CrowdAnalytics

Redundancy Clustering in Crowdsourced Data Analysis

Input Data
We asked workers to generate explanations for 12 charts drawn from 3 datasets covering a range of public-interest data types. We showed each chart to 10 different workers, for a total of 120 analysis microtasks. 93 workers participated, producing a corpus of 156 explanations.

Goal
Help the analyst identify plausible hypotheses by clustering together redundant explanations.

Methodology
We cluster the explanations using three methods:
• K-means clustering of various bag-of-words representations of the explanations.
• K-means clustering of representations consisting of human generated pairwise similarity scores.
• Manual clustering by Mechanical Turk workers.

Results
We evaluate the quality of a clustering by comparing it to three experts’ clusterings using the F-measure similarity metric.

Conclusions and Future Work
Human workers do better than automatic methods in redundancy reduction at small scale. This may be reversed when dealing with large datasets, an inquiry we leave for future work.

In addition to identifying plausible explanations, an analyst may wish to determine whether hypotheses are trustworthy and cite a reputable source. We are currently investigating various ways to identify and surface this information.