Tecnológico Nacional de México
Instituto Tecnológico de Tijuana

Ing. en Informática

Materia: Calidad de los sistemas de la información

Unidad #1
Tema: Pair Coding

No.Control:
18210703

Alumnos:
Baltazar Moreno Brayan

It is a simple form of data compression that replaces the most common pair of consecutive data bytes with bytes that do not appear in the data. A replacement table is needed to reconstruct the original data. This algorithm was first publicly described by Philip Gage in the February 1994 C Users Journal article "A New Algorithm for Data Compression".

A variant of this technique has been found useful in a variety of natural language processing applications, including: B OpenAI GPT, GPT2, GPT3.

Ejemplo:

Suppose the data to be encoded is

aaabdaaabac
The byte pair "aa" occurs most frequently, so it will be replaced by a byte that is not used in the data, "Z". Now there is the following data table and replacement:

Zabd Zabac Z = aa
The process is then repeated with the "ab" byte pair, replacing it with Y:

ZYdZYacY = abZ = aa
The only remaining literal byte pair occurs only once, and the encoding might stop here. Or the process could continue with recursive byte-pair encoding, replacing "ZY" with "X":

XdXacX = ZYY = abZ = aa
This data cannot be further compressed using byte pair encoding because no byte pairs occur more than once.

To uncompress the data, simply do the replacements in reverse order.