# CALIFORNIA STATE SCIENCE FAIR
# 2015 PROJECT SUMMARY

**Name(s)**

Sanjana V. Shah

**Project Number**

**J1418**

**Project Title**

## Text Compression Algorithm Based on Popularity Using Global English Dictionary

**Abstract**

**Objectives/Goals**

The objective of this project is to invent a new algorithm that encodes texts found in documents, books, and novels, using popularity of words selected from English dictionary. The goal of this project is to demonstrate that the new algorithm can create a compressed file, whose size is smaller than the one produced by standard compression software such as gzip and bzip2.

**Methods/Materials**

Compression ratios highly depend on extent to which the redundancies are detected and the ways in which they are encoded in the output. Since traditional schemes such as Huffman encoding and data dictionaries use the frequency of words that are found only within the document, storing these dictionary values in the output file increases the compression overhead. In contrast, this project uses a word list of 86,600 most popular English words, and encodes every word in the document with its popularity value. Higher the popularity, smaller is the encoded value. These popular words are stored in a global dictionary, accessible by any document that needs to be encoded or decoded. Since this dictionary resides outside the final document, it not only reduces the file size significantly, but also allows the dictionary to be more adaptable to adding/removing/rearranging popular words over time. The version number of the dictionary is stored in the header of the encoded file. The encoded file is then compressed using gzip and bzip2. When decoding, the output file is compared byte by byte to the original file to guarantee that there was no data loss.

**Results**

This algorithm reduces file sizes of books and novels by 20-30% during the encoding phase. When compressed, it further reduces the file size by 8% over the file size produced by gzip. Statistics show that about 75% of the words got encoded with popularity codes. With specialized domain specific dictionaries, the compressed file sizes of medical transcripts got reduced by 15% over gzip.

**Conclusions/Discussion**

The major contribution of this work is the use of popularity word dictionary and its adaptability to rapidly changing popularity of English words. In addition to its use in improving the compression ratios, this algorithm allows storing the dictionary globally, thereby sharable by all users for any type of documents. The novel approach in this project is to combine the two areas - the need to compress a file and the availability of popular words in English literature.

**Summary Statement**

The purpose of this project was to develop a new text compression algorithm that uses popularity word list, and to demonstrate that the compression ratios are better than standard gzip software.

**Help Received**

My math teacher, Mr. Chang, guided me during the project. Local city library helped me learn Java.