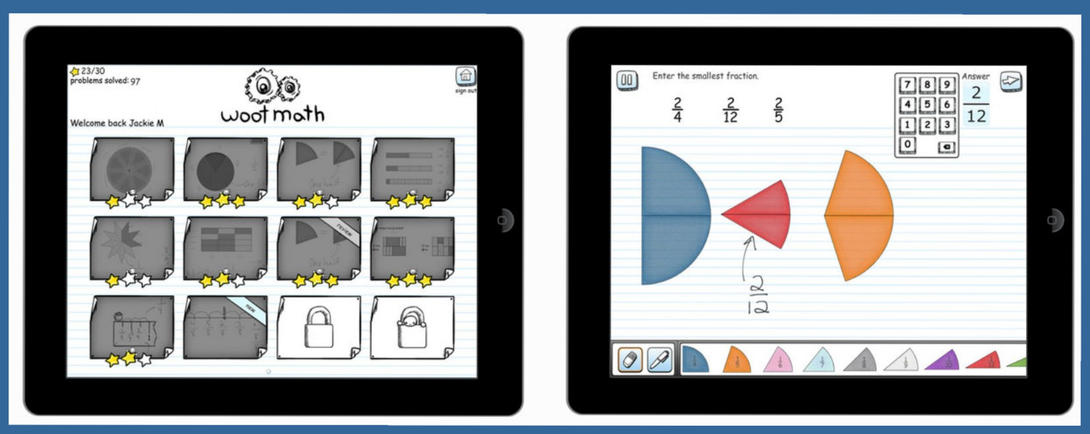
Woot Math Project Collaboration for CSCI 5622 – Machine Learning Unsupervised Learning Task

**Shane Lockwood, Brian Lubars, Brian McKean,
Carl Mueller, Geoffrey Sutcliffe**

Introduction

Woot Math is a Boulder-based education software startup focusing on helping teachers customize learning for students struggling with core math concepts in grades 3-8. The project goals were to identify behavioral traits (patterns of mistakes and misunderstandings) in student submissions and/or identify relationships between traits.

FIGURE 1



Left: Woot Math software's lesson choice menu emulating Angry Bird leveling.
Right: Sample problem displaying software elements/tools and submission areas.

Feature Engineering

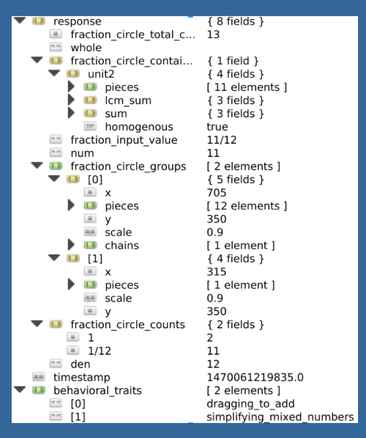
Woot Math supplied a 20gb .bson file containing 10+ million student submissions.

Feature Engineering approaches:

1) Generating strings from response object data combined with the problem's text to create bag-of-words features for each submission.

2) Directly using response key-value pairs as categorical or numerical features for each submission to be used in two different clustering models.

FIGURE 2



Response subfield of student submission.

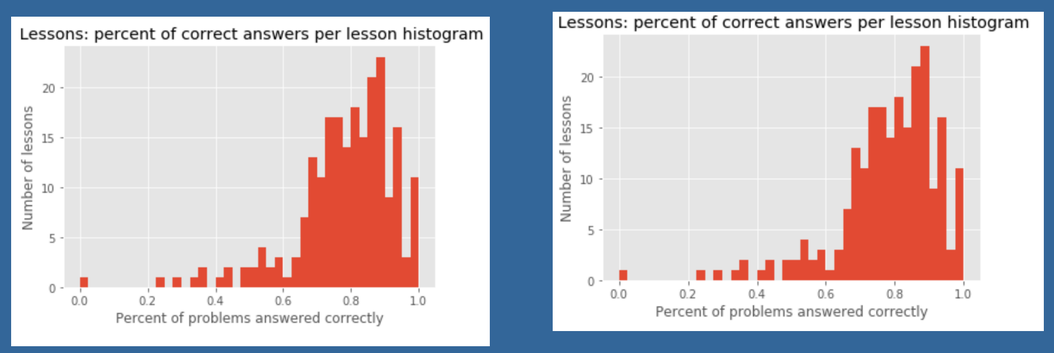
Key-value pairs represent the software components of tools used by students at the time of submission.

This was the predominant piece of data used to generate features.

Exploratory Data Analysis (EDA)

EDA was performed on the entire data set to gain general insight and to analyze clusters.

FIGURE 3



Examples of the types of EDA techniques used for understanding the data set and recognizing important features to be used in our analyses of our clusterings.

Unsupervised Clustering Approaches

Approach 1: Cluster across all lesson types

- Bag-of-words feature vectors
- Term Frequency – Inverse Document Frequency filtering
- K-means clustering using Scikit-Learn

Approach 2: Cluster within lesson types

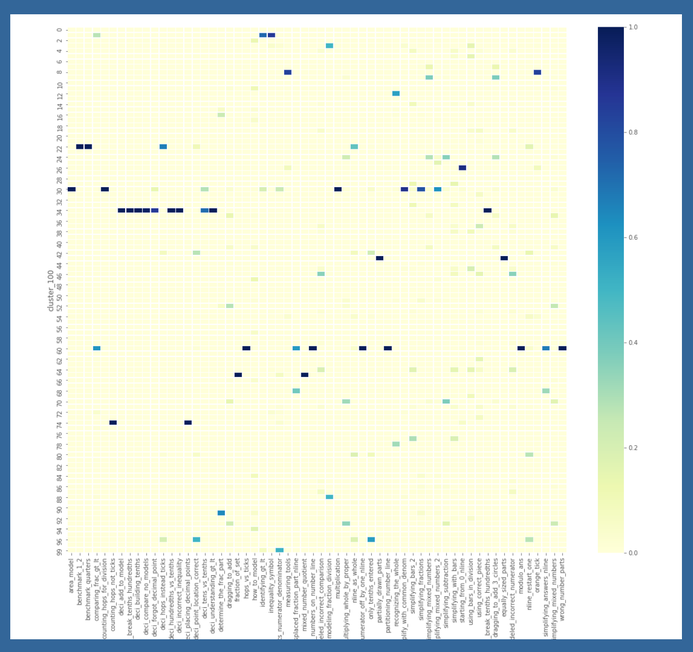
- Partition by lesson/problem type using 'qual_id'
- Kmodes Discrete Clustering
- Kmeans Numerical Clustering

Results

Approach 1:

First approach clustered submissions/documents directly, with no partitioning.

FIGURE 4



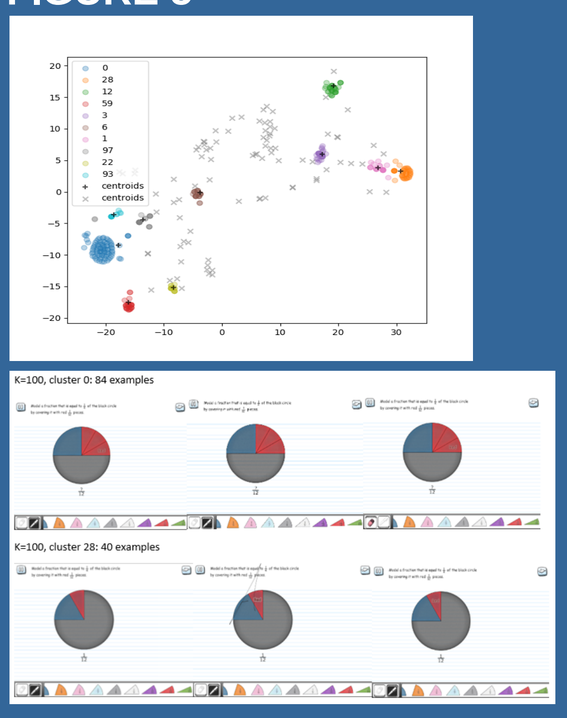
Heat map showing the relative density of a given behavioral trait (x-axis) within a given cluster (y-axis).

- Clusters were generated with high K values of 100.
- Some clusters entirely contained all responses for a given behavioral trait.
- Other clusters contained only a few traits, or none at all.
- Potential relationships between behavioral traits in dense clusters and potentially new behavioral traits in sparse clusters warrants further investigation.

Approach 2:

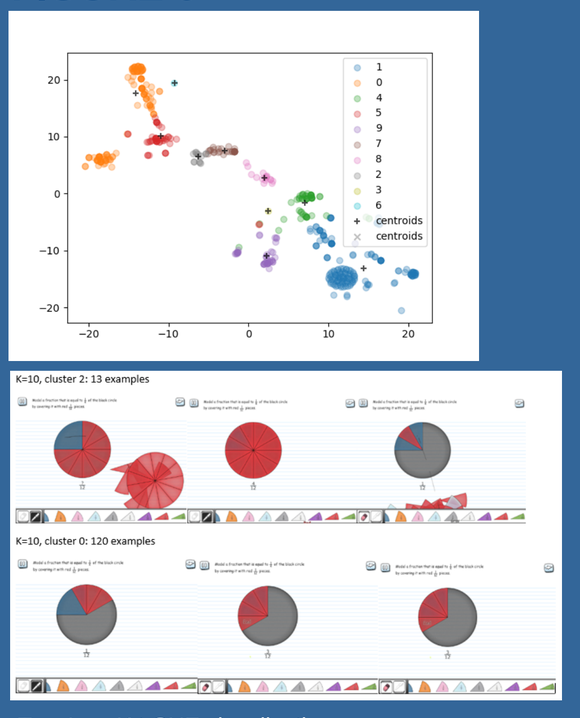
The second approach clustered student submissions/documents conditioned on a data value called the qual_id which represents problem types. The two clustering models, kmodes and kmeans, produced similar results. This approach was taken to narrow the scope of the clustering in an attempt to find more granular differences in our clusters.

FIGURE 5



Top: High K t-SNE visualization
Bottom: High K submission images

FIGURE 6



Top: Low K t-SNE visualization
Bottom: Low K submission images

- High K values produced many sparse clusters and a few highly dense clusters. Within the dense clusters, there was much greater homogeneity (Figure 5).
- Low K values for both models produced more uniformly populated clusters but with little intracluster homogeneity (Figure 6).

Challenges and Conclusions

This project presented a unique opportunity to provide meaningful insight into Woot Math's software platform. The clustering results do not appear random, but deeper analysis within clusters is needed to evaluate whether the clusters meaningfully represent important and actionable features within the data set.

The data provided, while abundant and clearly labeled, proved difficult to understand in terms of its importance when engineering features. However, we obtained a decent baseline with which Woot may further explore their data using unsupervised learning techniques.