Predicting the focus of the Data Literacy exam by analysing the word frequency of lectures

Dennis Grötzinger
Matrikelnummer 6010696
dennis.groetzinger98@gmail.com

Jakob Macke
Matrikelnummer 987654321
dummymail@uni-tuebingen.de

Abstract

We do nice stuff and get nice results

1 Introduction

1. We want to use statistics of the transcribes of the lectures as a proxy for determining which topics will be important for the exam (or which ones will be difficult)
2. Analysis is severely limited by certain assumptions, i.e. that word frequency can be used as a proxy for importance and also that the importance that prof henning has is also important for the phd students that design the exam

2 Data collection

1. huggingface didnt work so we used otter.ai
2. otter. ai is a cloud service similar to google cloud services that has speech to text as a service

3 Overall word frequency

1. here we analyze which words are most common overall, as a proxy for importance
2. we only use nouns that we think are sensible and not filler words

4 Word frequency per lecture

1. most frequent words in every lecture (top 50 and then only the sensible ones)
2. we can therefore see the most important topics by lecture and the data does not get skewed when one lecture has a lot of topics

5 cross importance

1. the most frequent words when we exclude the lecture it was introduced in (so the lecture where it is mentioned most often)
2. this will show us which topics are very important over the whole semester and not just because it was mentioned particularly often in the introductory lecture

6 Complexity/entropy of a lecture

1. use various scores from textstat to estimate the complexity of the lectures. This might show us which lectures will be especially difficult and should be a focus when studying

Project Report for *Data Literacy* 2021/22