Motivation

Domain specific
Earlier research, concerning domain specific articles, by **?**], 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. These findings both indicate that an in-domain corpus improves the performance of word embeddings for the specific domains.

Problems in validation
To assess the performance of the embeddings, validation methods are used. This are tasks designed to produce a metric that gives an indication of the usability of the provided embeddings. **?**] found that the validation method indicates only the quality of an embedding for a specific task. There is (yet) no methtod that can asses the usability of an embedding on all possible tasks, since each task may require other information to be embedded into the embedding. Validation methods use either labelled or unlabelled data. Labelled data is data that is in some way marked, so that the correct answer can be derived from it. Unlabelled is the opposite, this data is not marked.
The usage of labelled data is common practice for validation methods, since the results produced by this data can be easily checked checked. Unpublished results of the study by Truong encounter this problem, they show high error rates on the validation scores, presented in Table **??**. However, the word embeddings created correct document clusterings[**?**], this seems to indicate that the word-vectors are able to represent the words correctly but that the available validation sets cannot confirm this.
Furthermore, a study by **?**] found that the quality of embeddings are tasks specific, *different tasks favour different embeddings*. They also found that the embeddings encode information about word frequency, even in models that are created to prevent this. *This casts doubt on the common practice of using vanilla cosine similarity as a similarity measure.*
Therefore, we propose the validation of domain specific word-embeddings through a classification tasks, using multiple vector-distance calculations. This eliminates the need for labelled data in the validation of these domain specific word embeddings, will validate the quality of word embeddings for domain specific texts, and will validate the impact of different vector-distance measures on a categorization task.

	WordSim	Men	RareWords
Best results from the research by Truong:	0.49	0.61	0.32
Average results from the research by Truong	0.45	0.59	0.32

Table 1: Results for the different validation sets of word simularity validations on domain specific texts from the study by **?**]

Research Questions

- Have word-embeddings a higher accuracy for academic texts than TF-IDF for article classification?
- Which metric(s) can be used to measure the accuracy of word embeddings for scientific articles?