Incorporating Real-Time Clustering of Student Responses into an E-Learning System

Philipp Burckhardt, Christopher Genovese, Rebecca Nugent, Ronald J. Yurko

ISLE

Carnegie Mellon University
Statistics & Data Science

Background

The integrated statistics learning environment (ISLE) is an e-learning platform that we are developing at CMU to

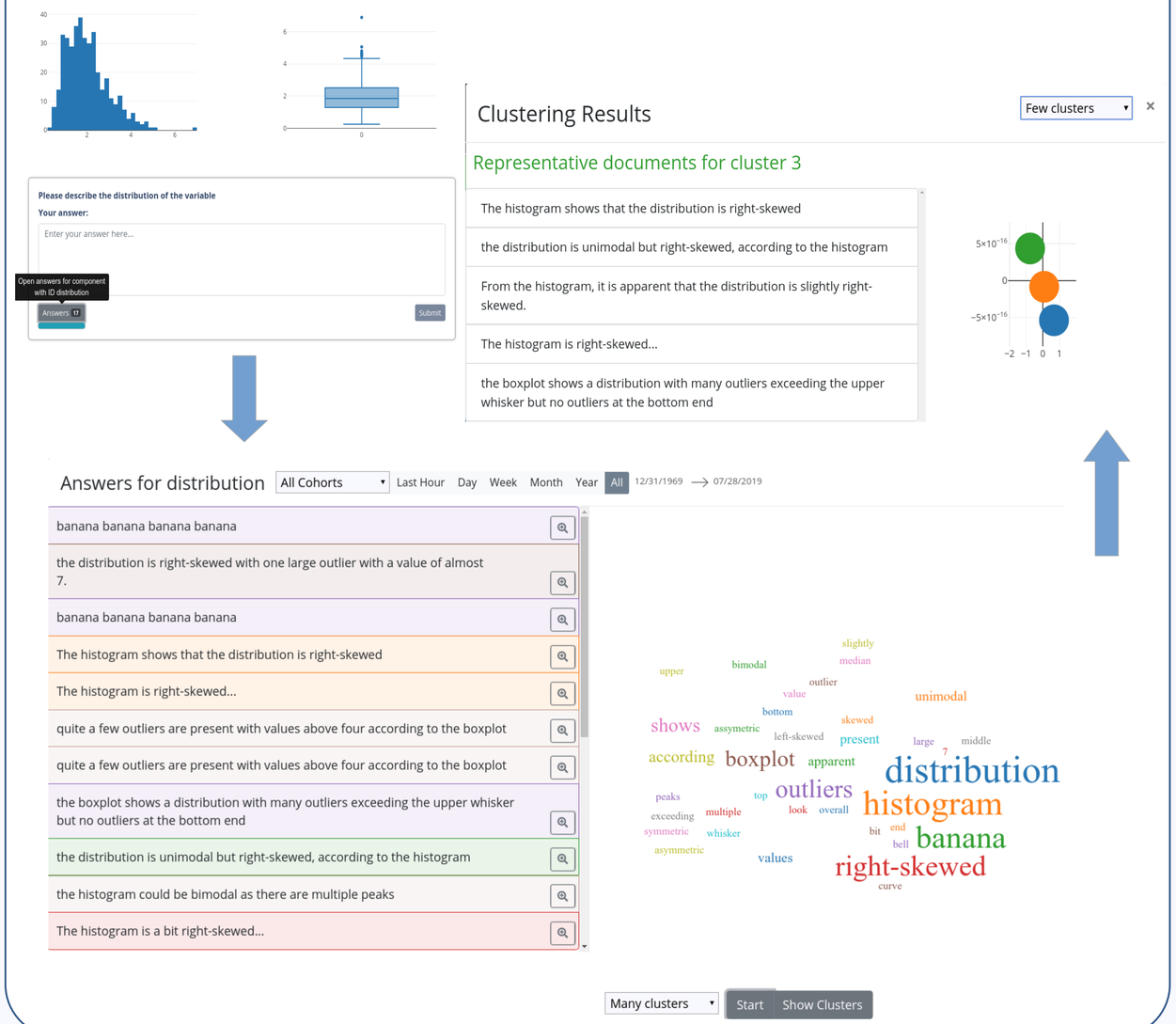
- let teachers integrate interactive learning modules into their courses (lectures, labs, etc.)
- allow monitoring and nudging of student behavior
- permit real-time evaluation of student progress
- enable students to interactively explore statistical concepts
- allow instructors to easily reuse and remix existing material

Specific Challenge

To promote active learning, students in the lab sessions for our introductory statistics class will often submit free-text answers, which are then discussed in class.

How can we give instructors an overview of the different types of answers submitted and help them select questions to discuss?

Solution: Real-Time Text Clustering



Methods

Pre-Processing:
Turn to lowercase, expand contractions, remove punctuation, remove stopwords, calculate "bag of n-grams" as features for our model.

Model:
Spherical k-means / k-means with cosine similarity.
Sequential version:

- Make initial guesses for centroid locations $m_1, m_2, \ldots, m_k$
- Set counts $n_1, n_2, \ldots, n_k$ to zero
- Until interrupted
  - Acquire next observation x
  - If m_i is closest to x,
    - Increment n_i and replace m_i with $m_i + (1/n_i)(x - m_i)$

Feature Hashing:
Vocabulary not known in advance. Use of "hashing trick": Words are directly mapped to indices by applying a hash function and then restricting the resulting hash value to the range [0, k-1] using the modulo function.