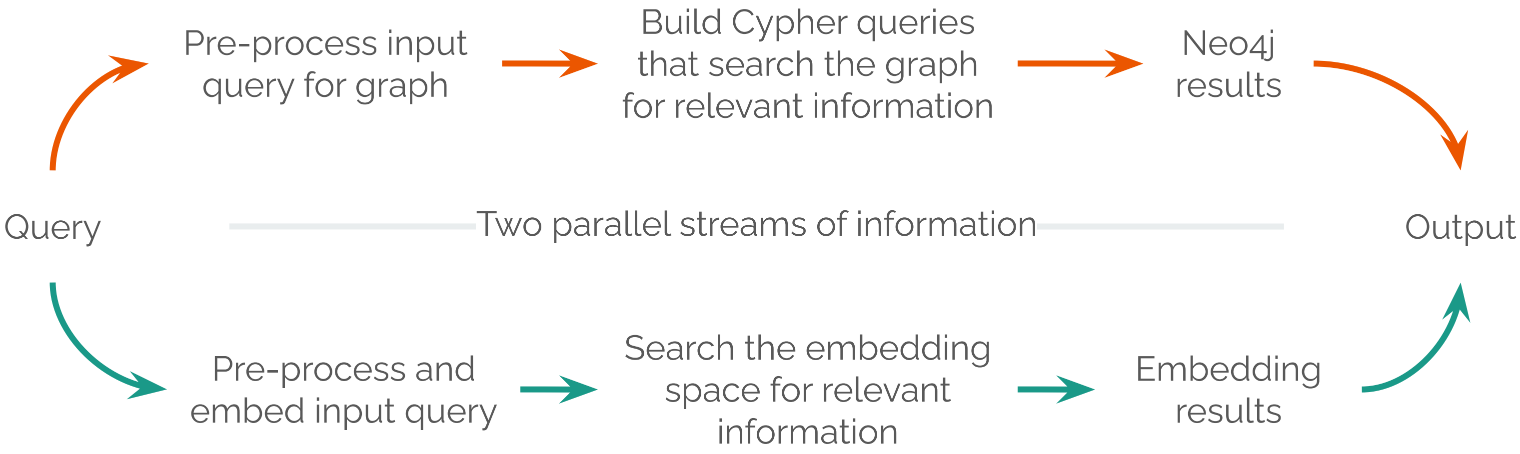


How do we query this?



The query pipeline

A bird's eye view



The query pipeline

Example

User query: “*climate change*”

The Graph side

"Experts"

Score function to quantify expertise

⟶ Recency of publications

⟶ Quantity of publications

⟶ Penalty for general topics

$$r_p = 1 - \frac{\texttt{publication year}}{2020 - 1930},$$

$$tsw = \texttt{count}(w_t)$$

$$tsp = \texttt{count}(p_t),$$

$$a_t = \begin{cases} 1, & \text{if } \texttt{count}(p_{a,t}) \geq 10 \\ \left(\frac{\texttt{count}(p_{a,t})}{10}\right)^{0.75} & \text{otherwise,} \end{cases}$$

$$\texttt{EXPERT\_ON weight} = \texttt{AVG}(r_p \cdot p_t) \times \left(\frac{9}{tsw}\right)^{0.75} \times \left(\frac{63}{tsp}\right) \times a_t$$

The Graph side

I lead the climate physics group and do research and teaching on many topics related to climate change.

These include long term projections, scenarios, the 2°C target, uncertainties in projections, climate model evaluation, model weighting, natural climate variab-

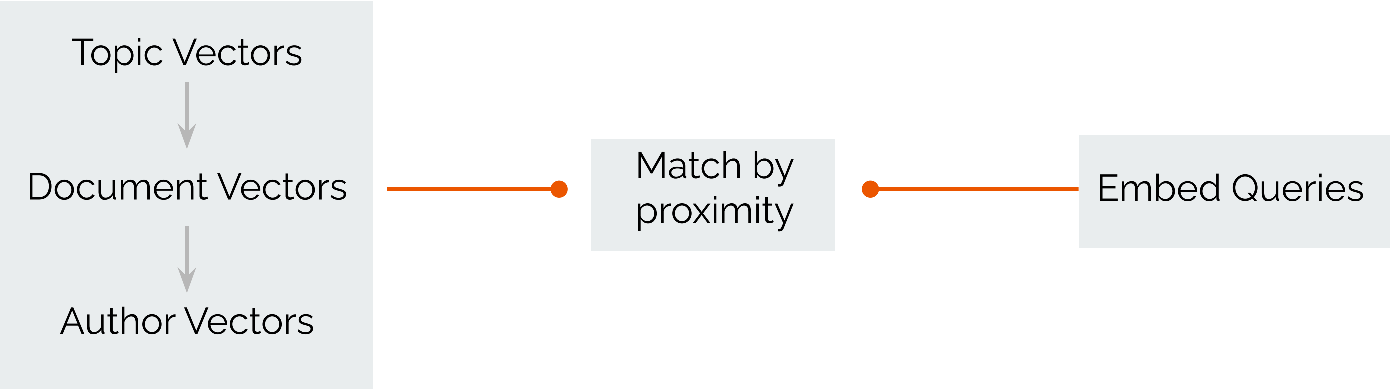
Pre-process input query for graph → Build Cypher queries that search the graph for relevant information → Neo4j results

```
preprocess('Climate Change')
['climat', 'chang']
```

```
WITH ['climat', 'chang'] as words
MATCH (w:Word)-[r1:IS_IN]-(t:Topic)
WHERE w.name in words
WITH t,  size(words) as inputCnt,
    count(DISTINCT w) as cnt, AVG(r1.weight) as s
WHERE cnt = inputCnt
WITH   t,s
MATCH (t:Topic)-[r3:EXPERT_ON] - (p:Person)-[r8]
                                  -(o:Organisation)
WITH p, SUM(r3.score_ipf*s) as s2, o
RETURN p.name, s2, o.name
ORDER BY s2 DESC
```

1. Lohmann, Ulrike
2. Knutti, Reto
3. Gruber, Nicolas
4. Peter, Thomas
5. Buchmann, Nina

The Embedding side



The Embedding side

Combination of weights given by topic modelling

1. Words: GloVe embeddings
2. Topics: Convex combination of word embeddings
3. Documents: Convex combination of topic embeddings
4. Authors: Convex combination of document embeddings
5. Queries: Average of word embeddings

Custom-made scoring function describing the relevance of a publication to a given author

Scoring function based on:

- publication type
- publishing date
- number of authors

The Embedding side

Pre-process and embed input query

↓

Search the embedding space for relevant information

↓

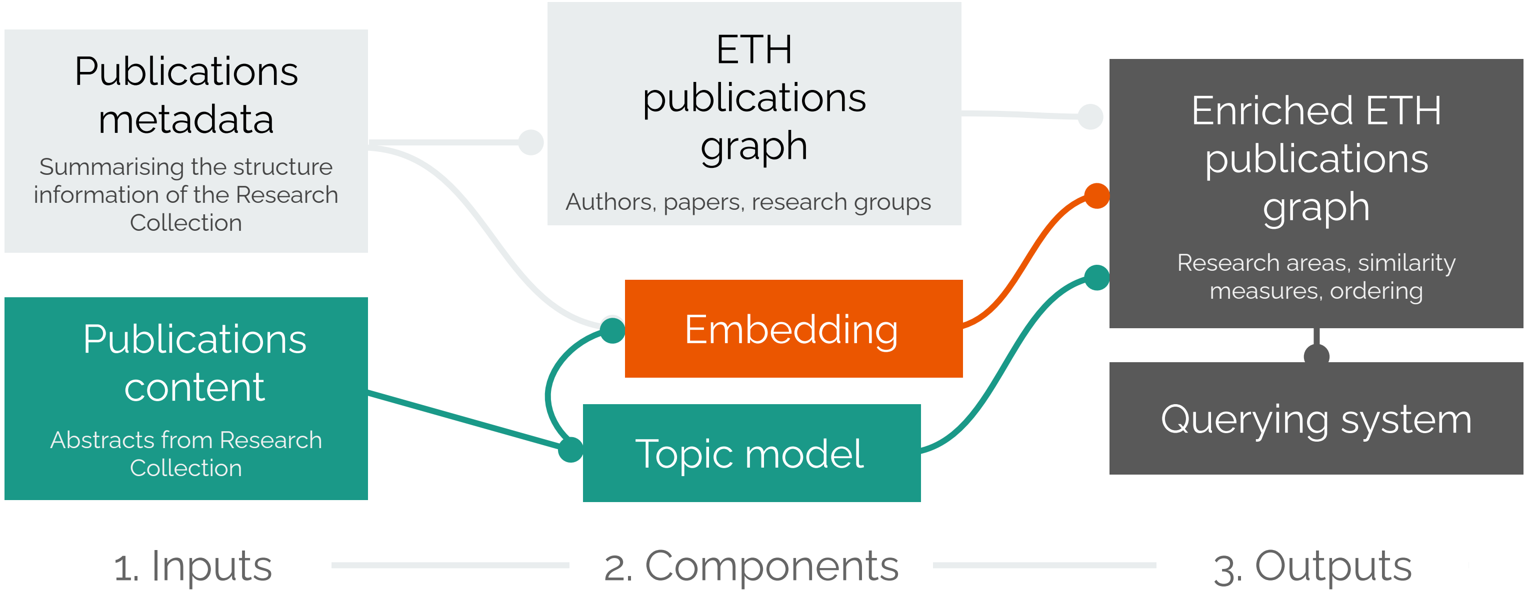
Embedding results

```
query = "climate change"
query_embs = get_list_embeddings_query(query, glove_vocab, glove_embedding)
# aggregating the query embeddings with a mean
query_emb = np.mean(query_embs, axis=0)
```

```
def visualise_most_similar_docs_query
```

```
Query: climate change
----------
10 most similar documents in the collection

Document 6540
Determining the time of emergence of climates altered from their natural state by anthropogenic influences can help inform th
e development of adaptation and mitigation strategies to climate change. Previous studies have examined the time of emergence
of climate averages. However, at the global scale, the emergence of changes in extreme events, which have the greatest societ
al impacts, has not been investigated before. Based on state-of-the-art climate models, we show that temperature extremes gen
```

Summary of technical approach



Future Work

- Handle bi- and trigrams
- Clustering of topics
- Refining score function
- Representation Learning instead of convex combination
- Improve Quantity and Quality of Data
- Extend use of Embedding Space (Querying and beyond LDA)
- NER (Named-Entity Recognition) for querying

Thank you

Questions?