Teaching data science in school: Digital learning material on predictive text systems

Stephanie Hofmann, Martin Frank

Teaching data science in school: Digital learning material on predictive text systems

Stephanie Hofmann[1] and Martin Frank[2]

[1]Karlsruhe Institute of Technology, Steinbuch Centre for Computing, Germany;
stephanie.hofmann@kit.edu

[2]Karlsuhe Institute of Technology, Steinbuch Centre for Computing, Germany;
martin.frank@kit.edu

Data science and especially machine learning issues are currently the subject of lively discussions in society. Many research areas now use machine learning methods, which, especially in combination with increased computer power, has led to major advances in recent years. One example is natural language processing. A large number of technologies and applications that we use every day are based on methods from this area. For example, students encounter these technologies in everyday life through the use of Siri and Alexa but also when chatting with friends they are supported by assistance systems such as predictive text systems that give suggestions for the next word. This proximity to everyday life is used to give students a motivating approach to data science concepts. In this paper we will show how mathematical modeling of data science problems can be addressed with students from tenth grade or higher using digital learning material on predictive text systems.

Keywords: mathematics education, Jupyter Notebooks, natural language processing, mathematical modeling, data science.

Motivation and classification of the research area

The use of mobile devices has increased enormously in recent years. This results in a high demand for fast and reliable input methods. One of these assistance systems for typing is word prediction. It makes suggestions for the next word based on expressions already written to save the time of the user when typing. But how does this assistance system know what the user wants to write next? How can word suggestions be generated in such a way that they suggest the desired word with high probability? Answering these questions is the goal of the interactive workshop.

Word predictions are predictive text systems, which are developed within the scientific field of natural language processing. The main goal of natural language processing is to study, how humans understand and use language in order to develop software programs, that allow computers to simulate human behavior (Chowdhury, 2005, p. 51). Predictive text systems also attempt to mimic the language of the user. Specifically, they involve the prediction of letters, words, or even entire word sequences. However, predictive text systems are not only used in typing, but also in speech recognition, spell correction, machine translation, and handwriting recognition (Ghayoomi & Momtazi, 2009, p. 5233). Predictive text systems are thus "one of the important tasks in most natural language processing applications" (Ghayoomi & Momtazi, 2009, p. 5233) and are encountered by students in their everyday life. This makes it an authentic and relevant topic.

In order to give predictions for a text, predictive text systems work with large text data to extract knowledge from it. The N-gram concept, which is a simple decision tree model, is used in the

workshop to bild suitable word suggestions out of the text data (Bahl et al., 1998, p. 1002). This makes it a typical problem in the field of machine learning and data science.

Since data now play a central role in all areas of our lives, it is necessary to give learners an understanding of data science issues and to train them in the critical use of data (Gould, 2021; Opel et al., 2019). Therefore, based on the mathematical and technical knowledge in the modeling of data science questions, digital learning material for high school students should be developed and tested on real problems. The topic of predictive text systems was seen as a particularly suitable example to convey machine learning and data science concepts in an easily accessible way in school lessons. In a design-based research approach, learning material on this topic should be designed, developed and improved in cycles of implementation and redesign with systematic feedback from students.

Background information on the learning material

In the workshop, high school students work out how occurrence frequencies of word sequences can be estimated using large data sets and how they can be used to generate word suggestions. On several worksheets, they collect first ideas for developing a prediction model and then develop different basic models for generating word suggestions. Finally, they apply these models to a large training data set.

The learning material is suitable for students from tenth grade and higher. The workshop assumes prior knowledge of relative and absolute frequencies, as well as prior knowledge of the concept of functions. In addition, students should understand and be able to calculate probabilities and perform and evaluate simple multistage random experiments. Programming skills are not required. The learning material can be used in a compact one- or two-day workshop or divided into several lessons of a teaching unit, both in presence but also in online classes.

A problem-oriented workshop

The developed learning material uses the example of word prediction to show how real-world problems solved using big data can be prepared for students. It approaches the topic in a problem-oriented manner and introduces mathematical content when it is needed to find and understand the solution of the problem. Thus, the problem is the focus and mathematics is experienced as a tool that helps to understand the world, so that the students do not only learn mathematical methods but can recognize the usefulness of mathematics for everyday life. This strategy is part of all workshops that are created within the CAMMP (Computational and mathematical modeling program) project of the Karlsruhe Institute of Technology[1] and the RWTH Aachen University[2]. The overall goal of the project is to promote competencies in mathematical modeling among learners through a variety of learning opportunities and to highlight the importance of mathematical modeling and simulation science for society. CAMMP aims to achieve this goal by providing education and training for teachers in mathematical modeling and by continuously developing and testing new learning material. All workshops created in the project are implemented as digital learning material.

---

[1] http://www.scc.kit.edu/forschung/CAMMP

[2] https://blog.rwth-aachen.de/cammp/

Digital learning material

Digital learning material in school context is already a much discussed topic due to the advancing technological development of society and is currently gaining even more importance for schools teaching due to the Covid-19 crisis and the distance learning associated with it. For mathematical modeling, digital learning material has a special significance, as it enables the solution of real-world problems with large amounts of data in school context and can be a useful support for learners, especially in complex reality-based problems (Geefrath & Siller, 2018, p. 9-10).

This learning material is therefore implemented in the form of digital worksheets, so-called Jupyter Notebooks (see Figure 1), which can be accessed via a cloud platform hosted by the Karlsruhe Institute of Technology. The material can thus be edited directly in the web browser. The login process to the platform is described at www.cammp.online/english/214.php. The digital working material contains many different building blocks clearly arranged in a single file. Instructions, formulas, illustrations, but also code fields can stand directly next to each other and facilitate the learners work in the workshop. Digital differentiation material, such as staged help as fold-out text or in the form of a link to a separate file, as well as consolidation tasks, make the learning material very suitable for heterogeneous learning groups. Additional differentiation is provided by information sheets that are linked to the worksheets and can be called up by the learners if required. For example, learners who have no experience in programming are supported by information about for-loops or if-statements. The subdivision of the problem into smaller tasks as well as adaptive, automated feedback on the solutions enable the learners to work through the material very independently. A more detailed description of the form of the learning material can be found in Gerhard et al. (in press).

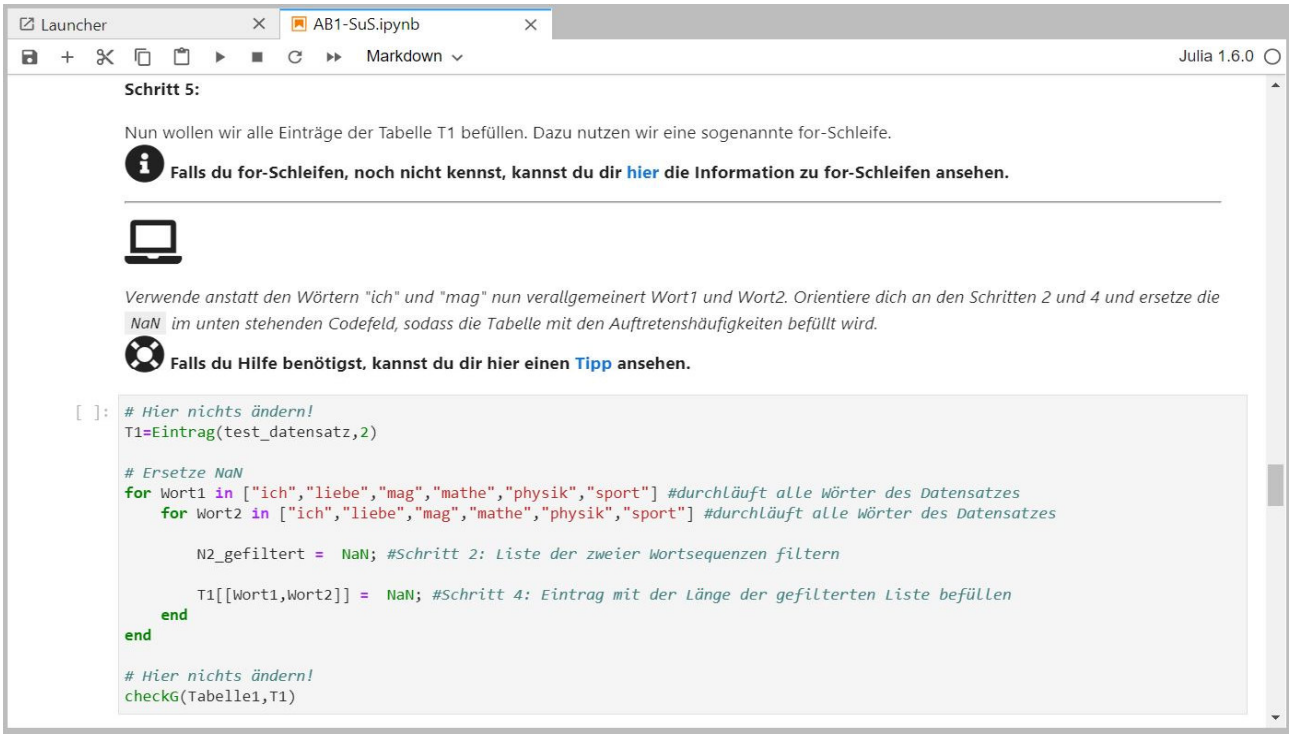


Figure 1: Screenshot of a digital worksheet from the learning material on word predictions

Didactic preliminary considerations

In the development of the learning material, particular attention was paid to the independent implementation of the prediction model by learners. At the end, the solution of the problem should not only be understood theoretically, but also worked out practically. Only minor technical details, which are less relevant for the basic mathematical understanding, are executed in the background.

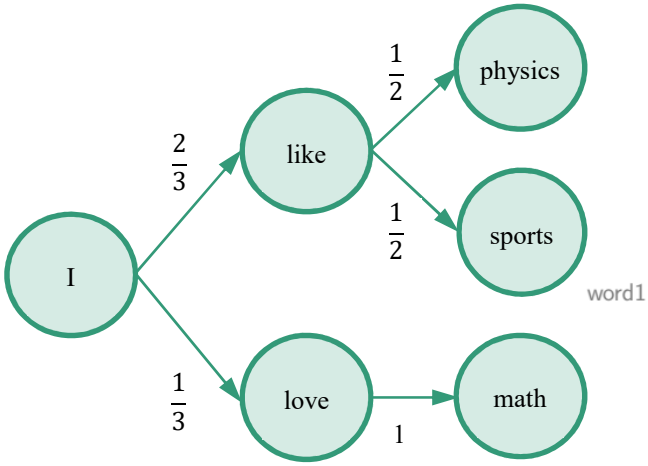
Furthermore, care was taken to avoid unnecessary technical terms, like “Markov chain” or “matrix”. Merely the nth-order Markov assumption is relevant for the workshop. Students will only learn about the applied assumption to the word prediction example. Here it means that the frequency of occurrence of words depends only on the n previous words. In the workshop the assumption thus simplifies the estimation of the occurrence frequency. Also, the term matrix, since it is not known by some learners depending on their previous schooling, is circumvented by paraphrasing it with tables. Since the matrix is only used as a storage form, this can be done without problems in the workshop.

Detailed description of the material

Worksheet 1: A first prediction model

To understand how word predictions are generated, the workshop will first look at a small example. As data basis, which will be used later for comparison, the three sentences "I like physics. I like sports. I love math." will be used. This represents the so-called training data of the prediction model and, in contrast to the training data sets used in practice, which comprise several thousand words, is chosen to be very small in order to be able to clearly illustrate the principle of word prediction. With the help of this data set, the word that the user would most likely write next should be suggested, if possible. It is first assumed that the user has already typed the word "I". Ideas and thoughts about possible suggestions for the next word and how to build them based on the data set, are first collected in a short brainstorming session in plenary. So far, the students always discovered independently that one possible strategy for generating the word suggestions is to compare the already typed word sequence, in our case the word "I", with the data set. In this way, we can determine which word has already been typed in the past after this word sequence and use this information for our prediction.

In our small training data set, the word "I" is found in three places. Now the words which follow can be identified. These are the words "like" and "love". One can therefore assume that these are popular



word1 / word2

	I	love	like	math	physics	sports
I	0	$\frac{1}{3}$	$\frac{2}{3}$	0	0	0
love	0	0	0	1	0	0
like	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
math	0	0	0	0	0	0
physics	0	0	0	0	0	0
sports	0	0	0	0	0	0

Figure 2: Transition graph (left) and transition table (right) of the bi-gram model

subsequent words after the word "I" and thus also represent meaningful suggestions for the next word. The word "like" follows the word "I" more often than the word "love". The probability of the word "like" occurring after the word "I" should thus be estimated as higher according to our training data. Quantitively, this so-called transition probability of the word "like" occurring after the word "I" $P(I \rightarrow \text{like})$ can be estimated via the occurrence frequency of the two-word sequence (I, like) as well as the occurrence frequency of the two-word sequence (I, any word) and results in

$$P(I \rightarrow \text{like}) = \frac{N(I, \text{like})}{N(I, I) + N(I, \text{like}) + N(I, \text{love}) + \cdots} = \frac{2}{0 + 2 + 1 + 0 + 0 + 0} = \frac{2}{3}.$$

These two-word sequences are also named bi-grams, which explains, why the corresponding model is called the bi-gram model. In order to display and store the different transition probabilities in a suitable form, the students learn two different possibilities. The visual representation as a transition graph (see Figure 2, left side) serves mainly as a visual support for the learners, while the transition table (see Figure 2, right side) is especially suitable for a larger data set as a storage location for the transition probabilities. The row of the transition table represents the already typed word (word1), the so-called word history, and the column represents the following word (word2). If a word is now typed, the next word with the highest probability can be found in the corresponding row and indicated as a suggestion.

Based on this, learners develop a general strategy for estimating transition probabilities from a training data set using the occurrence frequencies of two-word sequences and automate the process for all possible transitions so that the calculations no longer need to be done by hand. Now the model can be tested on a larger data set. A larger data set is important so that meaningful suggestions for the next word can be made for as many already typed words as possible. As soon as the already typed word does not occur in the training data set, no suggestion can be made using the bi-gram model. The training data set used for this purpose consists of the German-language texts of the corpus "What's up, Switzerland" (Stark et al., 2014-2020) and the texts of the category "Belletristik" of the corpus "LIMAS" (Research group LIMAS, 1970-1971). The training data set contains more than 300,000 words. From this, part of the data is retained for later testing of the model.

The model is first trained with the help of sample texts, the so-called training data set. Subsequently, the correlations detected from these can be used for prediction on an unknown data set. Bi-Gram models therefore use typical machine learning strategies.

Worksheet 2: Uni-gram and tri-gram model

When testing the bi-gram model with different word histories, learners are tasked with identifying various problems of the model and coming up with possible model improvements. Among other weaknesses of the model, learners recognized that the prediction model uses only the last word for word prediction and that the word history needs to be extended to more than one word for more context-based suggestions. To do this, students use the learning material to develop a tri-gram model that takes the last two written words into account to build a suggestion.

Another problem, which learners will identify through examples in the workshop, occurs when the word already typed does not appear in the data set. In this case, there is an option to suggest the words

that occur most frequently overall in the training data set. This model is called a uni-gram model, because the occurrence frequencies of one-word sequences are determined instead of two-word sequences, as in the original model. The transition probability is calculated by dividing the occurrence frequency of the single word by the total number of words at

$$\mathrm{P}(\text{word 2}) = \frac{\mathrm{N}(\text{word 2})}{\mathrm{N}(\text{every word})}.$$

In the workshop, learners now collect advantages and disadvantages of the three n-gram models, to finally realize that all models have different strengths and weaknesses and that they need to be combined for the best possible prediction of the next word.

Worksheet 3 and 4: Combined models and model evaluation

The combination of the n-gram models makes it possible to realize the probability estimation more reliably and at the same time context-based. For the combination of the models, the back-off procedure or the interpolation are mainly used in natural language processing (Wendemuth et al., 2004, p. 30-31). The back-off procedure is a fall-back strategy. In the case of an unseen bi-gram as a word history, the tri-gram model does not give any suggestions for the next word. Therefore, the bi-gram model is used. If the last word in the word history also does not appear in the training data set, the uni-gram model is used and the words that appear most frequently in the data set are suggested. The students implement the back-off procedure with the help of a simple if statement. As a result, the students learn different combinations of possible n-gram models, while they are improving their basic programming skills.

However, with the back-off procedure, the model may count on the tri-gram probability, which is often not as reliable. This is because there are much more tri-grams than bi-grams or uni-grams. Therefore, many tri-grams do not appear in the training data set, or appear only very rarely. This means that the tri-gram likelihood is often based on very little data compared to the bi- or uni-gram model. It is therefore always best to use all transition probabilities of the individual models and combine them with a weighted sum to produce an overall transition probability. The new estimator of the transition probability of word2 with word history (word0, word1) is thus given by

$$\widetilde{\mathrm{P}}(\text{word 0}, \text{word 1} \rightarrow \text{word 2}) = \mathrm{g}_1 \cdot \mathrm{P}(\text{word 2}) + \mathrm{g}_2 \cdot \mathrm{P}(\text{word 1} \rightarrow \text{word 2})$$
$$+\mathrm{g}_3 \cdot \mathrm{P}(\text{word 0}, \text{word 1} \rightarrow \text{word 2}).$$

The interpolation weights g_1, g_2 and g_3 are initially determined by the learners in a range selected by logical considerations. Later, the weights can be optimized by minimizing an error measure. In this context, the distinction between training data, which are used to estimate the transition probabilities, and test data, which are used to determine the goodness of the language model, but also the weights of the interpolation, is important. Here, the students have the opportunity to independently develop an optimization procedure for minimizing the error measure as a function of the weights or, guided by another worksheet, to learn about one possible procedure.

At the end of the workshop, the initially qualitative considerations regarding the advantages of combining the n-gram models compared to a single n-gram model can also be confirmed quantitatively by calculating an error measure.

Additional tasks

Another optional task for learners is to use the prediction model, they have developed, to generate whole texts. Learners can choose different prediction models (a combined model or a single smoothed n-gram model) and observe differences in text generation. The higher the n-gram length is chosen, the better the generated text will be in both grammatical and contextual terms. At higher orders, the generation is comparable to copying individual sentence strings from the training data set.

Furthermore, the workshop is followed by some social questions about the benefits and dangers of these assistance systems, which can be debated with the learners in an open exchange. Questions such as "Do assistance systems, like word prediction, influence writing behavior?" or "How do biased word suggestions arise and what are their effects?" are particularly suitable for critical discussion. On the one hand, dealing with these questions serves to dive deeper into the topic. On the other hand, it shows a second perspective on the problem from a socio-scientific point of view and, by critically illuminating the assistance systems in plenary, it contributes to empowering the students to form independent critical judgments. The learning material is therefore suitable for application in interdisciplinary lessons, for example in a seminar or a project week on natural and social sciences.

In addition, another worksheet is in progress, which will give students the opportunity to use different data sets as training data for the prediction model. In this way, the influence of the training data on the language model can be highlighted and thus attention can be drawn to the importance of a suitable training data set.

Experience and summary

Data science in school is desirable and possible. This is demonstrated by the learning material described above, but also by other projects such as ProDaBi (Opel et al., 2019), in which a data science curriculum was developed and tested in high school, or the experience of Narges Norouzi et al. (2020) with learning material in machine learning and natural language of the COSMOS Summer School for high school students.

In this workshop, students learn how knowledge can be generated from data and used to support decisions when typing. In doing so, they use typical machine learning strategies. The learning material has already been conducted in an online workshop with 28 learners and will be tested more frequently and continuously improved in the future. The conduction took place as a two-day event during the summer vacations with students from grades ten to thirteen. Additionally, the workshop was conducted and examined from a didactic point of view with 29 teacher trainees of the RWTH Aachen University in one day as part of a didactic seminar. Especially the independent work of the high school students on the open tasks to optimize the interpolation weights and to generate their own text using the prediction model on the second workshop day shows how quickly the students were able to familiarize themselves with a topic that was new to them. After conducting the online workshop with students from grades 10 to 13, the students were asked to fill out an online questionnaire in order to use the results and past experiences to iteratively improve the workshop. The students named machine learning, mathematical modeling, as well as programming basics, among others, as areas in which they could see an increase in learning after the workshop. The workshop was rated with an average grade of two (good).

During the conduct, the learners contributed interesting ideas to the mathematical model, but also to the socio-critical discussion. The high oral participation of the learners shows how interested they are in understanding and solving the problem and the positive feedback from the learners also speaks in favor of the project: “I found the insight into generating word suggestions very interesting and it was a very good example of how mathematical modeling is applied in everyday life.” (Participant’s answer in the evaluation of an online workshop, personal communication, August 27, 2021).

References

Bahl, L. R., Brown, P. F., Souza, P. V. de, & Mercer, R. L. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(7), 1001–1008. https://doi.org/10.1109/29.32278

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Gerhard, M., Hattebuhr, M. Schönbrodt, S. & Wohak, K. (in press). Aufbau und Einsatzmöglichkeiten des Lehr- und Lernmaterials. In M. Frank & C. Roeckerath (Eds.), *Neue Materialien für einen realitätsbezogenen Mathematikunterricht 9*. Springer Spektrum.

Ghayoomi, M., & Momtazi, S. (2009, October 11–14). An overview on the existing language models for prediction systems as writing assistant tools. In *2009 IEEE international conference on systems, man and cybernetics* (pp. 5083–5087). IEEE. https://doi.org/10.1109/ICSMC.2009.5346027

Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics*, *43*(S1). https://doi.org/10.1111/test.12267

Greefrath, G., & Siller, H.-S. (2018). Digitale Werkzeuge, Simulationen und mathematisches Modellieren. In G. Greefrath & H.-S. Siller (Eds.), *Realitätsbezüge im Mathematikunterricht. Digitale Werkzeuge, Simulationen und mathematisches Modellieren* (pp. 3–22). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21940-6_1

Norouzi, N., Chaturvedi, S., & Rutledge, M. (2020). Lessons learned from teaching machine learning and natural language processing to high school students. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), 13397–13403. https://doi.org/10.1609/aaai.v34i09.7063

Opel, S., Schlichtig, M., Schulte, C., Biehler, R., Frischemeier, D., Podworny, S., & Wassong, T. (2019). *Entwicklung und Reflexion einer Unterrichtssequenz zum Maschinellen Lernen als Aspekt von Data Science in der Sekundarstufe II.* https://doi.org/10.18420/infos2019-c14

Research group LIMAS (Ed.). (1970-1971). *Corpus LIMAS (Linguistik und Maschinelle Sprachbearbeitung).* University Bonn, University Regensburg.

Stark, E., Ueberwasser, S., & Göhring, A. (2014-2020). *Corpus "What’s up, Switzerland?"*. www.whatsup-switzerland.ch

Wendemuth, A., Andelic, E., Barth, S., Dobler, S., Katz, M., Krüger, S., Maiwald, M., Mamsch, M., & Schafföner, M. (2004). *Grundlagen der stochastischen Sprachverarbeitung*. Oldenbourg Wissenschaftsverlag GmbH. https://doi.org/10.1524/9783486595000