# Semantic Textual Similarity of Course Materials at a Distance-Learning University

Niels Seidel
FernUniversität in Hagen
niels.seidel@fernuni-hagen.de

Moritz Rieger
FernUniversität in Hagen
moritz.rieger@posteo.de

Tobias Walle
FernUniversität in Hagen
tobias.walle@student.fernuni-hagen.de

## ABSTRACT

Choosing computer science courses from a wide range of courses is not an easy task for students - especially in the first semesters. To overcome the shortcomings of course descriptions and vague recommendations by acquaintances, we provide a method to identify and visualize semantic similarities between courses using textual learning materials. To achieve this goal, a complete set of course materials (94 courses, 572 course units / PDF textbooks) from the Faculty of Mathematics and Computer Science at FernUniversität in Hagen was vectorized as document embeddings and then compared using the cosine similarity of the vectors. The process can be fully automated and does not require labeled data.
The results were compared with the semantic similarity assessed by domain experts. Also the similarity of consecutive courses and sections within the same course have been evaluated against the average similarity of all courses.
The presented approach has been integrated into a course recommendation system, a course dashboard for teachers and a component of an adaptive learning environment.

## Keywords

NLP, Semantic Textual Similarity, Document Embedding, Educational Data Mining

## 1. INTRODUCTION

Before each semester, students are faced with the question of which course to take. In order to achieve the goal within a course of study, the examination regulations contain information on the optional and compulsory modules and courses. Study plans of the Student Advisory Service flank this framework with recommendations on the number, sequence and selection of courses for the individual semesters. Ultimately, the dates of the courses result in further organizational requirements with which the individual timetable must be brought into line. Despite these organizational restrictions, the internal autonomy of the universities opens up many options for selecting courses according to content criteria and interests. However, only module handbooks and course websites are usually available for decision-making purposes. Learning materials published in advance as textbooks or in the sense of OER are the exception. In both cases, however, the amount of information is difficult to manage. The linear format of the module manuals, which often contain more than one hundred pages, makes it difficult to identify courses that are similar in content and build on each other. Moreover, the concise descriptions of the modules represent only a fraction of the learning content. Courses that are not assigned to the course of study will, of course, not appear in the module handbook. Many students therefore seek advice from friends and fellow students or follow recommendations from teachers. However, prospective students and first-year students do not yet have these contacts. This challenge becomes particularly clear when looking at the example of the FernUniversität in Hagen. With over 74,000 registered students and a course offering of over 1,600 courses, the distance-learning university is the largest university in Germany. The Faculty of Mathematics and Computer Science alone accounts for 134 courses, of which 94 are courses and 40 are seminars or internships. For students at the faculty, choosing from this large number of courses is a particular challenge. In contrast to attendance universities, it is usually not possible to benefit from the experience of fellow students. Furthermore, the authors or supervisors are usually not personally known to the students, so that contacts with lecturers can hardly make the decision easier. Students can use the short descriptions of course contents and learning goals in the module manuals (approx. 100-200 words) as well as short readings of one chapter of the script for decision-making. For universities with a very large number of courses and a very wide range of options, the planning of the study program is therefore time-consuming and complex.

Teachers who wish to avoid redundancies to other courses and who wants to build on previous knowledge or develop the same for other courses when planning and creating learning materials face a similar hurdle. In view of the large number of courses, however, the people concerned do not always know exactly what their colleagues teach in detail in their courses. Consequently, overlaps and gaps in content remain undetected and potential for cooperation in the field of teaching is not recognized.

In this paper a method for the analysis of semantic similarity of courses using text-based learning materials is presented.

In the second section, related works regarding methods for determining the semantic similarity of texts as well as on course selection recommendation systems will be presented. Subsequently, the method document embeddings used here for the analysis of semantic similarities is presented in section 3 using the example of a corpus of 94 courses of the FernUniversität in Hagen. The results will be evaulated in section 4. Based on the semantic relations of the course materials, we present three prototypical applications in section 5: i) a tool for exploration and recommondation of courses, ii) a teacher dashboard, and ii) an adaptive course recomondations for long study texts. The article ends with a summary and an outlook.

## 2. RELATED WORKS

The processing of natural language using Natural Language Processing (NLP) techniques has made enormous progress in recent years. Conventional NLP methods generate from a text document by Bag of Words (BOW), frequency-based methods like Term Frequency Inverse Document Frequency (TF-IDF), Latent Dirichlet Allocation (LDA) etc. vectors and calculate the distance between the vectors [19]. However, these methods cannot capture the semantic distance or are very computationally intensive [21] and usually do not achieve good results. Newer machine learning methods achieve much better results in the analysis of semantic text representations [10]. A central challenge is the determination of Semantic Textual Similarity (STS), which is used in machine translation, semantic search, question-answering and chatbots. With the help of developments in the field of distributed representations, especially neural networks and word embeddings such as Word2Vec [21] and Glove [23], semantic properties of words can be mapped into vectors. Le and Mikolov have shown with Doc2Vec that the principles used can also be applied to documents[17].

The similarity of extensive book collections has so far been investigated in only a few works. The SkipThrough Vectors presented by Kiros et al. train an encoder-decoder model that attempts to reconstruct the surrounding sentences of an encoded passage [16]. However, the experiments were based on a relatively small body of only 11 books [30]. Spasojevic and Pocin, on the other hand, determined the semantic similarity at the level of individual pages and entire books for the corpus of Google Books, which contains about 15 million books [26]. The similarity of two books was determined from the Jaccard index of the permuted hash values of normalized word groups (features). Liu et al., however, point out that the semantic structure of longer documents cannot be taken into account in this way and therefore propose the representation as a Concept Interaction Graph [20]. Keywords are determined from a pair of documents and combined into concepts (nodes) using community detection algorithms. These nodes are connected by edges that represent the interactions between the nodes based on sentences from the documents. Although the method seems very promising, it has so far only been investigated on the basis of news articles. The SemEval-2018 Task 7 [12] pursues a similar goal as this paper, for example, with regard to STS, where semantic relations from abstracts of scientific articles are to be found. The gold standard used for evaluation is based on named entities (persons, places, organizations), which cannot be annotated with reasonable effort for large amounts of text.

Brackhage et al. had experts manually keyword module descriptions of several universities and visualized these data together with further metadata in a web application as a forced layout graph and adjacency matrix heatmap and made them searchable with the help of complex filters [7]. However, keywording proved to be extremely time-consuming and has to be updated frequently. Baumann, Endraß and Alezard used study history data to visualize "on the one hand the distribution of students across the modules in a study program and on the other hand the distribution of students in a module across different study programs" [4], without, however, concretizing their benefit for the intended support in the choice of courses. Askinadze and Conrad used examination data from one study program to illustrate the progress and discontinuation of studies in various visualisations [3]. However, there is a large number of applications and approaches to recommend courses to students. Lin et al. used the sparse linear method to develop topN recommendations based on occupancy data of specific groups of students [19]. With the help of K-Means and Apriori Association Rules, Aher and Lobo presented a recommendation system for courses in the learning management system Moodle [2]. Zablith et al. present several recommendation systems based on linked data from the Open University UK[1] [29]. The Social Study application, for example, suggests courses to learners based on their facebook profile, while Linked OpenLearn offers media and courses to the distance learning university's OER to learners. The recommendations are based on course-related metadata and links to other courses and media, but do not consider the semantics of the courses. D'Aquin and Jay try to reconstruct the missing semantics with the help of different linked data sources (e.g. DBpedia) in order to trace frequently occurring course occupancy (frequency sequences) [11]. The analysis of semantic similarities on the basis of the complete learning materials does not only provide insights into the content relations of learning resources but also opens up the possibility to understand the temporal structure and patterns of course assignments for the decision making process when choosing a course.

From the perspective of course planning, Kardan et al. have developed a prediction model for the number of course bookings with the help of a neural network [15]. Ognjanovic et al. have also modeled the course occupancy for several semesters in advance [22]. However, the authors of this paper could not find any contributions in the literature for a didactically motivated use of occupancy statistics. The same applies to the use of these data for the modeling of learners within adaptive or at least personalized learning environments.

## 3. DETERMINATION OF THE SEMANTIC SIMILARITY OF TEXTS

In this section, a procedure for analyzing the semantic similarity of courses using text-based learning materials is presented using the example of the study texts of the Faculty of Mathematics and Computer Science of the FernUniversität in Hagen. This procedure consists of four steps, which are mainly based on the work of [blinded]. First a corpus of course materials is created. Then these data are vectorized to determine the similarity in the third step. Finally,

[1]See `http://data.open.ac.uk/` (last accessed 15.06.2020).

an evaluation with a gold standard and other comparison parameters is carried out.

A corpus is a collection of related documents. In order to create a corpus, source data of 94 courses from all 20 subject areas of the faculty were available. A course consists of 3 to 10 documents, that we call course units or units. The course units were available as PDF documents that have between 20 and 60 pages. The PDFs differed in terms of their format (e.g. PDF/A, PDF/X), the PDF versions and the tools used to create them. The formatting of the type area was also not uniform. For these reasons, a programmatic extraction of chapters using regular expressions and PDF outlines proved to be unreliable and had to be discarded. The cover pages as well as redundant tables of contents and keyword indexes within a course were removed. The PDF documents were therefore first converted to text and divided into sentences and words using NLTK [5]. To avoid errors with mathematical formulas and dotted lines in the table of contents, the text was also normalized. The resulting corpus contains 572 course units, consisting of 654,367 sentences with a total of 9,507,770 words. The vocabulary contains 179,078 different words. Document Embeddings, also called Paragraph Vectors (PV) by Le and Mikolov, were used to vectorize the documents [17]. Since document embeddings are based on word embeddings, they must be created first. For this purpose, the words in one-hot-encoding enter a neural network. This serves an estimation task, whereby the word most likely to be in the context of a word is to be estimated. The neural network is trained with tuples from the text. For this purpose, a window is pushed through the entire corpus and the resulting tuple combinations are noted within the window. By feeding back the estimation error into the re-estimation, the weights of the weight matrix $W$ are optimized. This has the consequence that the weights of the estimation task for semantically close words assume similar values, since comparable tuples were used in the training. The weights of the estimation task represent the word embeddings.

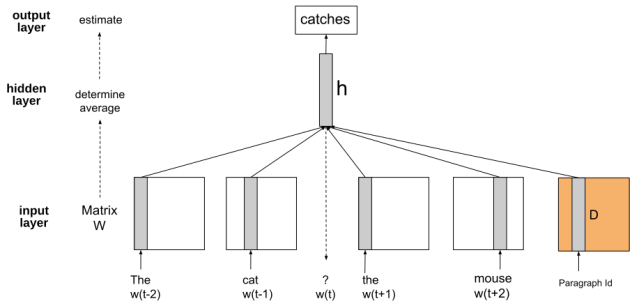


**Figure 1: Continuous Bag-of-Words model as well as the paragraph ID (orange), which is included in the estimate of w(t) in addition to the context words for document embeddings.**

In order to be able to represent whole documents semantically as vectors, the idea of word embeddings is extended to whole texts. For this purpose, a paragraph vector, a column of another weight matrix $D$, is combined with word vectors to estimate the next word from a given context (see Fig. 1). Since the word vectors capture the semantics of the words as an indirect consequence of the estimation task, this is done in a similar way with Document Embeddings. One can imagine the training of the PV as the training of another word. A PV acts as a kind of memory that contains information about missing words in the context within a document. For this reason this model is also called Distributed Memory Model of PV. Building on the Word Embeddings, PVs have been trained to represent entire documents. Now the PVs can be processed as characteristics of the documents to recognize semantic similarities of the documents. Before the documents are compared with each other, the semantic similarity of texts is first examined in general. To find a commonality of all terms, similarity has to be thought of as a "complex network of similarities" [28] of different entities. This complex form of similarity of natural language means that two documents cannot be considered semantically similar on the basis of common features, but that similarity is to be understood as the interaction of many direct and indirect relationships between the words contained in them [13]. This concept of similarity is taken into account in the training of word embeddings. The weight matrix, which ultimately contains the word embeddings, is the result of the use of the words in all contexts of the entire text corpus and thus represents the complex network of similarities described by Wittgenstein. To compare Distributed Representations, the cosine similarity is usually used as a metric [27, 13]. For normalized PV there is a linear relationship to the Euclidean distance.

# 4. EVALUATION OF NLP SYSTEMS

Since vectorizing the documents as PV is an Unsupervised Machine Learning method, there is no underlying test data against which the system can be tested. Following the SemEval competitions, a gold standard was therefore developed, which consists of a set of test and training data. However, this gold standard could not be generated by crowdsourcing [1], as is the case with many SemEval tasks, since a high degree of competence in the respective fields of knowledge is required for the assessment of semantic similarity. For this reason, three experts, which are authors of course texts themselves, were asked to compare one of their own courses with a course that they thought was similar. As an incidence they selected 6 unique courses. By evaluating documents that are related in terms of topic or content, monotonous gold standards that do not show any similarities could be avoided. Each of the three experts evaluated two courses which consisted of 4 and 7 units each. Each evaluator had thus made 28 comparisons. The similarity was indicated on a continuous scale from 0 (not similar) to 100 (identical). A nominal gradation of the scale was omitted due to expected problems of understanding with regard to the valence and equidistance of the scale values. Half of the gold standard data was used for training different hyperparameters, as shown in Fig. 3. The hyperparameters were composed of the window size of the Continuous Bag of Words, the dimension of the PV and the minimum frequency of occurrence of the words considered. The values for the individual parameters are based on plausibility tests and are within the value ranges known from literature (e.g. [21]). To avoid overfitting, the values for each parameter were only roughly graded. The minimum mean square error could be determined for a window size of 20, a dimension of

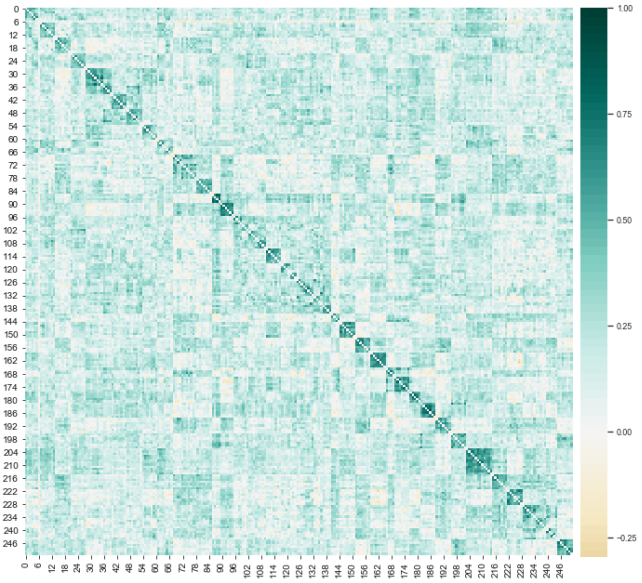

**Figure 2: Adjacency matrix of course unit relations. Courses are represented by a running number along the axis. The darker the boxes, the greater the semantic similarity. The dark colored rectangular artifacts along the diagonals indicate the high similarity of units of the same course.**

PV of 140 and a value of 20 as the lower limit for the frequency of occurrence of words (see solid blue dot in Fig. 3). Based on these hyperparameters a model was trained and tested with the second part of the gold standard (test data). Pearson's r as a measure of the linear relationship reached a value of 0.598. However, since in the present case whole documents were compared instead of individual sentences, the gold standard and the PV are more fuzzy. Fine fluctuations in cosine similarity are not reflected in the gold standard. However, in view of the subjective assessment on the continuous scale, which can be freely interpreted by the evaluators, the value for Pearson's r must be regarded as high. To establish the monotonous relationship between cosine similarity and the gold standard, Kendall's $\tau$ was determined as a rank correlation coefficient with a value of 0.451. In general, smaller values of correlation are obtained for Kendall's $\tau$ compared to Pearson's $r$. However, the low value is also due to the individual definition of the concept of similarity and the individual mapping of the subjectively perceived similarity to the scale. Looking at the areas of high similarity shown in Fig. 4, the correlation is more obvious.

In addition to the gold standard, the NLP system was checked for plausibility of the results. Two hypotheses were put forward for this purpose:

H1 Course units of consecutive courses are more similar than units of other courses.

H2 Course units of one course are more similar than units of other courses.

In order to test the first hypotheses, eight courses were initially identified which, given the numbering contained in the course title, clearly build on each other. The mean cosine similarity of the consecutive courses is 0.32, which is above the average of the whole corpus (0.18). Hypothesis 1 is thus confirmed. The second hypothesis could already be recognized by the strongly colored rectangular artifacts along the diagonals in the adjacency matrix in Fig. 2. The mean similarity of course units is 0.51 and is thus significantly greater than the mean cosine similarity of the whole corpus (see Fig. 5). Hypothesis 2 is therefore also confirmed. A further part of the plausibility check consisted, among other things, in excluding undesired effects of the document size on the semantic similarity. There is no correlation between the difference in the word count of two documents and their cosine similarity ($r = 0.013$).

# 5. APPLICATIONS

## 5.1 Course exploration and recommendation

The hurdles in the choice of courses addressed in the introduction to this paper address an application in which learners can explore the semantic similarity of courses and course units by means of visualizations in the form of chord diagrams, forced layout graphs and heat maps. These node-link diagrams are primarily suitable for small graphs, since the visualization quickly becomes confusing due to overlapping edges. Heatmaps in particular, may contain many nodes, but require a lot of space. Their readability depends largely on the arrangement of the elements. Due to this limitation, it seemed necessary to realize the exploration over the entire set of courses not graphically, but textually. Besides the given structuring of the courses according to study programs, chairs and lecturers, we tried to identify overlapping topics. Using Latent Dirichlet Allocation 11 topics were determined based on the word distribution [24]. For each topic the 20 most weighted terms were displayed in a word cloud. After the user has made pre-selection (e.g. by choosing a topic), a limited set of up to 20 courses including their course units can be explored. For this pupose various interactive node-link diagrams were created as Data Driven Documents [6].

The recommendation of courses is based on two approaches. Firstly, other courses with a high cosine similarity were proposed for a course. The suggestions were justified by a list of the particularly similar course units (see Fig. 6). In this way, the algorithmic decision can be understood on the basis of the available texts.

Secondly, the Alternating Least Squares Algorithm by Hu, Koren, and Volinski [14] was used for collaborative filtering in order to create a recommendation system based on the courses that of other students have been enrolled to in the past. Collaborative filtering often works with explicit feedback based on user ratings. However, course enrollment data does not express an assessment but a learner's preference, which is called implicit feedback. By choosing a course, a student indirectly expresses his or her preferences. Students who have taken similar courses may be interested in similar courses in the future. The numerical result of the implied feedback indicates the confidence, but not the students' preference for a course. The user behavior can be used to deduce which courses the user is likely to prefer. Fig. 7 shows

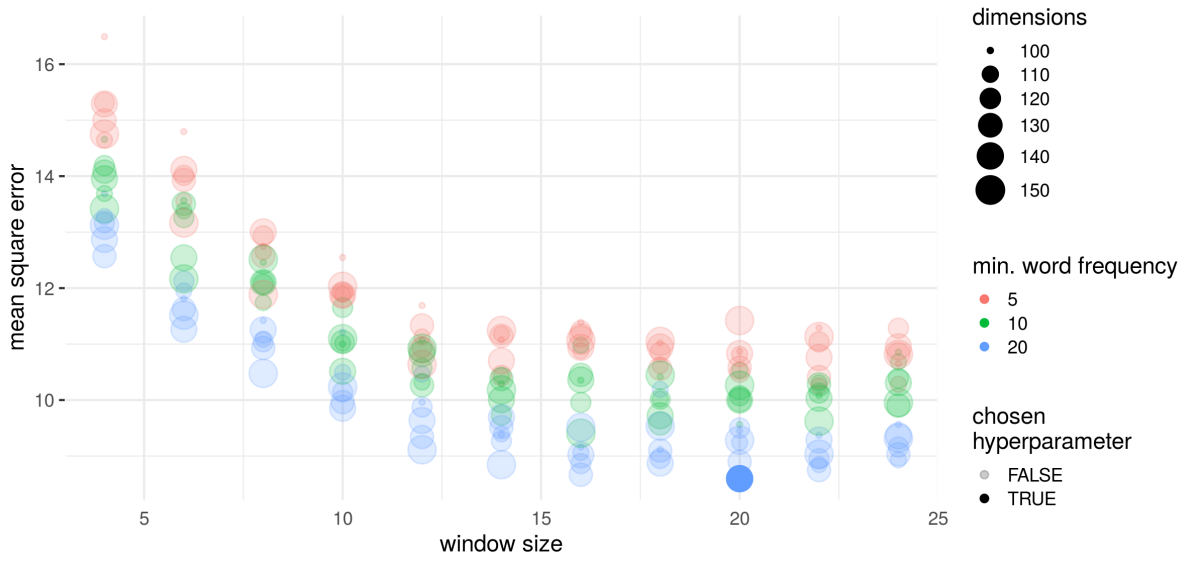


**Figure 3: Minimizing the mean square error for multiple configurations of hyperparameters. Each dot represents a hyperparameter configuration. The highlighted dot in solid blue represents the best parameter combination.**

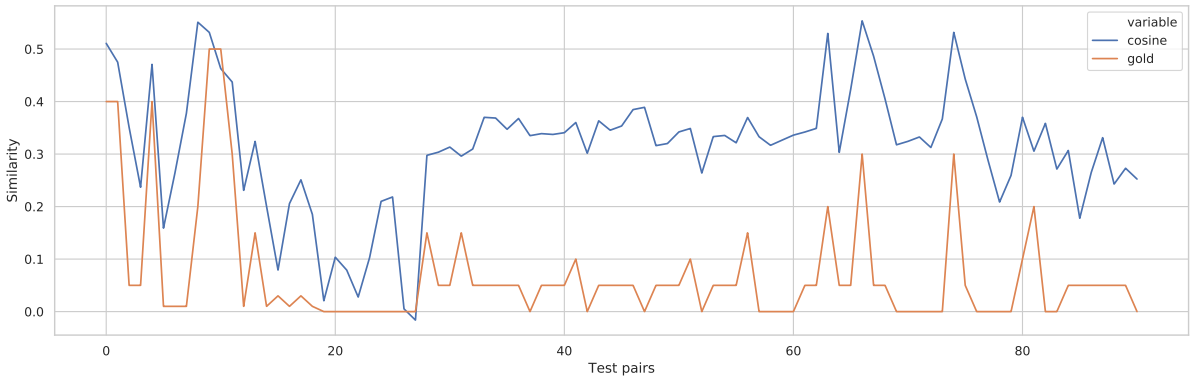


**Figure 4: Ratio of gold standard (orange) to cosine similarity (blue) for the individual test pairs**

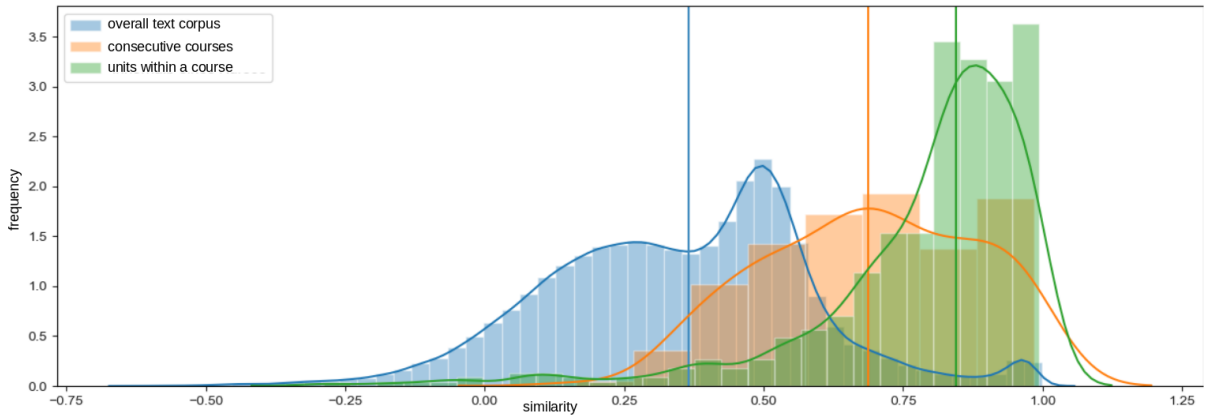


**Figure 5: Distribution and mean value of the cosine similarity in the entire corpus (blue), between the consecutive courses (orange), and the course units (green).**

a screen grab of the recommender system.

The filtering procedure described here only briefly has clear limits. For example, the order in which courses are taken is not considered. However, this can have a high relevance, as a student should not be recommended to take any more basic courses at the end of his studies. The method always interprets the attendance of a course as a positive factor. However, this is not always the case, for example, because a student attends a course but has not perceived it as interesting or valuable. Furthermore, there are compulsory modules in many courses of study, which must be attended in any case. However, this is a general disadvantage of recommendations based on implicit feedback. The chosen approach of collaborative filtering cannot make recommendations for prospective students who have not taken a course. In this case, however, the usual introductory courses of a degree program can be recommended. Besides the examination of certain subjects the course choice is not constraint by study regulations or other pre-requesites at our faculty. Such constraints might have to be considered for course recommender systems.


**Figure 6: Course details view with a list of related courses**

## 5.2 Teacher dashboard

The second application scenario is primarily aimed at teachers and authors of learning materials. In a Learning Analytics Dashboard [25] course occupancy statistics are linked with the semantic relations of the course materials. By including the semantic textual similarity of other courses and course chapters, responsible teachers can identify connections to other courses and identify possible content duplications. The dashboard consists of six tiles in a three-column layout: (1) An adjacency matrix shows the similarity of the course units contained in the course (Fig. 8, left). (2) The five most similar courses are shown in a matrix (Fig. 8, middle). (3) A line chart shows the course attendance of the last few years (Fig. 8, right). In addition, the dashboard contains statistics of the most frequently (4) previously, (5) simultaneously and (6) subsequently attended courses in the form of horizontal bar charts.

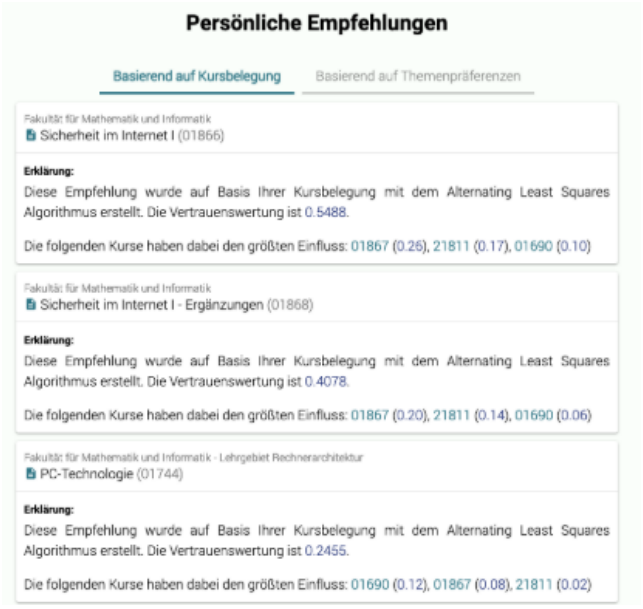


**Figure 7: Course recommendations based on the individual course of study and the data on the enrollment of all students in the study program**

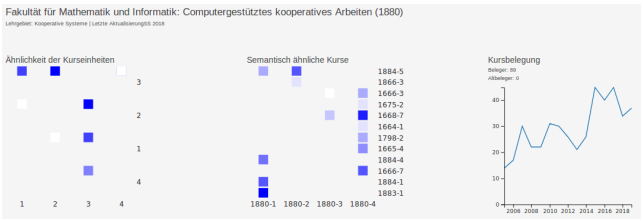


**Figure 8: Extract from the dashboard for teachers**

## 5.3 Adaptive course recommendations for long study texts

In the third use case, adaptive navigation support in the sense of direct guidance [8] was integrated in the online learning environment Moodle. The Moodle standard page plugin (mod_page) has been enhanced for the readability of long texts [18], so that the course texts, some of which are over 60 DIN-A4 pages long, can also be used on screen.

The marginal columns of the text are used to point readers to chapters of other courses that are very similar to the currently displayed text paragraph ,The recommendations are limited to two links per text paragraph. No recommendations are made for paragraphs of less than 100 words. The threshold value for the degree of similarity was chosen relatively high in order to avoid recommendations of courses that show only a little similarity.
In terms of adaptive learning it is taken into account whether the learner has already taken the recommended course. This information will be analyzed in relation to the learning progress in the current Moodle course. In case of a lower progress and comparatively low quiz results and only a few points achieved in the assignments we want to encourage the learner to make use of his previous knowledge, which he

has acquired in previous courses. Consequently, the recommended links point to courses that the learner already know and which are semantically related to the currently displayed text paragraph. In the second case high performing students or those who almost completed the current Moodle course will be provided with links to courses they have not enrolled so far. Often these are more advanced courses, if the students are in the beginning of their studies or if they have already enrolled to the primitive courses. In this way, we would like to encourage students to deepen their knowledge in a specific area through targeted course recommendations.

# 6. CONCLUSION AND OUTLOOK

An expandable corpus of the Faculty of Mathematics and Computer Science of the FernUniversität in Hagen was created. Special attention was paid to the fact that this corpus can be extended without manual effort. The corpus allows a storage-efficient access to single course units or to several units per faculty, chair and course, so that it can serve as a basis for further studies. Subsequently, methods for feature extraction of the documents were investigated. The focus was on the mapping of semantics in the vector representation. For the selected PV model from [17] it was shown that PV can map semantic information even in texts with several thousand words. The results were evaluated with a gold standard and show a high correlation to it. In relation to comparable studies (e.g. [9]), this paper compared much larger texts with several thousand sentences instead of just individual sentences, which can be more precisely semantically assigned. By means of Word and Document Embeddings, the similarity of two courses can be justified to the users of the system by considering the subordinate course units belonging to a course. In a next development step, a chapter-by-chapter or page-by-page analysis could make the relations of the units comprehensible by means of the relations of the chapters contained in the course. In order to improve the reliability of the evaluation, we have presented an approach to define a gold standard and two metrics (H1 and H2) for assessing STS for larger texts. However, the gold standard needs to be extended to make better conclusions about the quality of the approach. However, there is also a need for other metrics that can be determined with less effort in order to large text similarity.

In this article it was shown by way of example how the STS can be examined by extensive textual learning resources of a distance-learning university. However, the methods are also transferable to traditional universities, which work more with presentation slides and online resources. Furthermore, it is conceivable to compare courses and study programs of different universities [7] and thus facilitate the choice of study places. From the administrative perspective of course planning and accreditation further fields of application of the technology could arise. This only works as far as textual representations of learning materials such as presentation slides, video transcripts or online courses cover the content of a course.

The STS approach used here is subject to some limitations, which at the same time indicate a need for further research. In connection with documents embeddings, intrinsic information on the content of the documents has not been considered so far (see [10] and LDA or LSA). Homonyms have not been considered either, but could be learned from labeled texts and applied to other texts. In order to be able to reproduce the learning materials of a course completely in the corpus, texts from diagrams and other visualizations should also be included. The possible applications shown in section 5 illustrate possible fields of application for the use of semantic relations of study texts, but require further investigation – especially user studies.

In all three use cases it becomes clear that the textual similarity of the learning materials alone is not sufficient to recommend courses, present comprehensive data for course authors or make meaningful recommendations in an adaptive learning environment. Apart from that, the identification of course duplicates and overlaps might be another interesting use case for the corpus of study materials. In order to enable further research of this kind, we are trying to publish the text corpus as research data.

# 7. REFERENCES


[1] E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre. SemEval-2012 task 6: a pilot on semantic textual similarity, 2012.

[2] S. B. Aher and L. Lobo. Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. *Knowledge-Based Systems*, 51:1–14, 2013.

[3] A. Askinadze and S. Conrad. Development of an Educational Dashboard for the Integration of German State Universities' Data. In *Proceedings of the 11th International Conference on Educational Data Mining*, pages 508–509, 2018.

[4] A. Baumann, M. Endraß, and A. Alezard. Visual Analytics in der Studienverlaufsplanung. In *Mensch & Computer*, pages 467–469, 2015.

[5] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, 2009.

[6] M. Bostock, V. Ogievetsky, and J. Heer. $D^3$ Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309, 2011.

[7] N. Brackhage, Carsten Schaarschmidt, E. Schön, and N. Seidel. ModuleBase: Inter-university database of study programme modules, 2016.

[8] P. Brusilovsky. Adaptive Navigation Support. In *The Adaptive Web: Methods and Strategies of Web Personalization*, pages 263–290. 2007.

[9] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. *arxiv.org*, 2017.

[10] A. M. Dai, C. Olah, Q. V. Le, T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1507.0, jul 2013.

[11] M. D'Aquin and N. Jay. Interpreting data mining results with linked data for learning analytics: motivation, case study and directions. In D. Suthers and K. Verbert, editors, *Third Conference on Learning Analytics and Knowledge, LAK '13, Leuven, Belgium, April 8-12, 2013*, pages 155–164. ACM, 2013.

[12] K. Gábor, H. Zargayouna, I. Tellier, D. Buscaldi, and

T. Charnois. Exploring Vector Spaces for Semantic Relations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1814–1823, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

[13] E. Grefenstette. *Analysing Document Similarity Measures*. PhD thesis, University of Oxford, 2009.

[14] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272, dec 2008.

[15] A. A. Kardan, H. Sadeghi, S. S. Ghidary, and M. R. F. Sani. Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, 65:1–11, 2013.

[16] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-Thought Vectors. *CoRR*, abs/1506.0, 2015.

[17] Q. V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. *jmlr.org*, 2014.

[18] Q. Li, M. R. Morris, A. Fourney, K. Larson, and K. Reinecke. The Impact of Web Browser Reader Views on Reading Speed and User Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[19] J. Lin, H. Pu, Y. Li, and J. Lian. Intelligent Recommendation System for Course Selection in Smart Education. *Procedia Computer Science*, 129:449–453, 2018.

[20] B. Liu, T. Zhang, D. Niu, J. Lin, K. Lai, and Y. Xu. Matching Long Text Documents via Graph Convolutional Networks. *CoRR*, abs/1802.0, 2018.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. 2013.

[22] I. Ognjanovic, D. Gasevic, and S. Dawson. Using institutional data to predict student course selections in higher education. *The Internet and Higher Education*, 29:49–62, 2016.

[23] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[24] R. Rehurek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

[25] B. Schwendimann, M. Rodriguez-Triana, A. Vozniuk, L. Prieto, M. Boroujeni, A. Holzer, D. Gillet, and P. Dillenbourg. Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, PP(99):1, 2016.

[26] N. Spasojevic and G. Poncin. Large Scale Page-Based Book Similarity Clustering. In *ICDAR 2011*, 2011.

[27] S. M. Weiss, N. Indurkhya, and T. Zhang. *Fundamentals of Predictive Text Mining*. Texts in Computer Science. Springer London, London, 2015.

[28] L. Wittgenstein. Philosophical Investigations. In *New York: The Macmillan Company*, page 272. Blackwell, 1953.

[29] F. Zablith, M. Fernandez, and M. Rowe. Production and consumption of university Linked Data. *Interactive Learning Environments*, 23(1):55–78, 2015.

[30] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv e-prints*, page arXiv:1506.06724, jun 2015.