Identifying and Characterizing Near-Duplicates in Big Clinical Note Datasets

Authors: Rodney A Gabriel; Sanjeev Shenoy; Tsung-Ting Kuo; Chun-Nan Hsu, University of California San Diego

Abstract: The widespread presence of near-duplication in clinical data is problematic when training predictive algorithms that model the language or attributes in notes. For example, duplicates may cause a predictive model to erroneously identify correlations between symptoms if these are repeated many times. We developed scalable algorithms to characterize sources of near-duplication in clinical notes as: exact copies, common output notes (e.g., device output), or general templates.

We developed a method to identify clusters of highly similar notes – it uses an approximation algorithm to minimize pairwise comparisons and consists of three phases: 1) Minhashing (first used in AltaVista search engine to detect duplicate webpages from the entire World Wide Web) with Locality Sensitive Hashing; 2) a clustering method using tree-structured disjoint sets; and 3) classification of near-duplicates via pairwise comparison of notes in each cluster. The algorithm can be used to analyze large clinical note corpora with finite available memory space. Results demonstrated that from 10,902,795 notes from UCSD, there were a total of 14,539 clusters, in which 6,945 notes were exact copies, 11,509 were common output notes, and 44,218 were general templates. There were no false positive clusters when clusters were created for notes with Jaccard similarity of 100%.