Predicting Important Topics for the Data Literacy Exam by Analyzing the Word Frequencies

Dennis Grötzinger
Matrikelnummer 6010696
dennis.groetzinger98@gmail.com

Simon Heuschkel
Matrikelnummer 6016180
simon.heuschkel@student.uni-tuebingen.de

Abstract

We use the lecture recordings and slides of the lecture "Data Literacy" at the University of Tübingen 2021/2022 to predict which topics of the lecture will be especially relevant for the exam. To this end, we analyze the word frequencies of relevant topic words in the lectures and use this as a proxy for judging the importance of the topics. Among other things, we find that the topics regarding "gaussian", "bayes" and "pca" will be a focus of the exam. Furthermore, we argue that additional time should be spent on the topics "maximum likelihood estimate" and "log likelihood" as those could be topics that are especially hard to grasp. The results can be found in the repository: https://github.com/Eoli-an/Exam-topic-prediction.

1 Introduction

University classes are aiming to educate their participant and transfer knowledge to an interested audience. In the end of a semester, students mostly need to prove their understanding of its content by passing an assignment, which is often an exam. In order to pass the assignment, the lecturer hands out information in several forms, like lecture recordings and handout slides.

Learning the all given information in order to pass the exam is a hard and tedious task, which takes a lot of time. Therefore, it is advisable to set learning emphasizes in order to maximize the return of investment on the exam grade in terms of study time.

In this paper, we will introduce a method that will help students set such an emphasis. Our method is based on the core assumption, that the lecturer will focus especially on those topics that will be most relevant in the exam, which will be reflected in the learning material. If that is true, the exam topics can be predicted based on the material given by the lecturer.

To approximate and quantify the focus of topics in the learning material, we use the word frequency of the lectures and the slides. To this end, we conduct two analysis in our paper. The first experiment will examine the overall word frequency of relevant topics in the lecture recordings. This will show us which topics are overall the most relevant for the exam. The second experiment will contrast the word frequency of the lectures to the word frequency of the slides. This gives us a further data point for the relevance of topics and allows us to determine topics that the lectures has spent a lot of time explaining, which in turn indicates that this topic is hard to understand.

2 Data Collection

The basis of our analysis is the course "Data Literacy" at the university of Tübingen in the winter terms 2021/22. The lecture was held by various people of the methods of machine learning faculty at the university of Tübingen [2].

As our first resource, we use recordings of lectures. The recordings of lecture one was too big to transcribe with our tools, and lecture thirteen was not yet available at the time of our analysis. This lead to an exclusion of these lectures from the analysis. To analyse the word frequency, we needed to convert the lecture recordings to written text. In our pre-registration we stated that we will use the wave2vec2 model of huggingface for transcribes. However, even the largest available model gave very poor results that were unusable for our analysis. Instead, we used the cloud-servive provider Otter.ai [4] to generate the transcribes. Our final result were eleven lecture transcripts in .txt format as data for our analysis.

Our second data source are handout slides. These were used by the lecturer during the lecture presentations and afterwards handed out to the students. We extracted the textual data of the pdfs with the python port of apache tika [1].

3 Data Pre-processing

In a first attempt to calculate the word frequency of topics in the lecture material, we tried to automatically extract the most frequent terms from the transcribes and the slides. However, we soon realized that automatic methods were not getting us the desired results, as many of them were filler or irrelevant words such as "basically" or "Yeah". [1]

Instead, we hand-crafted a set of relevant words that correspond to topics in the lecture. We then searched the material for occurrences in the learning material. In order to account for different expressions of the same topic, we often used the word stem of a topic in our analysis. For example, we used "estimat" to both capture the word "estimator" and "estimation" which both express the same topic.

The result is a collection of the relevant words in combination with their frequencies in both transcribes and lecture slides.

4 Experiments

We conduct two experiments that analyze the most frequent and therefore most important topics of the lecture.

4.1 Overall Word Frequency

It is often the case that if a topic is important, one will mention it more often. This is our core assumption for this experiment. We use the relevant word frequencies in the all of the transcribes as defined in Section 3. The results of this analysis can be seen in Figure 1. We can observe that the most used word by far is "gaussian" with almost double the occurrences compared to the second most frequent word. Other frequent words include "pca", "bayes", "linear regression" or "estimation" (we used the word stem "estimat" here for our analysis). If our assumptions are correct, those will be the most relevant topics in the exam. In contrast, topics that were not mentioned very often and might therefore not be very important in the exam include "regularization", "linear discriminant analysis", "hypothesis test" or "entropy".

4.2 Word Usage in Slides vs. Transcribes

In this experiment, we explore the relationship of rlevant word frequency of the slides in contrast to the word frequency of the transcribes. This allows us to draw more nuanced conclusions. For example, if a topic is mentioned a lot in both slides and transcribes, it is probably quite relevant for the lecture and vice versa. If a topic gets mentioned a lot in the transcribes but less so in the slides, this is an indication that the lecturer had to spend a lot of time explaining those topics. Therefore, this might be a rather difficult topic and worth spending some extra time for in the exam preparation. In contrast, if a topic gets mentioned a lot in the slides but less so in the transcribes, the topics might be easier to explain and therefore less time has to be spent on them during the exam preparation.

[1] Fun trivia: The most used word in the lecture recordings was "Okay", which was said a total of 434 times.

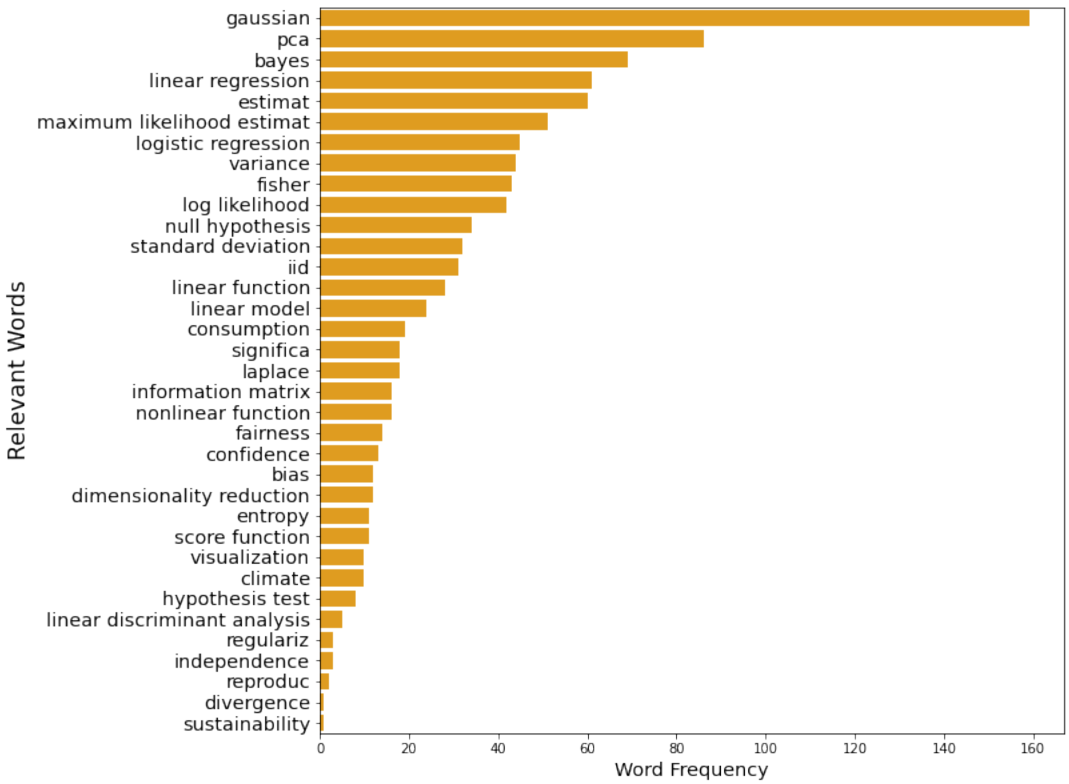

Figure 1: The frequency of relevant topics over all lecture transcribes.

To highlight the relationship of the word frequencies, we opted for a ranked representation in a scatterplot, inspired by [3]. The plot can be seen in Figure 2. The x-axis shows how frequent a word is in the transcribes and the y-axis how frequent it is in the slides. However, instead of using absolute occurrences as the frequency like in 4.1 here we only use the ranking of the frequencies. This means that the least frequent word gets assigned to be rank 1 the second least frequent word gets rank 2 and so on. This procedure discards the relative frequency distance of two words. It enables us to create a dense representation of the data, where it is only important which rank a certain word has. This leads to an evenly distributed plot. We can observe that the topics "gaussian", "estimation" or "pca" are frequent in both the slides and the lectures, indicating a high importance for the exam. Furthermore, we see that topics like "confidence", "sustainability" or "fairness" are topic that are frequent in the slides but not very frequent in the transcribes. This is an indication that those topics are not as hard to explain and therefore do not need as much time in the exam preparation. In contrast, topics like "maximum likelihood estimation", "standard deviation" or "log likelihood" are topics that the lecturer has spent a lot of time explaining. Those might be rather hard topics that the lectures wanted to especially emphasize, in turn warranting additional preparation time for the exam.

5 Summary

In this paper, we have used word frequency analyses to predict, which topics will be relevant for the Data Literary exam. In the first experiment, we have explored the most frequently used relevant words in the transcribes and argue that this is a good proxy for exam relevance. In the second experiment, we have analyzed the relationship of the word frequencies in the slides versus the word frequencies in the transcribes. This enables us to predict which topics will be relevant or are harder to grasp, warranting additional preparation time.

Most notably, we advise taking special emphasis on thoroughly learning the topics "gaussian", "estimation", "pca" and topic related to "bayes", as those were the topic that were the ones that come up most in the slides and the transcribes. We also point out that it could be worthwhile to have a good

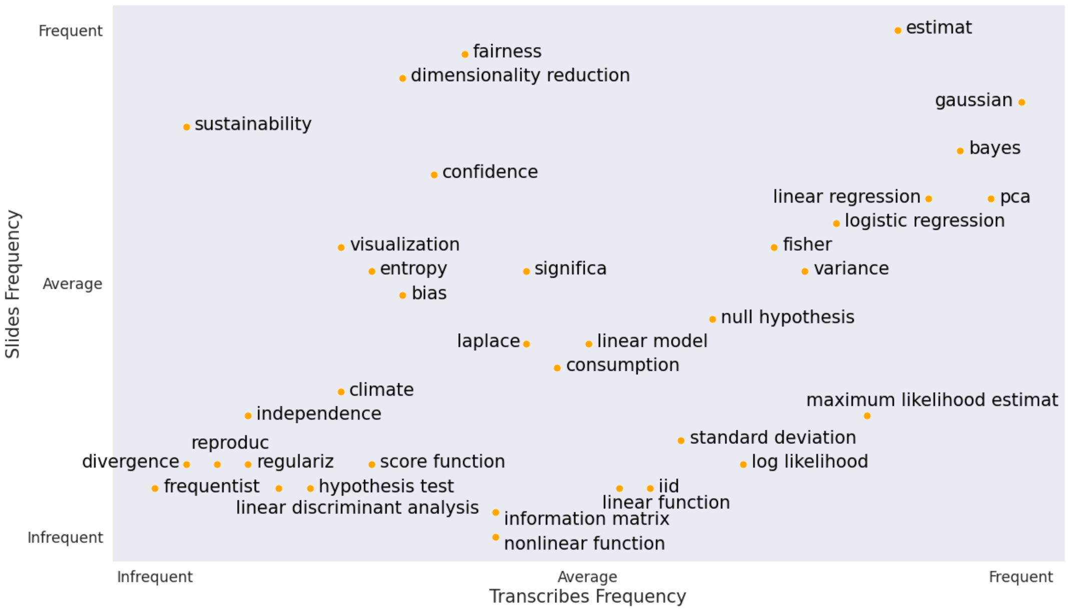

Figure 2: The frequency of relevant words in the slides vs. in the transcribes. We only use the rank of the words in terms of frequency, any other relative information is discarded.

look at the topics "maximum likelihood estimation" and "log likelihood" as those could be topic that are difficult to understand deeply.

However, our conclusions and predictions are based on a number of assumptions, which introduce limitations to our work. In the next and last section, we will outline those limitations and provide an outlook on how to overcome them in further work.

6 Limitations and Further Work

First, our whole analysis bases on the assumption that the word frequency of a topic is a relevant proxy for the relevance of that topic. While this certainly appeals to our intuitions, we found no concrete evidence that this the case for our analysis. However, one obvious way to put this to the test would be actually check whether our predicted topics are also the ones that come up in the exam. A further limitation is the fact that the lectures were held by different instructors. The same is true for some slides. This could potentially skew the data. Further work could explore if there is substantial difference between the lectures and correct for them. Another limitation in our conclusion is the variety of terms classified as relevant. Some terms are distribution, others are explicit algorithms or mathematical terms. This makes comparisons hard.

In conclusion, we believe that we have provided a good proxy to divide up the limited time for the exam preparation on the different topics of the lecture. Time will tell if our predictions were reasonable.

References

[1] *Apache Tika*. 6.02.2022. URL: https://tika.apache.org/.

[2] *Faculty for Methods of Machine Learning, University of Tübingen*. 6.02.2022. URL: https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/methoden-des-maschinellen-lernens/.

[3] Jason S. Kessler. “Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ”. In: (2017).

[4] *Otter Voice Meeting Notes*. 5.02.2022. URL: https://otter.ai/.