MATCHLESS: A Linear Algebraic Approach to Duplicate File Identification

Litt, Michael

Many algorithms currently exist to identify duplicate files, using techniques such as hashing, fuzzy matching, size comparison, and character by character comparison. The purpose of this project is to demonstrate a novel method of file comparison using singular value decomposition (SVD), a linear algebra concept involving matrices. The technique is applicable to any file that can be represented as an array of values, such as an image or spreadsheet. The approach takes advantage of the orthogonal property of SVD to check for differences between files using vector-matrix multiplication. The method is demonstrated on sample images and other types of files. In addition, other applications of this method are explored. This novel method succeeds in identifying duplicate files and does so elegantly and efficiently. This algorithm can be used to reduce clutter in smartphone and computer directories, and can also be implemented in apps to detect incoming duplicate files, among other applications.

Awards Won:
Mu Alpha Theta, National High School and Two-Year College Mathematics Honor Society: First Award of $ 2,500
National Security Agency Research Directorate : Honorable Mention "Science of Security"