IFLA WLIC 2019 Athens

**Submitted on:** 26.09.2019

*Title of the Satellite Meeting: Data intelligence in libraries: the actual and artificial perspectives*
*Date: 22 – 23 August 2019*
*Location: Deutsche Nationalbibliothek, Frankfurt, Germany*

# An innovative approach to scalable semantic embedding

**Rob Koopman**
Global Engineering Department, OCLC, Leiden, The Netherlands.
E-mail address: rob.koopman@oclc.org

**Shenghui Wang**
Research Department, OCLC, Leiden, The Netherlands
E-mail address: shenghui.wang@oclc.org



## Abstract:

*Embedding words, entities and documents in compact, semantically meaningful vector spaces allows for computable semantic similarity/relatedness which could make search more intelligent and benefit other tasks conducted in libraries, such as entity disambiguation, de-duplication, clustering, recommendation, subject prediction, etc. Deep learning models are powerful but require high computing power and careful tuning hyperparameters for optimal performance. In our quest for practical solutions to support libraries in this field, we revisit the global co-occurrence based embedding methods and propose a conceptually simple and computationally lightweight approach. Our experiments show highly competitive results with a few state-of-the-art embedding methods on different tasks, including the standard STS benchmark and a subject prediction task, at a fraction of the computational cost. We will show the potentials of this scalable semantic embedding method for other applications such as entity disambiguation, citation recommendation, clustering and collection exploration.*

**Keywords:** Semantic Embedding, Random Projection, Subject Prediction.

## Introduction

Being able to measure similarity or relatedness is important to modern digital library systems. Tasks including information retrieval, entity disambiguation, de-duplication, clustering, recommendation, subject prediction, etc. all make use of similarity/relatedness one way or another. Keyword-based or string similarity-based methods cannot fulfil these tasks in a satisfactory manner. Big search engines currently benefit from semantic embedding technologies for better information retrieval. These embedding technologies enable us to

represent words, entities, bibliographic records in compact, semantically meaningful vector spaces. This allows for computable semantic similarity/relatedness which would benefit the various tasks mentioned above.

Much semantic embedding research has adopted the notion of *Statistical Semantics* (Furnas et al., 1983; Weaver, 1955) based on the assumption of "a word is characterized by the company it keeps" or in Linguistics the *Distributional Hypothesis* (Harris, 1954; Sahlgren, 2008): words that occur in similar contexts tend to have similar meanings. Various distributional semantic models have been proposed to represent (embed) words in a continuous vector space where semantically similar words are mapped to nearby points ("are embedded nearby each other").

The recent success of local context predictive models such as Word2Vec (Mikolov et al., 2013) have initiated the development of more complex and powerful deep learning models (Bojanowski et al., 2016; Peters et al., 2018). Deep learning models produce compact and discriminating embeddings, however, have substantial computation requirements if applied on large bibliographic collections. They also need careful tuning for optimal hyperparameter settings. Unfortunately, most libraries and even large-scale aggregators do not have the processing capacity nor the skills to embrace powerful deep learning. They either stick with the traditional keyword-based approach, or use embeddings pre-trained on large corpora and plug into various downstream tasks (clustering, classification, etc.). However, for the latter option, such transfer learning might fail to capture crucial domain-specific semantics for applications such as medical information retrieval, special collection exploration, etc.

In our quest for practical solutions to support libraries in this field, we revisit the global co-occurrence based embedding methods and propose a conceptually simple and computationally lightweight approach. Our method extends random projection by weighting and projecting raw term embeddings orthogonally to an average language vector. As a result, the discriminating power of the term embeddings is increased, and meaningful document embeddings can be built by assigning appropriate weights to individual terms. In the paper, we describe how updating the term embeddings online, as we process the training data, results in an extremely efficient method (orders of magnitude faster than Word2Vec/fastText). Our experiments show highly competitive results with a few state-of-the-art embedding methods on different tasks, including the standard Semantic Textual Similarity (STS) benchmark and an extreme multi-label classification (automatic MeSH subject prediction) task, at a fraction of the computational cost.

## Ariadne semantic embedding method

### (1) Fast and efficient random projection

Let a document be a set of words and entities for which co-occurrence is relevant. In general, a document could therefore be a sentence, a paragraph, a fixed-size window, or, in our case, a bibliographic record. Let $n_E$ be the total number of *frequent* entities – which could be terms (words or phrases), subjects, authors, citations – we want to embed, and $D$ the chosen dimensionality of the embedding vectors. An entity is considered frequent when it occurs in more than $K$ documents in the corpus, where $K$ is flexible depending on the size of the corpus.

Building on the previous work (Koopman et al. 2015, 2017, 2019) we embed the relevant entities by Random Projection (Achlioptas 2003, Johnson and Lindenstrauss 1984) of their weighted co-occurrences, as shown in Figure 1.

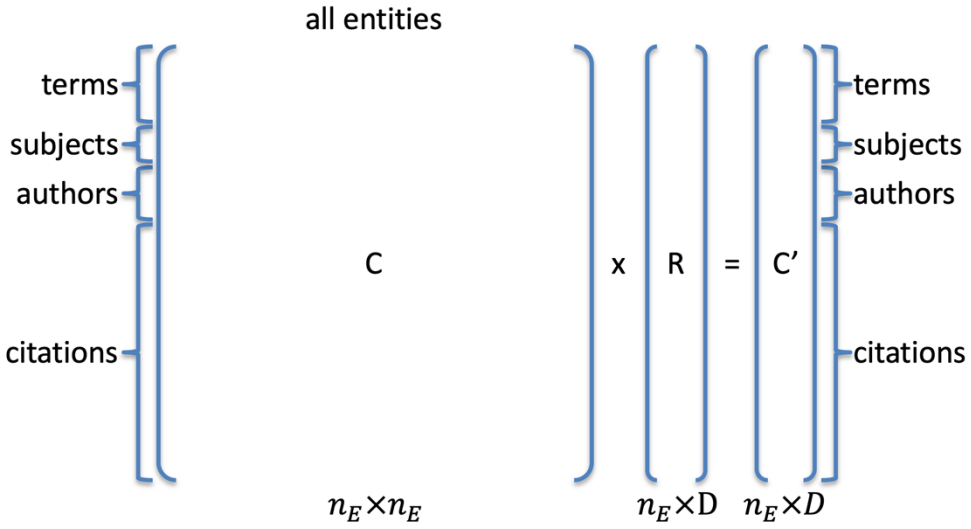


**Figure 1. Random projection**

Here, $C$ is the co-occurrence matrix of different types of entities, $R$ is a random matrix of +1 and -1. $C'$ is the matrix of final embedding vectors, each row representing an embedded entity.

Traditional random projection starts by computing the co-occurrence matrix $C$ of size $n_E \times n_E$. This matrix contains, for each pair of entities the number of documents (or paragraphs, or sentences) of the corpus in which both entities occur. Using a matrix of random projection vectors $R$ of size $n_E \times D$ ,we can then project our $n_E$ dimensional representation of each entity to a lower $D$ dimensional space. By leveraging the linear nature of the matrix multiplication, we can update $C'$ directly as we go through the corpus, without ever explicitly representing $C$. Koopman et al. (2019) has shown that this method is simple but highly efficient and scalable compared to other complicated methods while the competitive results are achieved at a fraction of the computational cost.

**(2) Orthogonal projection and weight assignment**

One crucial step after the standard random projection as described above is to remove the noise collected from the corpus to increase the discriminating power of the final embeddings. Traditionally very frequent words (so-called "stop words") are removed explicitly beforehand. In our approach, we give a continuous weight to all entities based on how frequently they occur and compute the "average vector" of the corpus, $\overrightarrow{v_a}$ , the sum of all the rows of $C'$. Entities are increasingly more informative as they differ more from this average vector. By this reasoning, we project entity vectors $\overrightarrow{v_e}$ on the orthogonal hyperplane to $\overrightarrow{v_a}$:

$$\overrightarrow{v_e^*} = \overrightarrow{v_e} - (\overrightarrow{v_e} \cdot \overrightarrow{v_a})\, \overrightarrow{v_a} \ ,$$

resulting in a representation where the uninformative component of entities is eliminated and normalise the vectors to have unit length.

Before computing document vectors, we calculate the weights for each entity according to its original similarity to $\overrightarrow{v_a}$, as

$$w_e = 1 - \cos(\overrightarrow{v_e}, \overrightarrow{v_a}).$$

This way, the more informative entities, i.e., those sharing less with the average vector, gets higher weights when calculating the document embeddings. Koopman et al. (2019) has illustrated the importance of this step in terms of getting distinctive document embeddings.

**(3) Document embedding**
With the embeddings of the frequent simple entities and their proper weights, we can compute embedding of a bibliographic record as the weighted average of its component entities' embeddings. The component entities could be its title words, authors, publishers, citations, and any entity that has been embedded.

**Usage of semantic embedding**
Till now, entities and documents are embedded in the same semantic space, all having the unit length, making similarity computations elegant and effective. We can now calculate the similarity/relatedness between any pairs of or different types of entities or between an entity and a bibliographic record. We have been applying our semantic embedding method in various applications. We summarise our work so far and present some potential usages for the future:

- Given a query, retrieve the most related bibliographic records, or authors, publishers, citations, etc. (Wang and Koopman 2017b);
- Based on the similarity scores, identify the duplicates in the collection, or cluster the records into meaningful clusters (Wang and Koopman 2017a);
- Given an entity (author or subject), visually explore its closest context (Koopman et al. 2017, Castermans et al. 2018);
- For a new bibliographic record, automatically suggest subjects or classifications for indexing (Koopman et al. 2019, Wang et al. 2019);
- Find relevant paragraphs in full text collections (Wierst et al. 2018);
- Given a text, find the most related authors who may have written about the same topic or journals that have publish on the same topic;
- Generate personal recommendations based on user's interests;
- Help to select journals that could be suitable venues to submit an article;
- Match subjects from different thesauri, explicitly or implicitly;
- Classifications often cover only a subset of the collection. Semantic embedding help retrieving relevant documents even if they are not assigned with any classification code.

This list is of course not complete. Whenever the similarity/relatedness is needed, semantic embedding could always help.

**Experiments**

**(1) STS benchmark and computational efficiency**
We use the Semantic Textual Similarity (STS) Benchmark to measure the validity of our document embeddings. This is a SemEval task organized between 2012 and 2017. It consists of 8628 pairs of English sentences, selected from image captions, news headlines and user forums. The similarity between these sentence pairs was annotated using a five-point scale via crowdsourcing (Agirre et al., 2016). Participating systems calculate the similarity between these sentence pairs and are evaluated based on their Pearson correlation with the gold standard STS annotations. A higher score indicates more consistent between a system and human judgement.

We compared our method with Doc2Vec (Le and Mikolov, 2014), fastText (Bojanowski et al., 2016) and Sent2Vec (Pagliardini et al., 2018), on a subset of the MEDLINE dataset that consists of the metadata of $10^6$ MEDLINE articles selected from WorldCat.org. Each article

is written in English and has a title and an abstract. All experiments were carried out on the same server with 2 Intel Xeon Silver 4109T 8-core processors and 384 GB memory. The common hyperparameters were: a vector size of 256, a minimal number of word occurrences of 10, number of negative samples of 10, window size of 10, using hierarchical softmax, a learning rate of 1.0 and a number of threads of 16.

Table 1 shows the train times for different methods. Our method, that embedded $10^6$ MEDLINE articles and $340 \times 10^3$ unique words, required only 43 seconds, while the other methods spent more than 2 or 3 hours to finish the training. With the same training data, our method had the highest STS scores among all the methods.

**Table 1. STS scores and train times**

| Method | Dev | Test | Train time |
|---|---|---|---|
| Doc2Vec | 63.7 | 49.8 | 2h3m |
| FastText | 52.8 | 41.1 | 3h4m |
| Sent2Vec | 65.4 | 55.0 | 3h10m |
| Ariadne | 73.6 | 58.9 | 43s |

**(2) Subject prediction**

We also evaluated our embedding method on the use case of subject prediction. This remains a difficult problem and is a form of Extreme Multi-label Text Classification (XMTC) (Prabhu and Varma, 2014; Bhatia et al., 2015; Liu et al., 2017), where the prediction space normally consists of hundreds of thou- sands to millions of labels and data sparsity and scalability are the major challenges. In our MEDLINE dataset, there are more than 324,619 MeSH headings indexing 896,300 articles (the other articles do not have any subjects) with on average 16 headings per article. However, only 102,484 MeSH headings are used to index more than 10 articles.

We took two approaches of using semantic embedding for subject prediction, as described in Wang et al. (2019). The first approach is to directly leverage the article-entity similarities to predict the subjects that have the highest similarity scores to the article to be indexed. The second approach is non-parametirc, that is to collect the subjects that were used to describe the closest neighbours of the article to be index and use them to index it.

This task is to provide a shortlist of potentially relevant entities to the document at hand. It is important to present a ranked shortlist of candidate entities and to evaluate the quality of the prediction with an emphasis on the relevance of the top portion of such lists. Therefore, we use rank-based evaluation metrics such as precision and recall at top $n$. Precision@$n$ is the proportion of the predicted entities in the top $n$ list that are actual entities of the test document, while Recall@$n$ is the proportion of the correctly predicted entities over all actual entities of the test document.

Figure 1 shows the Precision@$n$ and Recall@$n$ of three methods, where *Ariadne* represents the straightforward predictions based on entity-document similarities and Ariadne+NPSP represents the non-parametric algorithm on top of Ariadne embeddings. We can see that the quality of the predicted subjects from our similarity-based prediction are comparable with

those generated by fastText. The precision of fastText is higher than our Ariadne method for low values of n while it quickly decreases to be worse than ours. Up to top 20 candidates, the recall for both Ariadne and fastText are more or less the same, but our method is able to predict more actual subjects at lower ranks, where the recall outperforms fastText.

The clear winner is the NPSP method. The precision and recall are both consistently higher than the other two methods. At n = 100, the recall is nearly 20% higher than the fastText predictions. More correct subjects are predicted at lower ranks, which explains the much slower decrease of precision with increasing rank.

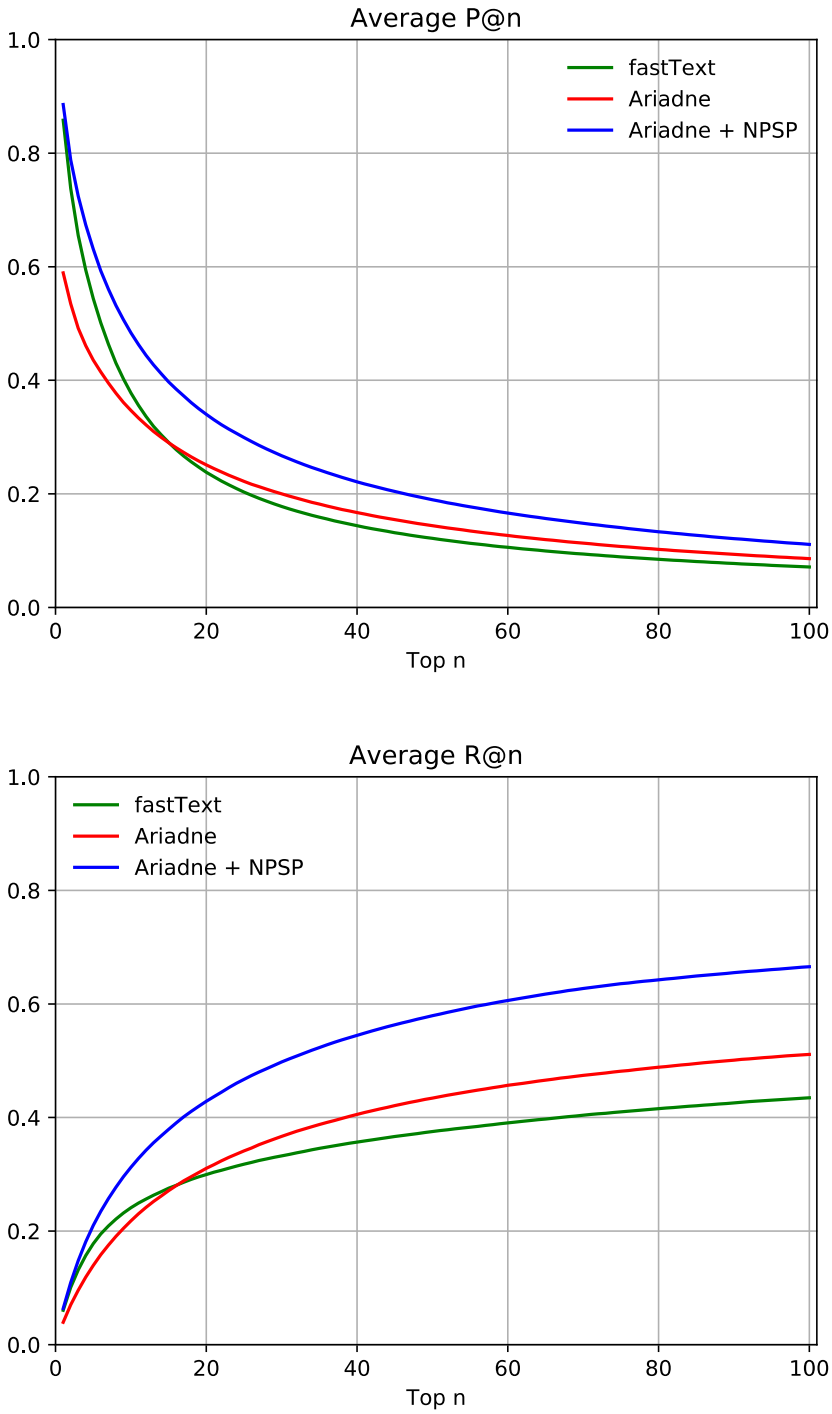


**Figure 1. Average precision and recall at n for subject prediction**

As presented in Wang et al. (2019), Ariadne embedding helps to identify more specific subjects that may not have a lot of training examples and therefore more difficult for a classifier to make the right predictions. This is important for producing useful subject prediction, which identify not only frequent general subjects, but also more specific subjects.

**(3) Dataset sensitivity**

Due to the high cost of applying deep learning embedding methods from scratch, transfer learning (using pre-trained embedding for downstream applications) is currently very common. However, it may fail to capture the domain-specific semantics which are important to the application at hand.

Table 2 shows the top 10 words that are most related to the word "young", calculated based on different datasets. The dramatic difference between "young" in the context of astrophysics and those in the other domains should raise some awareness for anyone who is trying to use pre-trained embeddings for specific domains.

**Table 2. Different datasets, different embeddings, different neighbours**

| Data set | Top 10 words related to "young" |
|---|---|
| WorldCat | people, children, adolescents, nobleman, christians, pianists, siblings, vietnamese, clergyman, housekeeper |
| Medline | adults, children, people, women, men, adulthood, infants, athletes, girls, leaves, patients, mania, boys, chicks, calves |
| Art library | people, children, persons, adults, lady, women, gentlemen, artists, readers, folks, americans, memorial, girls, architects |
| Astrophysics | stars, supernova, stellar, clusters, massive star clusters, brown dwarf |

## Conclusion

We have described a novel, simple, effective and efficient method for term and document embeddings. As we have shown, our method has important practical benefits: 1) it is fast and has low hardware requirements, having linear time complexity and constant space complexity in function of the number of documents, resulting in very short run-times in practice. 2) It computes semantically discriminative term embeddings and weightings with a single pass through the training data, and has the capacity to effectively include very rare words. Our experiments show it outperforms state-of-the-art methods in terms of the STS benchmark and subject prediction when trained on the same datasets, while at the same time being computationally cheaper by orders of magnitude.

We showed the usages of semantic embeddings in the subject prediction task that is common in the bibliographic world. We argue that semantic embeddings can help with other tasks conducted in libraries, but also be aware of the influence of the training data on the learned embeddings. Consider using a method like Ariadne to create useful embeddings for small coherent data sets rather than generating generic but less precise embeddings using larger heterogenous data sets.

**References**


Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences 66*(4), 671–687.

Agirre, E., C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wieb (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the SemEval-2016*.

Bhatia, K., H. Jain, P. Kar, M. Varma, and P. Jain (2015). Sparse local embeddings for extreme multi- label classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28*, pp. 730–738. Curran Associates, Inc.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Castermans, T., K. Verbeek, B. Speckmann, M.A. Westenberg, R. Koopman, S. Wang, H. van den Berg, and A. Betti. SolarView: Low Distortion Radial Embeddings with a Focus. In *IEEE Transactions on Visualization and Computer Graphics*, 2018

Furnas, G. W., T. K. Landauer, L. M. Gomez, and S. T. Dumais (1983). Statistical semantics: Analysis of the potential performance of keyword information systems. *Bell System Technical Journal 62*(6), 17531806.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146162.

Johnson, W. and J. Lindenstrauss (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Math. 26*, 189–206.

Le, Q. V. and T. Mikolov (2014, 5). Distributed Representations of Sentences and Documents. *Interna- tional Conference on Machine Learning - ICML 2014 32*, 11881196.

Liu, J., W.-C. Chang, Y. Wu, and Y. Yang (2017). Deep learning for extreme multi-label text classifica- tion. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, New York, NY, USA, pp. 115–124. ACM.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, USA, pp. 3111–3119. Curran Associates Inc.

Koopman, Rob, Shenghui Wang, Andrea Scharnhorst, and Gwenn Englebienne. 2015. "Ariadne's Thread: Interactive Navigation in a World of Networked Information." In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. New York: ACM, 1833–38.

Koopman, Rob, Shenghui Wang and Andrea Scharnhorst. 2017. "Contextualization of Topics: Browsing Through the Universe of Bibliographic Information." *Scientometrics* 111(2):1119–39.

Koopman, Rob, Shenghui Wang and Gwenn Englebienne. 2019. "Fast and Discriminative Semantic Embedding." In: *Proceedings of the 13th International Conference on Computational Semantics*, Long Papers, edited by Simon Dobnik, Stergios Chatzikyriakidis and Vera Demberg. Gothenburg: Association for Computational Linguistics, 235–46.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). Deep contextualized word representations. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Prabhu, Y. and M. Varma (2014). Fastxml: A fast, accurate and stable tree-classifier for extreme multi- label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, New York, NY, USA, pp. 263–272. ACM.

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguistica 20*(1), 3353.
Weaver, W. (1955). Translation. In W. Locke and D. Booth (Eds.), *Machine Translation of Languages*, pp. 15–23. Cambridge, Massachusetts: MIT Press.
Wang, S., Koopman, R.: Clustering articles based on semantic similarity. Gläser, A. Scharnhorst, W. Glänzel (eds.) Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of *Scientometrics* (2017) DOI 10.1007/s11192-017-2298-x http://rdcu.be/pDZH
Wang, S., Koopman, R.: Semantic Embedding for Information Retrieval. In: *Proc. of the 5th Workshop on Bibliometric-enhanced Information Retrieval* (BIR 2017), CEUR-WS.org (2017) 122–132
Wang, S., R. Koopman and G. Englebienne. Non-parametirc Subject Prediction. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries* (TPDL2019). To appear.
Pauline van Wierst, Steven Hofstede, Yvette Oortwijn, Thom Castermans, Rob Koopman, Shenghui Wang, Michel A. Westenberg, and Arianna Betti. BolVis: Visualization for Text-based Research in Philosophy. In proceedings of 2018 Workshop on Visualization for the Digital Humanities (VIS4DH).