# STRATIFYING REPOSITORIES TO DETECT SUPERFLUITY

D.Ragu Nandhan , S.Raghav, Dr. Sheryl Oliver. MTech., Ph.D ,
Department of Computer Science
St.Joseph's College of Engineering Chennai, India

The amount of data we produce every day is truly astonishing. There are 2.5 quintillion bytes of data created each day, but that pace is only accelerating even more. These data being produced are not always redundant free. In case of redundancy, they have to be eliminated to ensure a good and fast working environment. Technologies which help us utilize the storage devices better have been favored ever since the birth of the Internet. It helps us to effectively manage the limited storages we have on our mobile phones and personal computers. Data deduplication is one such technology that helps us utilize storage devices and network utilization better. Data deduplication plays a vital role today in every domain, making the system work better and faster. Data deduplication takes up many forms. Virtual tape libraries, archive storage, disk storage systems, and applications such as email systems, content managers, backup systems and more, are examples of where data deduplication can be applied. It ultimately finds out the redundant files which are not necessary and provides an intimation to the user, so appropriate actions can be taken by the user. This project detects the duplicate text files and images by generating a hash code for the files and comparing it with the hash codes of other files to detect if there is a duplicate. The user can decide what to do with those duplicate files later. The directory is specified and a recursive file search is used to detect all the file duplicates that exist. Text files are said to be redundant when their contents are exactly the same. In case of images, the process of image manipulation is deployed. Image manipulation is the technique by which an image is transformed so as to convert it to user desired form and in this project, the images are reshaped and it is then converted to grey scale so as to augment the accuracy. Images are said to be redundant when all the pixel values of the compared images match and they are said to be similar when their properties like hue, saturation are altered. Similar images are also found with the help of the techniques like D-Hashing. Sometimes, text files can be very similar and they should also be considered as duplicates at certain places. By using a technique called rolling hash, it is quite easy to find how similar two files are.

 ISBN 978-93-81288-18-4