Improving Student Performance with N-gram Text Analysis

Team Members: Anushikha Joshi, Mudit Chandna, Khushil Nagda, and Shivraj Sambhus

Abstract

Text analysis involves deriving meaningful insights from text data by using methods such as n-gram analysis in a bag-of-words model. Using this model, we constructed a list of common words from the end of chapter questions found in the CourseKata textbook 'Advanced Statistics and Data Science I: A Modeling Approach (High School / ABC)' that each student performed poorly on. We believe that this list may improve student performance by presenting them with the concepts that they may not understand well enough. The students can then refer to these words in upcoming problems to further improve their comprehension of the material in each chapter.

Methods

We began by implementing n-gram analysis in a bag-of-words model to examine the high frequencies of words and pairs of words in the end of chapter questions for each student in the $\text{`checkpoints}_e oc.csv' file. The bag-of-words model shortens the questions to sets of words so that we can pick the$
$grams' and \text{`}2-grams'. An example of a high frequency 2-gram is \text{`}random sample' and has 261 total occurrence$

Results

We believe that students may benefit from reviewing the course chapters containing the aforementioned personalised list of words because they either did not engage enough with the chapter when completing the end of chapter questions or that they did not have enough study material in the chapter to study the specific concepts associated with these words. To visually represent the former case, we generated a heatmap that measures how much the students engaged with the end of chapter questions in terms of how poorly they scored on each set of questions. The heatmap allows us to infer that students tend to score a lower grade on their end of chapter questions when they do not engage with the questions for a sufficient amount of time.

Discussion

One possible implementation of our results is to create a summary page which follows each chapter and displays the students' personalised list of words once they have completed the end of chapter questions. This option can be enabled or disabled for any course that is coordinated on the CourseKata learning management system.

One issue that we ran into was that we could not repeat our experiment on exam data rather than end of chapter question assessment data since those datasets were inaccessible to us.

Conclusion

N-gram text analysis in a bag-of-words model allows us to understand students' performance on a given end of chapter assessment. Students typically score higher on end of chapter questions when they are engaged with the material and have practised it thoroughly. If students do not perform well on these assessments, we conclude that they did not engage with the chapter for long enough or did not have a sufficient amount of material in the chapter to study from.