# Building a Conference Recommender System based on SciGraph and WikiCFP

Semantics 2019

# Who we are

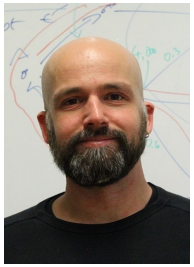
- Master students
- Business Informatics
- University of Mannheim

Andreea Iana

Steffen Jung

Philipp Näser

- Our supervisors



Heiko Paulheim

Sven Hertling

Aliaksandr Birukou
(Springer Nature)

# Imagine you write a paper ...

# Imagine you write a paper ...

Your research paper

Your recommendation

The goal of the project was to design a recommender system for scientific publications issued in conference proceedings by Springer Nature. The purpose of the system is to present recommendations of conference series to users who intend to submit their scientific work to a suitable conference.

Recommend Clear

**Recommendations**

| Rank | Conferenceseries | Confidence | Upcoming Date* |
|---|---|---|---|
| 1 | International Conference on Intelligent Text Processing and Computational Linguistics | 0.17 | |
| 2 | Asia Information Retrieval Symposium | 0.13 | |
| 3 | European Conference on Information Retrieval | 0.07 | 2019-04-14 |
| 4 | International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management | 0.04 | 2018-09-18 |

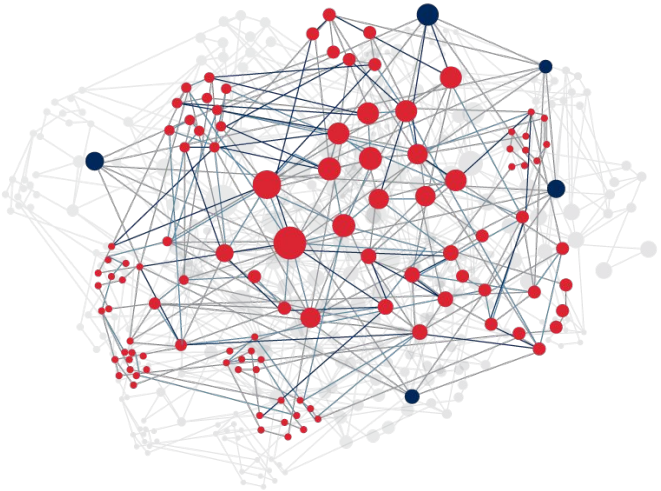
**Recommender System for Scientific Publication Opportunities**

Please insert your abstract here

Recommend Clear

# Datasets

1. Springer SciGraph
    - Data on past conferences
    - Used for model development

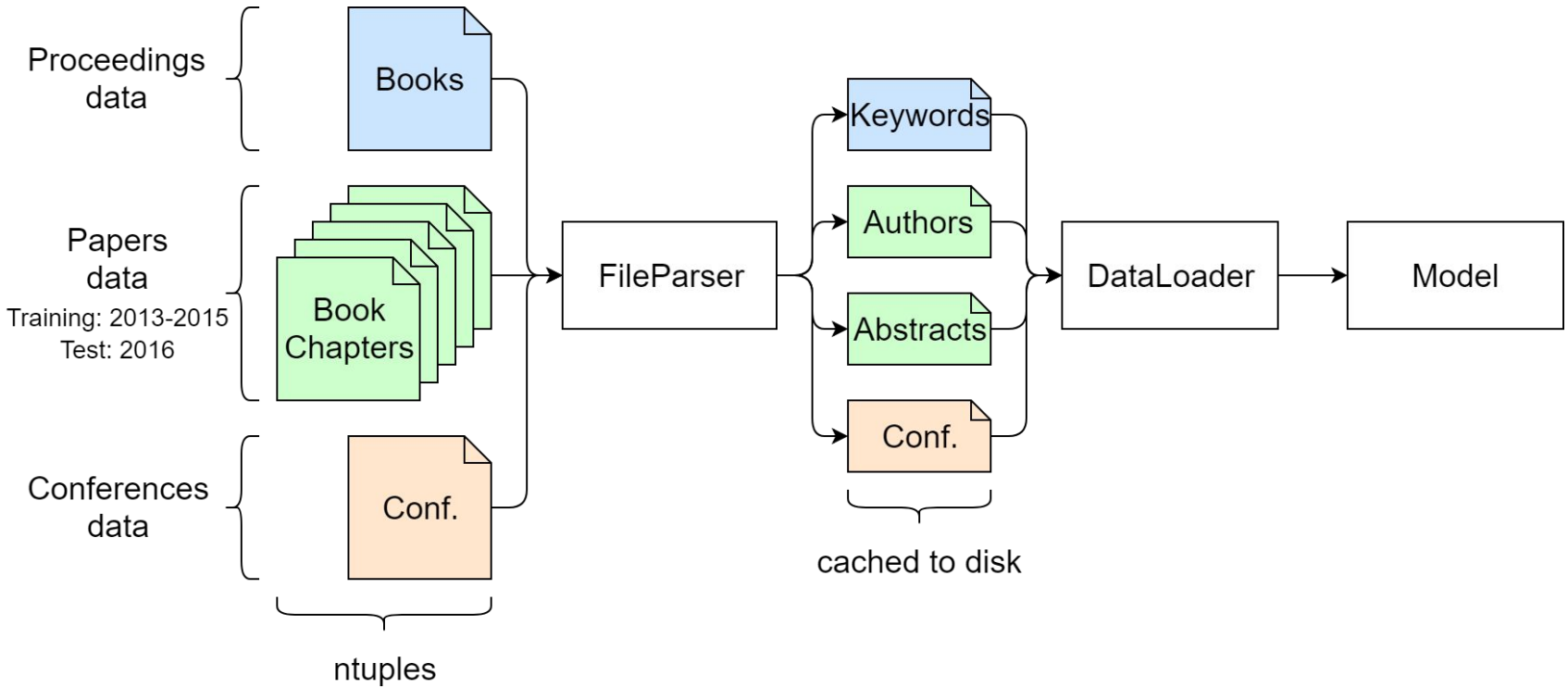


2. WikiCFP
    - Data on upcoming conferences
    - Used to enrich information in UI

WikiCFP

A Wiki for Calls For Papers

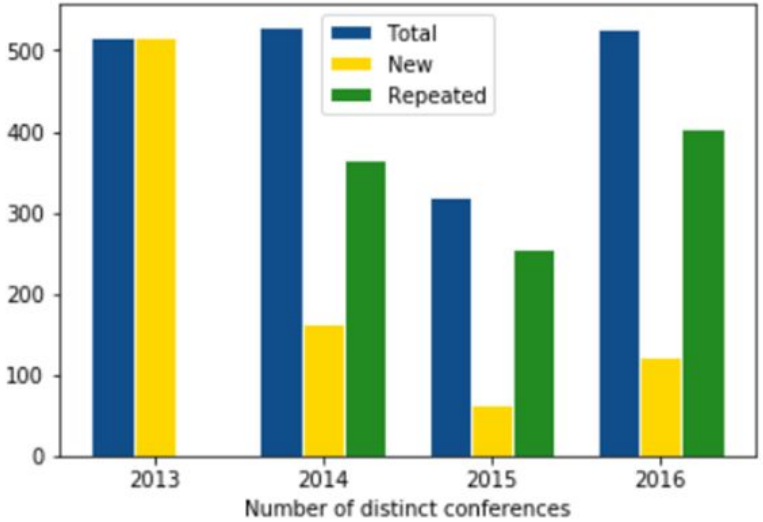
# SciGraph dataset



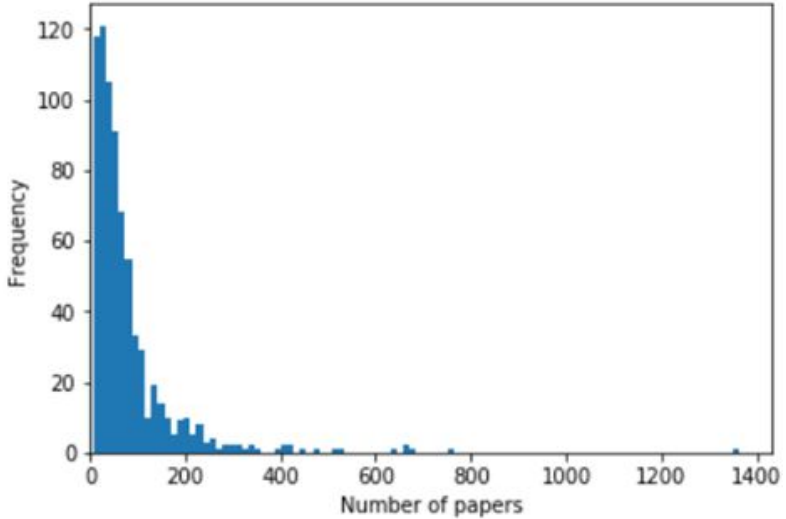
# SciGraph exploration: conference series

- Training data: 742 conference series
- Test data: 526 conference series (23% new in 2016)

Distinct conference series per year: [319 - 530]




Papers per conference: [6 - 1,364]



# WikiCFP exploration

- Total number of conferences crawled: 65,714
- Years: 2004-2017, and partially 2018-2021
- Linking to SciGraph data
  - Total number of conference series in SciGraph: 1,196
  - String similarity on conference names: 53.1% of SciGraph conference series linked

SEMANTiCS 2016 : 12th International Conference on Semantic Systems (Call for Workshop Proposals)

Link: http://www.semantics.cc/

| | |
|---|---|
| **When** | Sep 12, 2016 - Sep 12, 2016 |
| **Where** | Leipzig |
| **Submission Deadline** | Jun 23, 2016 |
| **Notification Due** | Jul 26, 2016 |

**Categories** semantic web ontology semantics linked data

# Evaluation metrics

**Basic setup:** 10 recommendations per model, 1 ground truth value

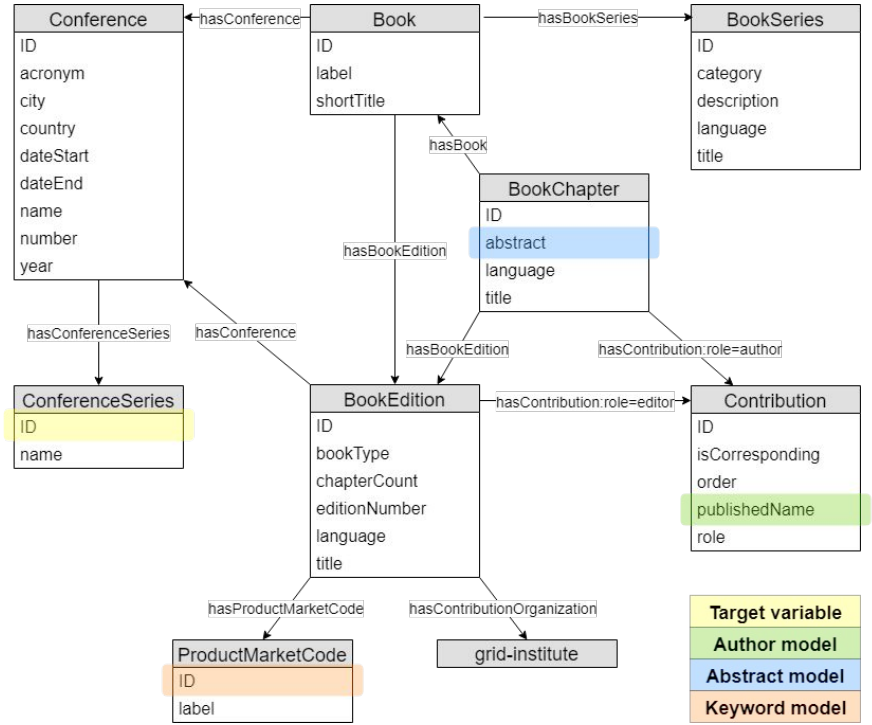
1. Recall: “How often did we get the truth value?”
2. Mean Average Precision (MAP): “How good is our ranking?”

**Best case scenario:**

1. Recall = 0.815 (truth value always contained)
2. MAP = 0.815 (truth value always on position 1)

# Recommendation models

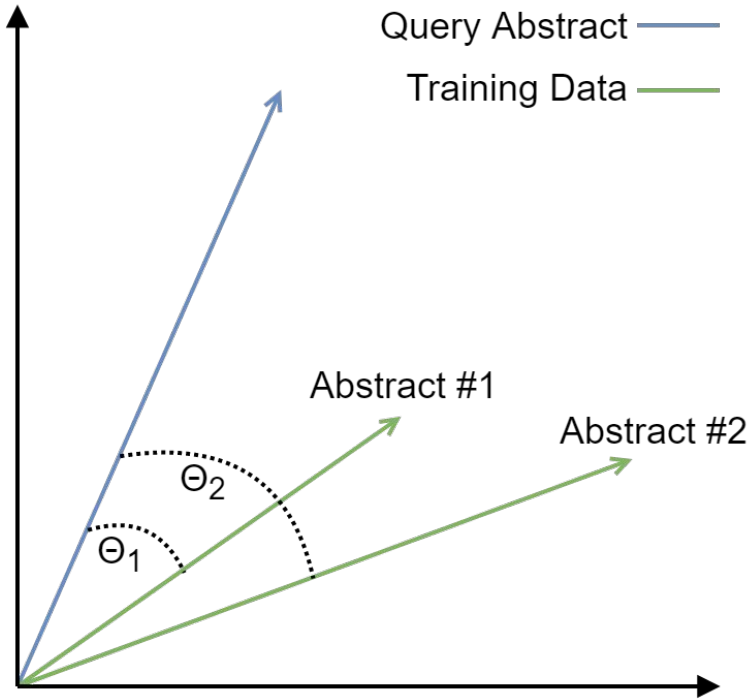
1. Models based on abstracts
2. Models based on keywords
3. Combined models



# Abstract models

**Basic idea**

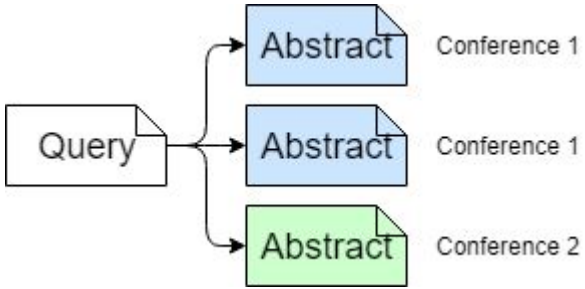
1. Convert an abstract into a feature vector
2. Compare with feature vectors in training data
3. Recommend closest conference series



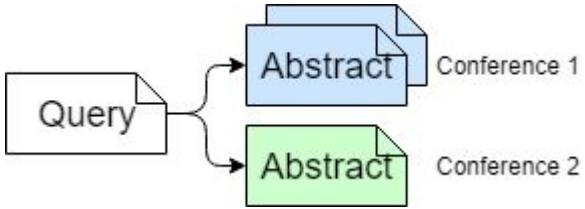
# Abstract models

**Two versions:**

1. Max




2. Concat




**Feature vector models**

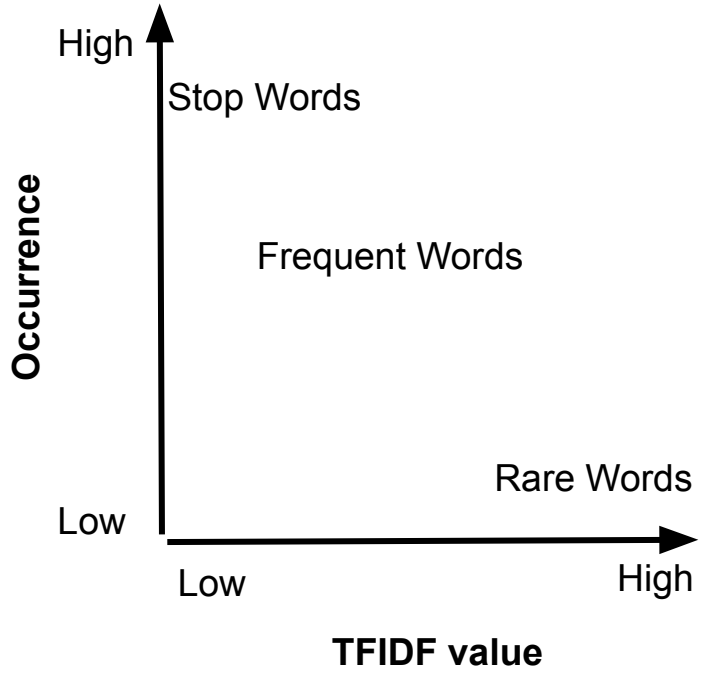
1. n-TFIDF
2. n-LSA
3. pLSA
4. Word embeddings

**Other models**

5. TFIDF + classifier
6. CNN

# n-TFIDF model

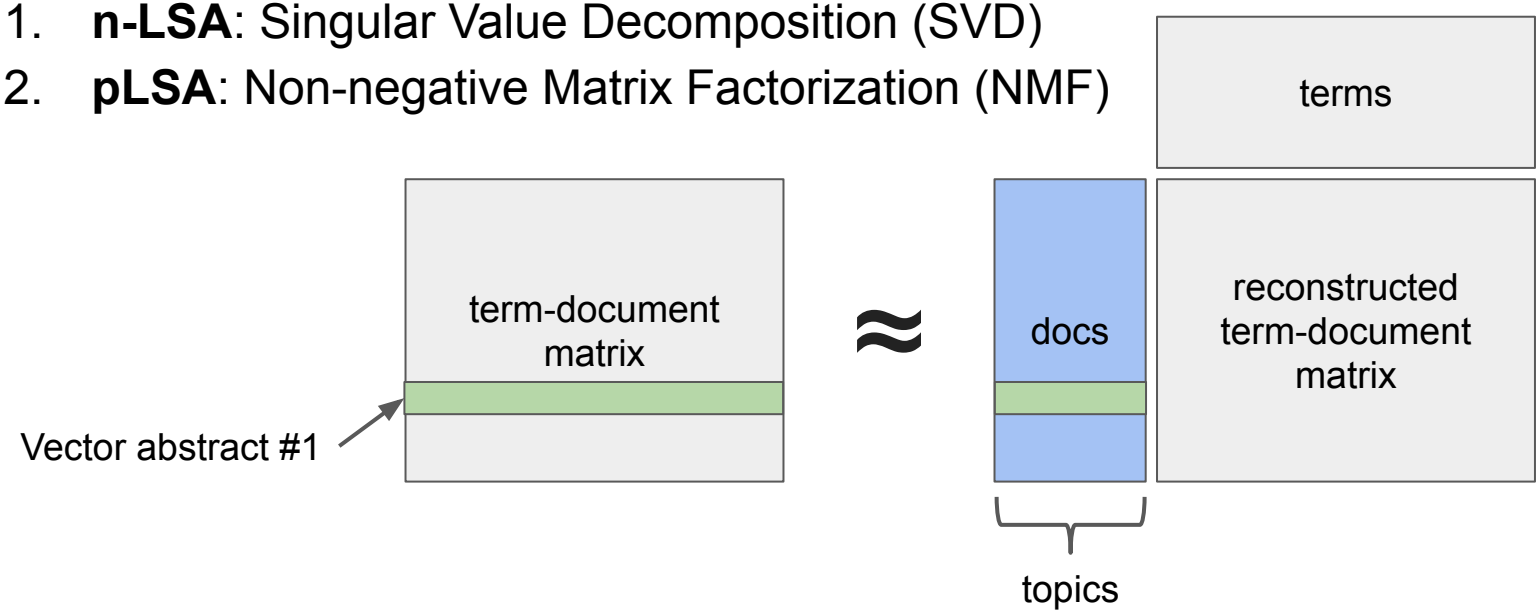
**Goal**: capture how relevant a word is in a given document from a corpus



# Dimensionality reduction

Different ways of reducing the dimensionality of feature vectors

1. **n-LSA**: Singular Value Decomposition (SVD)
2. **pLSA**: Non-negative Matrix Factorization (NMF)



# Word embeddings

**Goal**: capture the semantic similarity of words
**Basic idea**: predict the representation vectors of a word based on its context

**Model idea**

1. Uses vectors with word embeddings:
    a. Pretrained (i.e. Word2Vec, Glove, FastText) - low performance
    b. Trained on the abstracts of papers published in 2009-2015
2. Feature vector for the abstract computed by:
    a. Averaging the embedding vectors of all the words in the abstract
    b. Weighted average of the embedding vectors of all the words in the abstract (TFIDF weights)
    c. Using Doc2Vec: embed both the words and the entire document

# Abstracts model evaluation

**n-TFIDF**

| Model | Feat. | n_grams | max_feat | Recall | MAP |
|---|---|---|---|---|---|
| max | 1.6M | (1,2) | None | 0.357 | 0.175 |
| **concat** | **1M** | **(1,4)** | **1M** | **0.490** | **0.270** |

**n-LSA**

| Model | Topics | n_grams | max_feat | Recall | MAP |
|---|---|---|---|---|---|
| max | 1000 | (1,4) | 1M | 0.346 | 0.166 |
| **concat** | **1000** | **(1,4)** | **1M** | **0.490** | **0.270** |

← Converges to n-TFIDF

# Abstracts model evaluation

**pLSA**

| Model | Topics | Recall | MAP |
|---|---|---|---|
| max | 100 | 0.286 | 0.118 |
| **concat** | **500** | **0.369** | **0.172** |

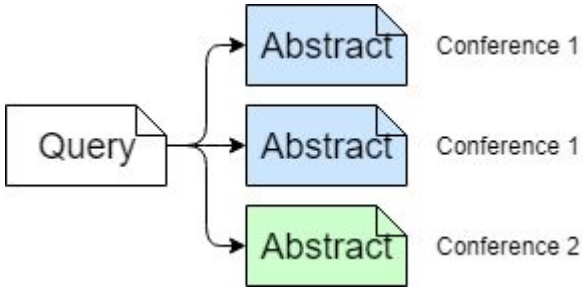
**Doc2Vec**

| Model | Vector Dimension | Recall | MAP |
|---|---|---|---|
| max | 100 | 0.312 | 0.131 |
| **concat** | **100** | **0.352** | **0.164** |

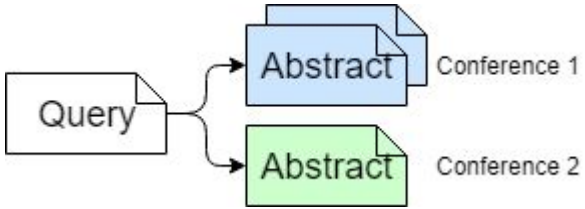
# Abstract models

**Two versions:**

1. Max




2. Concat



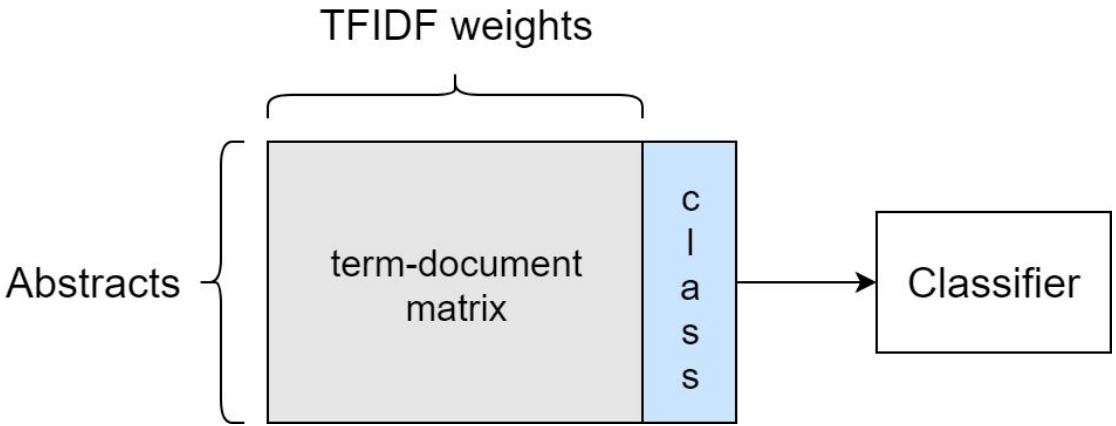

**Feature vector models**

1. n-TFIDF
2. n-LSA
3. pLSA
4. Word embeddings

**Other models**

5. TFIDF + classifier
6. CNN

# TFIDF + Classifier

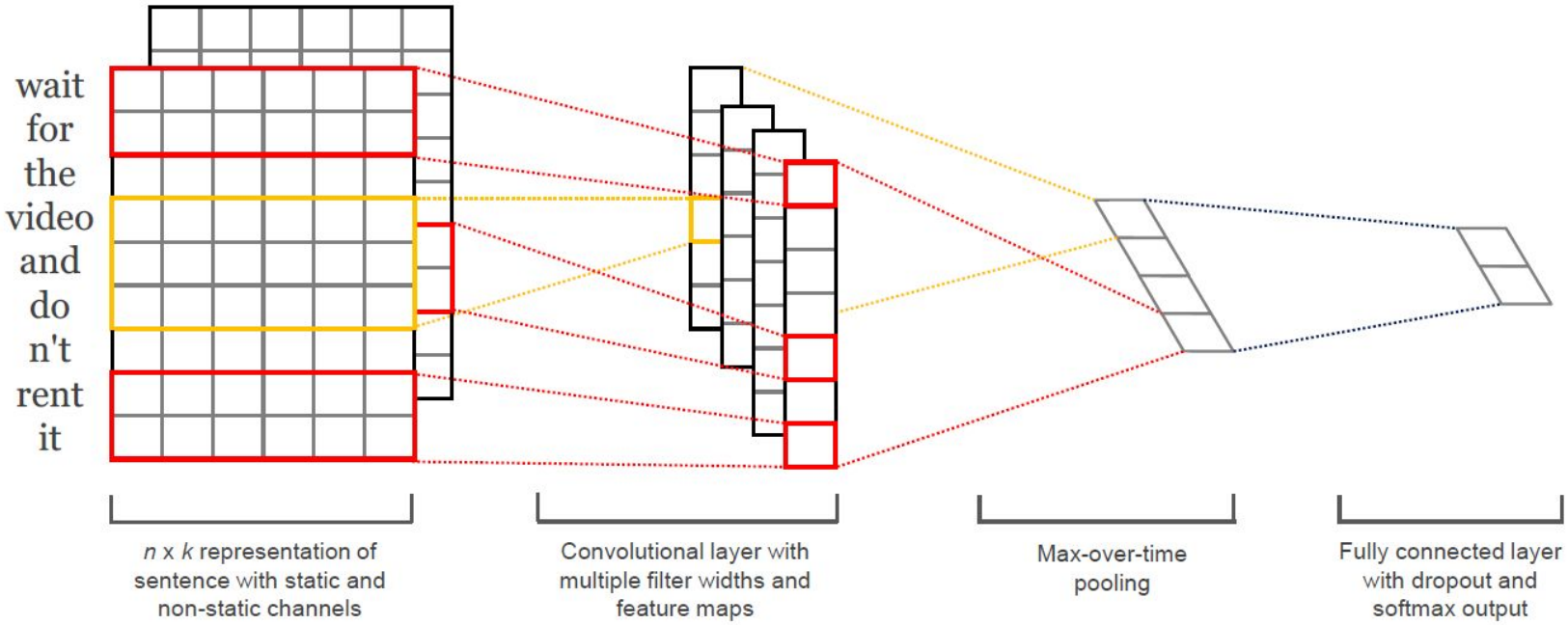



- Random Forest
- Multinomial Naive Bayes
- K-NN
- Logistic Regression
- AdaBoost

# CNN model

**Inspired by Y. Kim, 2014**

“Convolutional Neural Networks for Sentence Classification”



# Abstract models evaluation

**TFIDF + Classifier**

| Classifier | Notes | n_grams | max_feat | Recall | MAP |
|---|---|---|---|---|---|
| KNN (20) | max | (1,1) | None | 0.326 | 0.188 |
| **MultinomialNB** | **concat** | **(1,4)** | **750K** | **0.494** | **0.273** |

**CNN**

| Model | Filters | FC Layer | Recall | MAP |
|---|---|---|---|---|
| 100d-w2v-w10-SG | 100 * [3,4,5] | 300 → classes | 0.402 | 0.197 |
| **100d-w2v-w10-SG** | **100 * [3,4,5]** | **300 → 1024 → classes** | **0.405** | **0.201** |

# Keyword model: SciGraph Product Market Codes

**Model idea**

1. Extract keywords from proceedings
2. Generate TFIDF keyword vectors
3. At test time, compare vectors using cosine similarity

# Keyword model evaluation

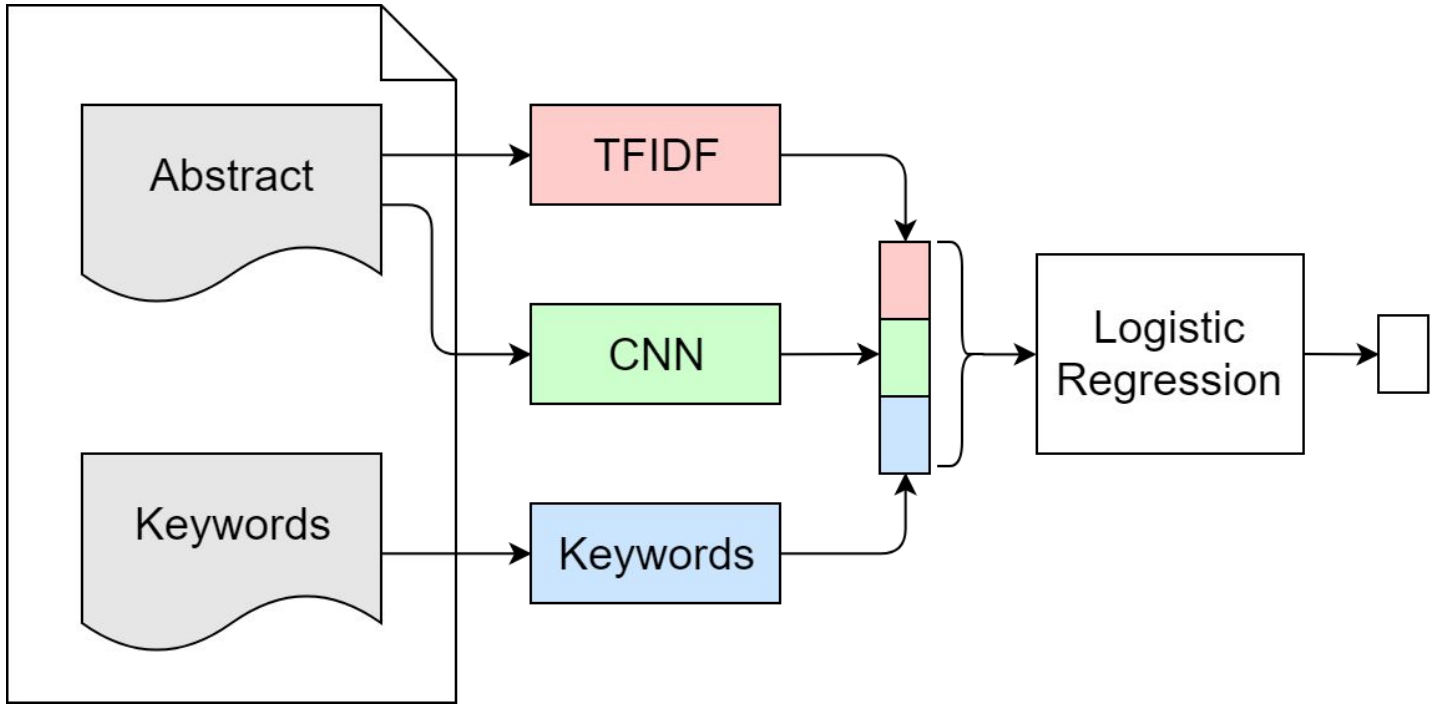
**SciGraph keyword model**

| Model | Recall | MAP |
|---|---|---|
| **max** | **0.665** | **0.522** |
| concat | 0.646 | 0.475 |

**But:** Product market codes are selected by Springer (not the authors)

# Ensemble model

- Combine the predictions of individual classifiers



# Ensemble model evaluation

| Type | Models | Recs | Recall | MAP |
|---|---|---|---|---|
| Addition | TFIDF + CNN | | 0.498 | 0.250 |
| Stacking | TFIDF + CNN + Keywords | 10 | 0.648 | 0.509 |
| **Stacking** | **TFIDF + CNN + Keywords** | **100** | **0.662** | **0.539** |
| Stacking | TFIDF + CNN + Keywords | 1000 | 0.661 | 0.540 |

| Model | Recall | MAP | Notes |
|---|---|---|---|
| TFIDF | 0.490 | 0.270 | concat |
| CNN | 0.391 | 0.195 | - |
| SciGraph Keywords | **0.665** | 0.522 | max |

# Limitations

**Models**

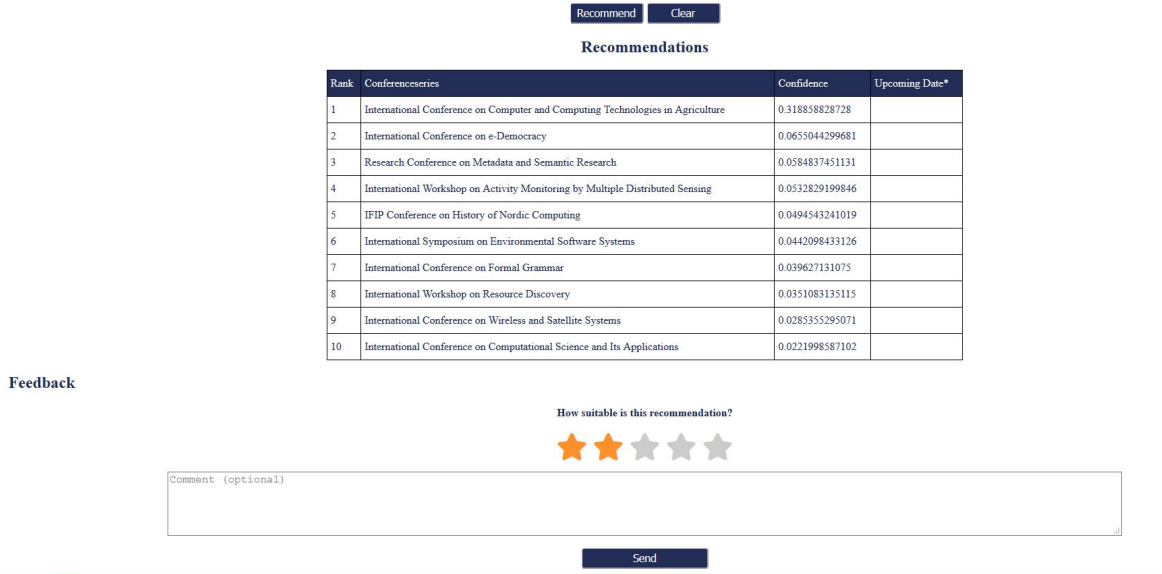
- Almost all proceedings in SciGraph are related to Computer Science: 99%
- There is a bias towards AI conferences, due to their strong representation
- Based solely on proceedings published by Springer
- Abstract models could be further improved

**External information (WikiCFP)**

- Not all conference series from SciGraph are contained in WikiCFP
- Missing information for some of the crawled WikiCFP conferences
- No continuous crawling of WikiCFP

# Outlook

- Models using the graph structure (e.g. graph embeddings)
- Consider additional features (e.g. citations: available now in SciGraph)
- User study to evaluate the quality of the recommendations



# Thank you for your attention!

The goal of the project was to design a recommender system for scientific publications issued in conference proceedings by Springer Nature. The purpose of the system is to present recommendations of conference series to users who intend to submit their scientific work to a suitable conference.

Recommend Clear

**Recommendations**

| Rank | Conferenceseries | Confidence | Upcoming Date* |
|---|---|---|---|
| 1 | International Conference on Intelligent Text Processing and Computational Linguistics | 0.17 | |
| 2 | Asia Information Retrieval Symposium | 0.13 | |
| 3 | European Conference on Information Retrieval | 0.07 | 2019-04-14 |
| 4 | International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management | 0.04 | 2018-09-18 |

http://confrec.dws.uni-mannheim.de/

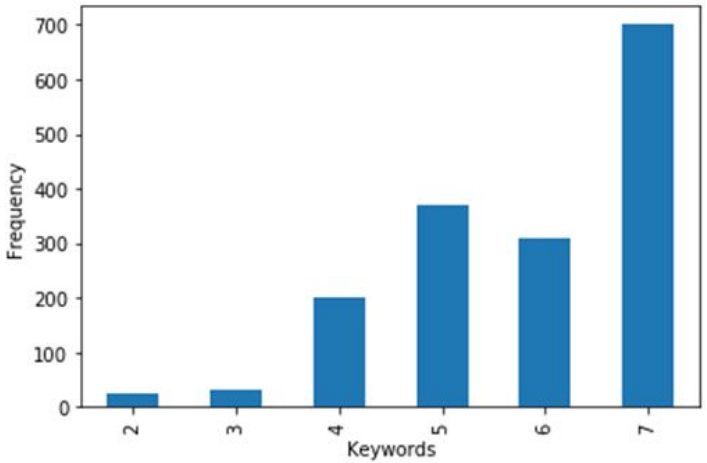
# References

1. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

# Back-up Slides

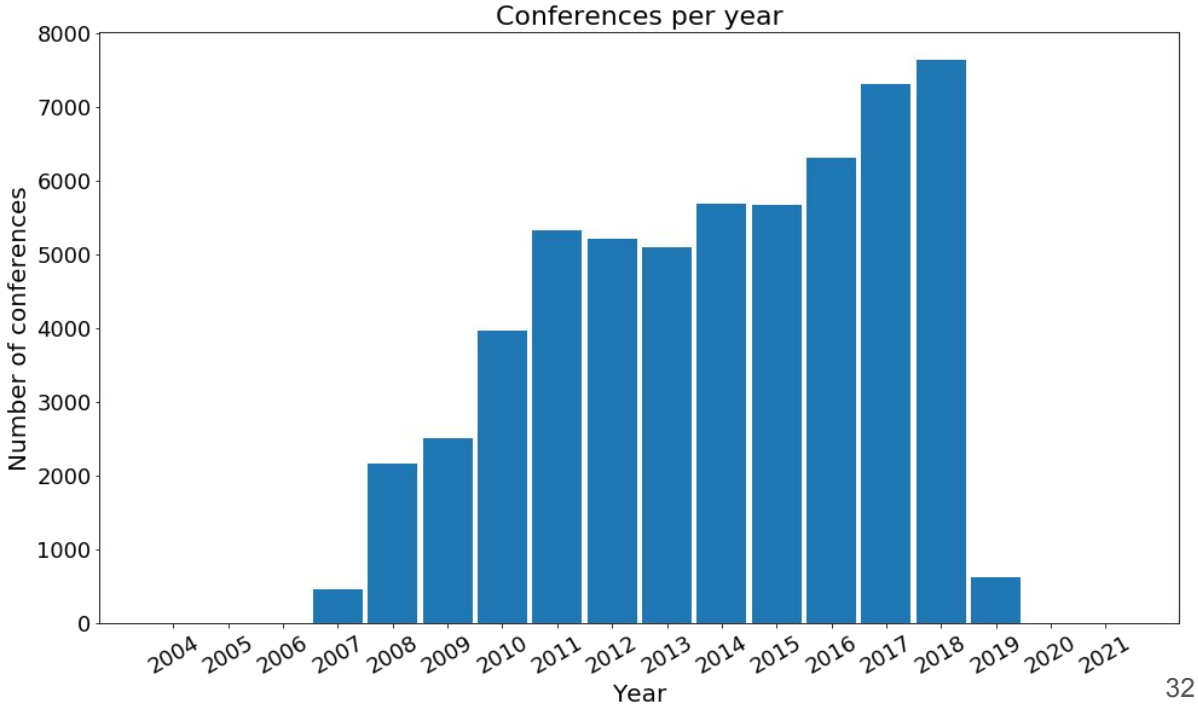
# SciGraph exploration: keywords

- Training data: 155 distinct tags
- Test data: 150 distinct tags, of which ~77% are in the training data
- Number of keywords per book: [2 - 7]



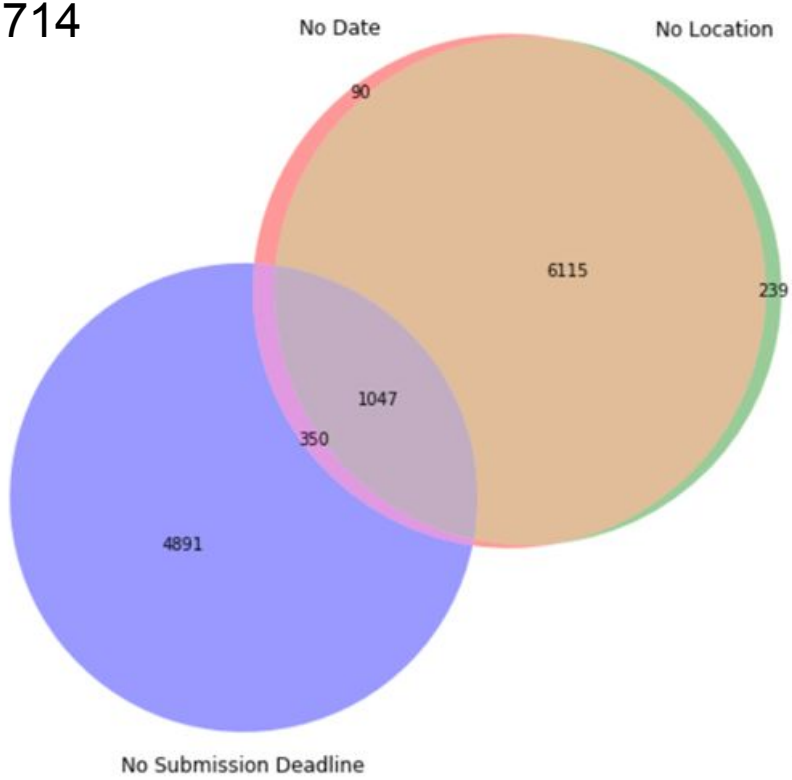
# WikiCFP exploration

- Number of conferences per year
  - Mean: 3,227
  - Median: 3,249



# WikiCFP exploration

- Total number of conferences crawled: 65,714
- Conferences with incomplete information
  - Missing start/end date: 7,602
  - Missing location: 7,430
  - Missing submission deadline: 6,317



# WikiCFP linking

**Development decisions**

- No continuous crawling during development
- Crawled on: 11/July/2018 (i.e. threshold date)
- UI only considers conferences taking place after threshold date: 223

**Recommendations**

| Rank | Conferenceseries | Confidence | Upcoming Date* |
|---|---|---|---|
| 1 | International Workshop on Artificial Neural Networks | 0.13 | 2019-06-12 |
| 2 | International Conference on Artificial Intelligence and Soft Computing | 0.13 | |
| 3 | International Conference on Engineering Applications of Neural Networks | 0.11 | 2018-09-03 |

# Evaluation metrics: Mean Average Precision

**Recommendations**

| |
|---|
| ECCV |
| ICCV |
| CVPR |
| GCPR |
| ... |

**Ground truth:** ECCV

**Precision@K:** 1/1 = 1.0

**Recommendations**

| |
|---|
| ICCV |
| ECCV |
| CVPR |
| GCPR |
| ... |

**Ground truth:** ECCV

**Precision@K:** 1/2 = 0.5

**Recommendations**

| |
|---|
| ICCV |
| CVPR |
| ECCV |
| GCPR |
| ... |

**Ground truth:** ECCV

**Precision@K:** 1/3 = 0.33

**Recommendations**

| |
|---|
| ICCV |
| ACCV |
| CVPR |
| GCPR |
| ... |

**Ground truth:** ECCV

**Precision@K:** 0.0

**MAP** = (1.0 + 0.5 + 0.33 + 0.0) / 4 = 0.46

# Author names model (baseline)

**Model idea**

1. Compare given author names with author names in the training data

2. Recommend conference series of papers with the highest number of matching author names

# Author names model evaluation

- Training data: 110,831 distinct author names
- Test data: 53,862 (62% new in 2016)

| Model | Recall | MAP | Notes |
|---|---|---|---|
| All rec. | 0.377 | 0.284 | 23.6% None |
| Top 10 rec. | 0.372 | 0.284 | 23.6% None |

- No recommendations in many cases
- Author names ambiguous

# TFIDF model evaluation

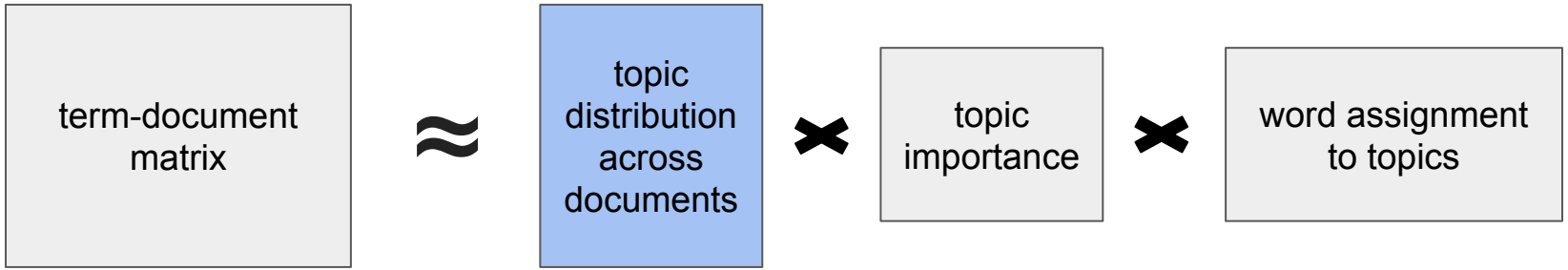
| Model | min_df | Recall | MAP | Feat. |
|---|---|---|---|---|
| max | 0 | 0.334 | 0.163 | 54,594 |
| max | 3 | 0.334 | 0.162 | 17,961 |
| max | 9 | 0.334 | 0.160 | 9,256 |
| **concat** | **0** | **0.461** | **0.237** | **54,594** |
| concat | 3 | 0.456 | 0.232 | 17,017 |
| concat | 9 | 0.450 | 0.227 | 8,478 |

# n-TFIDF model evaluation

| Model | n_grams | max_feat | Recall | MAP | Feat. |
|---|---|---|---|---|---|
| max | (1,2) | None | 0.357 | 0.175 | 1,652,643 |
| concat | (1,2) | None | 0.487 | 0.268 | 1,671,678 |
| concat | (1,3) | 500K | 0.490 | 0.267 | 500,000 |
| concat | (1,3) | 1M | 0.489 | 0.269 | 1,000,000 |
| concat | (1,4) | 500K | 0.490 | 0.267 | 500,000 |
| **concat** | **(1,4)** | **1M** | **0.490** | **0.270** | **1,000,000** |
| concat | (1,5) | None | 0.460 | 0.254 | 14,944,591 |

# n-LSA and pLSA models

**n-LSA:** uses SVD




**pLSA:** uses NMF

# LSA model evaluation

| Model | Topics | Recall | MAP |
|---|---|---|---|
| max | 50 | 0.305 | 0.126 |
| max | 200 | 0.317 | 0.143 |
| max | 500 | 0.330 | 0.155 |
| max | 1000 | 0.330 | 0.158 |
| concat | 50 | 0.372 | 0.173 |
| concat | 200 | 0.429 | 0.213 |
| concat | 500 | 0.449 | 0.225 |
| **concat** | **1000** | **0.461** | **0.237** |

Converges to TFIDF

# n-LSA model evaluation

| Model | Topics | n_grams | max_feat | Recall | MAP |
|---|---|---|---|---|---|
| max | 1000 | (1,4) | 1M | 0.346 | 0.166 |
| concat | 1000 | (1,3) | 300K | 0.470 | 0.248 |
| concat | 500 | (1,3) | 300K | 0.489 | 0.266 |
| concat | 1000 | (1,3) | 1M | 0.489 | 0.269 |
| concat | 1000 | (1,4) | 500K | 0.490 | 0.267 |
| concat | 500 | (1,4) | 1M | 0.466 | 0.248 |
| **concat** | **1000** | **(1,4)** | **1M** | **0.490** | **0.270** |

Converges to n-TFIDF

# pLSA model evaluation

| Model | Topics | Reg. | Init. | Recall | MAP |
|---|---|---|---|---|---|
| max | 100 | 6 | rand | 0.285 | 0.116 |
| max | 100 | 6 | nndsvd | 0.286 | 0.118 |
| concat | 100 | 0 | rand | 0.213 | 0.093 |
| concat | 100 | 0.5 | rand | 0.356 | 0.162 |
| concat | 100 | 0 | nndsvd | 0.300 | 0.132 |
| concat | 100 | 1 | nndsvd | 0.363 | 0.166 |
| **concat** | **500** | **1** | **nndsvd** | **0.369** | **0.172** |

# Word embeddings: simple averaging

| Embedding | Model | Vector Dimension | Window Size | Recall | MAP |
|---|---|---|---|---|---|
| Word2Vec | max | 50 | 10 | 0.329 | 0.142 |
| | **max** | **100** | **10** | **0.346** | **0.154** |
| | concat | 50 | 10 | 0.314 | 0.139 |
| | concat | 100 | 10 | 0.333 | 0.152 |
| FastText | max | 50 | 5 | 0.308 | 0.130 |
| | max | 100 | 5 | 0.331 | 0.144 |
| | concat | 50 | 5 | 0.285 | 0.122 |
| | concat | 100 | 5 | 0.312 | 0.138 |

# Word embeddings: TFIDF-weighted-averaging

- Word2Vec embeddings
  - Vector size: 100
  - Window size: 10

| Model | Recall | MAP |
|---|---|---|
| max | 0.346 | 0.154 |
| concat | 0.333 | 0.152 |

# Doc2Vec: model evaluation

| Model | Vector Dimension | Window Size | Recall | MAP |
|---|---|---|---|---|
| max | 100 | 5 | 0.312 | 0.131 |
| max | 100 | 10 | 0.273 | 0.112 |
| max | 300 | 5 | 0.281 | 0.122 |
| max | 400 | 5 | 0.279 | 0.121 |
| **concat** | **100** | **5** | **0.352** | **0.164** |
| concat | 100 | 10 | 0.300 | 0.129 |
| concat | 300 | 5 | 0.341 | 0.159 |
| concat | 400 | 5 | 0.333 | 0.154 |

# TFIDF + Classifier

| Model | n_grams | max_feat | Recall | MAP | Notes |
|---|---|---|---|---|---|
| MultinomialNB | (1,1) | None | 0.467 | 0.244 | concat |
| MultinomialNB | (1,1) | None | 0.177 | 0.066 | max |
| knn(5) | (1,1) | None | 0.236 | 0.146 | max |
| knn(20) | (1,1) | None | 0.326 | 0.188 | max |
| AdaB(100) | (1,1) | None | 0.108 | 0.038 | max |
| MultinomialNB | (1,4) | 500K | 0.493 | 0.273 | concat |
| **MultinomialNB** | **(1,4)** | **750K** | **0.494** | **0.273** | **concat** |

# CSO keyword model

## Model idea

1. Extract keywords from abstracts given by CSO.
2. Generate keyword vectors, including parent elements.
3. At test time, compare vectors using cosine similarity.

| Model | Recall | MAP | Notes |
|---|---|---|---|
| **CSO** | **0.201** | **0.081** | **max** |
| CSO | 0.027 | 0.008 | concat |
| CSO-IDF | 0.199 | 0.081 | max |
| CSO-IDF | 0.093 | 0.029 | concat |

# Recommendation models evaluation

| Model | Version | Recall | MAP |
|---|---|---|---|
| n-TFIDF / n-LSA | concat | 0.490 | 0.270 |
| pLSA | concat | 0.369 | 0.172 |
| n-TFIDF + NB | concat | 0.494 | 0.273 |
| Doc2Vec | concat | 0.352 | 0.164 |
| CNN | - | 0.391 | 0.195 |
| SciGraph keywords | max | 0.665 | 0.522 |
| **Stacking ensemble (TFIDF + CNN + Keywords)** | **-** | **0.662** | **0.539** |