1 Motivation

1.1 In-domain embeddings and validation

In previous research concerning domain specific articles, **?**] found that in-domain training of the word embeddings can improve the process of document clustering[1]. The usage of in-domain data is more important 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 clustering. However unpublished results[2] by Truong et al. state that the embeddings show high error rates on the word-relatedness validation scores. This seems to indicate that the word-vectors are of good quality due to the correct results on document clustering, but that the word-relatedness validation task has been unable to confirm this due to the domain-specific nature of the texts used by Truong. **?**] used multiple word similarity validations to asses the quality of the word embeddings. However, these sets are created to validate the generic embeddings; they fail to asses the quality of the domain-specific embeddings.

1.2 Research

To solve the problem of the limited availability of pre-labeled validation sets for domain-specific articles, we compare the embeddings to TF-IDF on a categorization task. This (a) indicates 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 than TF-IDF for academic texts on the task of article classification?

RQ. 1.1 Do embedding optimizations 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 texts for article classification?

RQ. 2 Can word embeddings, combined with PCA[3]-based TSNE[4], create a two-dimensional plot that preserves the relatedness between journals?

[1] Document clustering aims to discover underlying structures in texts**?**].

[2] Obtained from author.

[3] PCA is the abbeviation for principal component analysis

[4] TSNE (t-Stochastic Neighbor Embedding) was introduced by **?**]. The technique is used to reduce the dimensions of a vector

Methodology

RQ. 1 focusses on the classification results of both embedding-based techniques and TF-IDF. To asses the classification task, we use the rank of the class to which the item belongs. This transforms 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 ranking results of both techniques. To achieve optimal results, we also validate distance metrics, we compare 20 distance metrics on ranking performance on the embeddings. For this part of the research, we will use the following hypothesis:

> ***H. 1*** *Embedding based techniques give lower rankings (i.e. better performance) 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*** *The cosine similarity metric produces the has performance (i.e. lowest ranks) 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 or invalidating these hypotheses, we get an indication of the performance of the embeddings compared to TF-IDF, get insight into possible performance and resource trade-offs 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 confirmation. We expect to see clustering of journals in certain areas, which indicates a research subject. We also expect that journals which are visually close together are closely related by subject.