Motivation

In-domain embeddings and validation
In earlier research concerning domain specific articles, **?**], found that in-domain training of the word embeddings can improve the process of document clustering. This effect is even stronger than the number of training examples and the model architecture. **?**] found that the corpus domain is more important than the corpus size. Using an in-domain corpus significantly improves the performance for a given task, whereas using a corpus in an unsuitable domain may decrease performance. Truong et al. encountered a problem in the validation of these in-domain embeddings. The word embedding produced correct document clustering results, leading to the conclusion that these embeddings are of good quality, since they capture the document relatedness needed to create correct clusterings. However unpublished results by Truong et al. state that the embeddings show high error rates on the validation scores. This seems to indicate that the word-vectors are of good quality, but that the available validation metrics fail to confirm this. **?**] used multiple word similarity validations to asses the quality of the word embeddings. However, these set are created to validate the generic embeddings, they fail to asses the quality of the domain specific embeddings.

Research
To asses the problem of the limited availability of pre-labelled validation sets for domain specific articles, we compare the embeddings to TF-IDF on a categorization task. This A) gives an indication of the embedding quality for categorization tasks and B) contrasts the performance of embeddings to the performance of the more traditional TF-IDF approach. To ensure the quality of the embeddings for our research, we reuse the embeddings created in the research of **?**].

RQ. 1 Have word-embeddings a higher accuracy for academic texts than TF-IDF for article classification?

RQ. 1.1 Does TF-IDF weighting increase the performance of word embedding on academic texts for article classification?
RQ. 1.2 Can the usage of alternative distance metrics improve the performance of word embedding on academic articles for article classification?

RQ. 2 Can word embeddings, combined with pca[1]-based TSNE, create a two-dimensional plot that preserves the journal relatedness?

Embedding and TF-IDF
RQ. 1 focusses on the classification results of both embedding-based techniques and TF-IDF. To measure classification task we use the rank of the class to which the item belongs. Transforming the classification task from a binary metric to a ranking metric. For this task, we will use different versions of embeddings, to answer RQ 1.1, and different TF-IDF versions to not only compare the two techniques, but also look for the optimal results of both techniques. To achieve the optimal results, we compare 20 distance metrics from the SciPy library on ranking performance on the standard embeddings. For this part of the research, we will use the following hypothesis:

H. 1 *Embedding based techniques give lower rankings than the TF-IDF based techniques.*
H. 1.1 *TF-IDF weighted document embeddings outperform standard embeddings on the classification of academic articles.*
H. 1.2 *Cosine similarity based ranking results in the best performance for the classification of academic articles.*
H. 1.3 *Word embeddings use less memory while giving better ranking results than TF-IDF on the classification of academic articles.*

By validating our invalidating these hypotheses we get an indication of the performance of the embeddings compare to older techniques, get insight into possible performance and resource trade-off's and get insight into the performance of different distance calculation metrics. RQ. 2 concerns the visualization of word embeddings and the accuracy of this visualization. To answer this research question, we will use the following hypothesis:

H. 2 *Word embeddings, combined with pca-based TSNE can preserve the journal relatedness on a two dimensional plot.*

The validation of this hypothesis will rely on visual conformation. We expect to see clustering of journals

[1] principal component analysis

in certain areas, which indicates a "research subject". We also expect that articles which are visually close together are closely related by subject.