The Lempel–Ziv–Welch (LZW) Algorithm

Tom Magerlein

May 10, 2017

Asked during Ben's presentation on Huffman coding:

Can we replace common sequences of characters, like "the", with single codes?

Asked during Ben's presentation on Huffman coding:

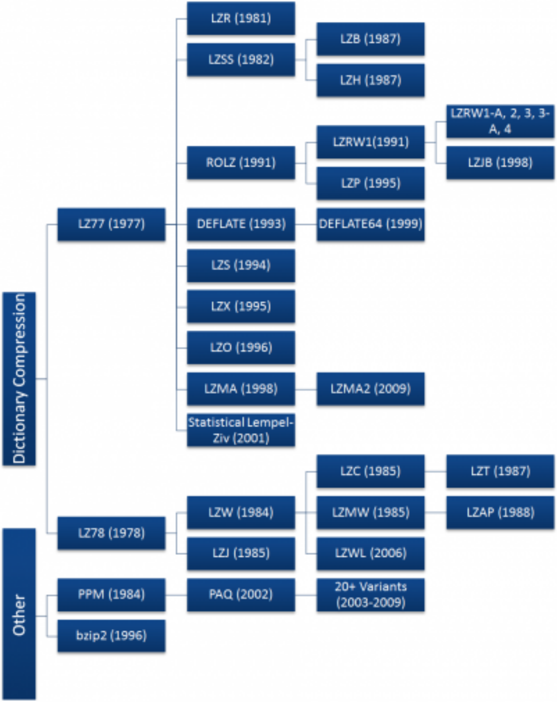
Can we replace common sequences of characters, like "the", with single codes?

Yes!

Origins

- 1977-78: Lempel and Ziv introduce LZ77 and LZ78
  - LZ77: Replace previously seen sequences of characters with references to previous appearances
  - LZ78: Instead of referencing earlier parts of file directly, builds a dictionary of previously-seen symbol sequences
- 1983: Sperry Corp. (later Unisys) files patent on original LZW implementation
- 1984: Welch publishes "A Technique for High-Performance Data Compression", describing the LZW algorithm

Origins



Dictionary Compression

LZ77 (1977)

LZR (1981)

LZSS (1982)

LZB (1987)

LZH (1987)

ROLZ (1991)

LZRW1(1991)

LZRW1-A, 2, 3, 3-A, 4

LZJB (1998)

LZP (1995)

DEFLATE (1993)

DEFLATE64 (1999)

LZS (1994)

LZX (1995)

LZO (1996)

LZMA (1998)

LZMA2 (2009)

Statistical Lempel-Ziv (2001)

LZ78 (1978)

LZW (1984)

LZC (1985)

LZT (1987)

LZMW (1985)

LZAP (1988)

LZWL (2006)

LZJ (1985)

Other

PPM (1984)

PAQ (2002)

20+ Variants (2003-2009)

bzip2 (1996)

Uses and Patent Troubles

- Saw use in some compression utilities, but most notable use was in CompuServe's GIF image format, introduced in 1987
- In 1993/4, Unisys discovers use of LZW in GIF format, attempts to claim licensing fees from software that handles GIF images
  - Leads to development of the patent-unencumbered PNG format and the widespread use of the DEFLATE compression algorithm, as well as use of the GIF format without compression
- Patent expired in 2003, but still not widely used except in GIF

LZ77: Overview

- Replaces previously seen data segments with a reference to where they last occurred, as a pair indicating offset and sequence length
- Compressor keeps an output history (typically a few kilobytes), called the "sliding window", and a lookahead buffer
- Algorithm
  - Find longest prefix of data in lookahead buffer which occurs in sliding window
  - If such a prefix exists, and it would save space to do so, output reference to its last occurrence; otherwise output first unit of data as a literal

LZ77: A Short Example

A	B	C	B	A	B	C	B	C	B	C

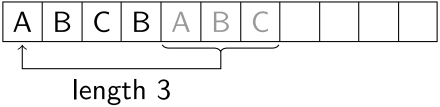
LZ77: A Short Example

A	B	C	B	A	B	C	B	C	B	C

A	B	C	B							

LZ77: A Short Example

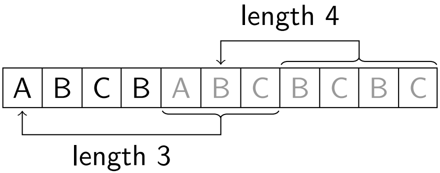
A	B	C	B	A	B	C	B	C	B	C



A	B	C	B	A	B	C				

LZ77: A Short Example

A	B	C	B	A	B	C	B	C	B	C



LZ78: Overview

- Replaces LZ77 sliding window with a dictionary, and backreferences with codes representing entries in the dictionary
- Compressor, decompressor agree on rules to build dictionary, so it does not need to be stored with compressed data
- Algorithm:
  - Find longest prefix of lookahead buffer in current dictionary
  - Output code for that prefix
  - Output code for first character after prefix
  - Add prefix followed by next character to dictionary, if dictionary is not full

LZ78: Example

A B C B C B A A B C A B C B B B B B B

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		

LZ78: Example

A B C B C B A A B C A B C B B B B B

[A] [B] [C] [B] [CB] [A]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B]

Dictionary			
A	0000	ABCB	1000
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B]

Dictionary

A	0000	ABCB	1000
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B]

Dictionary			
A	0000	ABCB	1000
B	0001		
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B]

Dictionary			
A	0000	ABCB	1000
B	0001	BB	1001
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B]

Dictionary			
A	0000	ABCB	1000
B	0001	BB	1001
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B]

Dictionary			
A	0000	ABCB	1000
B	0001	BB	1001
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B]

Dictionary			
A	0000	ABCB	1000
B	0001	BB	1001
C	0010		
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B] [BB] [B]

Dictionary			
A	0000	ABCB	1000
B	0001	BB	1001
C	0010	BBB	1010
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B] [BB] [B] [*EOF*]

Dictionary			
A	0000	ABCB	1000
B	0001	BB	1001
C	0010	BBB	1010
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B] [BB] [B] [*EOF*]

Dictionary

A	0000	ABCB	1000
B	0001	BB	1001
C	0010	BBB	1010
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZW: Overview

- Extends LZ78 to eliminate the requirement that the symbol at the end of a new dictionary entry be emitted as a literal, instead using it as first symbol of next prefix
- Now possible for decompressor to encounter codes before they are added to its dictionary:

LZW: Overview

- Extends LZ78 to eliminate the requirement that the symbol at the end of a new dictionary entry be emitted as a literal, instead using it as first symbol of next prefix
- Now possible for decompressor to encounter codes before they are added to its dictionary:

$$\text{AAA} \rightarrow \text{[A][AA]}$$

  - Unknown code must have been added to dictionary after encoding previously received sequence; must therefore code for the previously received sequence followed by one more character

LZW: Overview

- Extends LZ78 to eliminate the requirement that the symbol at the end of a new dictionary entry be emitted as a literal, instead using it as first symbol of next prefix
- Now possible for decompressor to encounter codes before they are added to its dictionary:

  $$\text{AAA} \rightarrow [\text{A}][\text{AA}]$$

  - Unknown code must have been added to dictionary after encoding previously received sequence; must therefore code for the previously received sequence followed by one more character
  - Last character of sequence must be same as first, since the new dictionary entry was last sequence followed by the last sequence again, followed by that character

LZW: Example

A B C B C B A A B C A B C B B B B B B

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		

LZW: Example

A B C B C B A A B C A B C B B B B B B

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		

LZW: Example

A B C B C B A A B C A B C B B B B B B

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC]

Dictionary			
A	0000		
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B]

Dictionary			
A	0000	BA	1000
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B]

Dictionary			
A	0000	BA	1000
B	0001		
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010		
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB]

Dictionary

A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011		
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100		
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101		
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101	BB	1101
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101	BB	1101
CB	0110		
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B] [BB]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101	BB	1101
CB	0110	BBB	1110
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B] [BB]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101	BB	1101
CB	0110	BBB	1110
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B] [BB]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101	BB	1101
CB	0110	BBB	1110
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B] [BB] [BBB] [*EOF*]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101	BB	1101
CB	0110	BBB	1110
BCB	0111		

LZW: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [BC] [B] [A] [AB] [C] [ABC] [B] [BB] [BBB] [*EOF*]

Dictionary			
A	0000	BA	1000
B	0001	AA	1001
C	0010	ABC	1010
EOF	0011	CA	1011
AB	0100	ABCB	1100
BC	0101	BB	1101
CB	0110	BBB	1110
BCB	0111		

LZ78: Example

A B C B C B A A B C A B C B B B B B B

[A] [B] [C] [B] [CB] [A] [AB] [C] [ABC] [B] [B] [B] [BB] [B] [*EOF*]

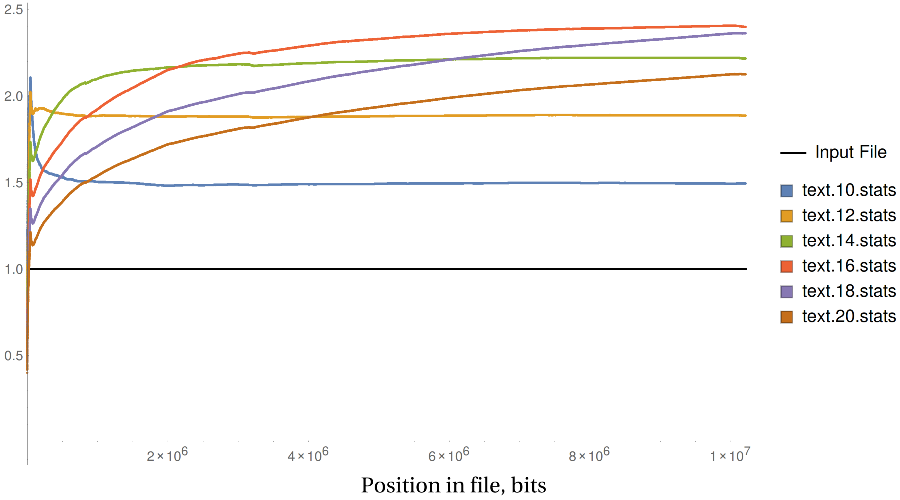
Dictionary			
A	0000	ABCB	1000
B	0001	BB	1001
C	0010	BBB	1010
EOF	0011		
AB	0100		
CB	0101		
CBA	0110		
ABC	0111		

LZW: Simple Variants

- Implementation that Welch proposed in his paper took 8-bit symbols to 12-bit codes, which remains common
- Welch also briefly mentions variable code length, where symbol width starts at the minimum necessary to represent all codes in the dictionary, and increases as the dictionary grows
- Some variants have a “clear” code, which encoder can emit to make decoder reset its dictionary, in case data being compressed changes such that previously built dictionary no longer matches input well
- Others will replace rarely-used codes with new ones when the dictionary fills up

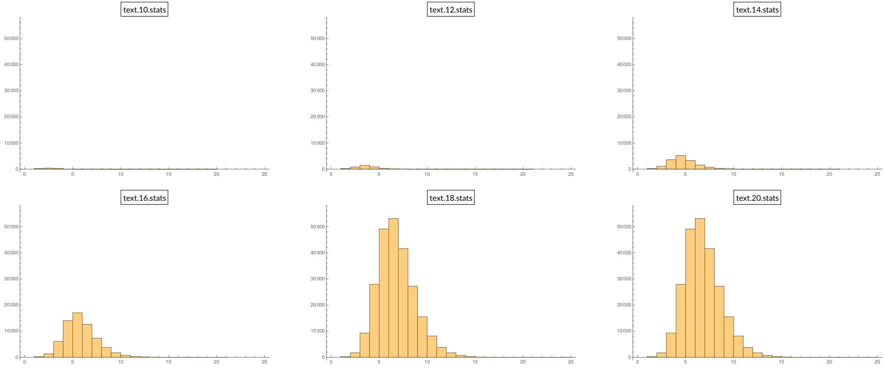
Analysis

Compression Ratios for English Text, Various Code Widths



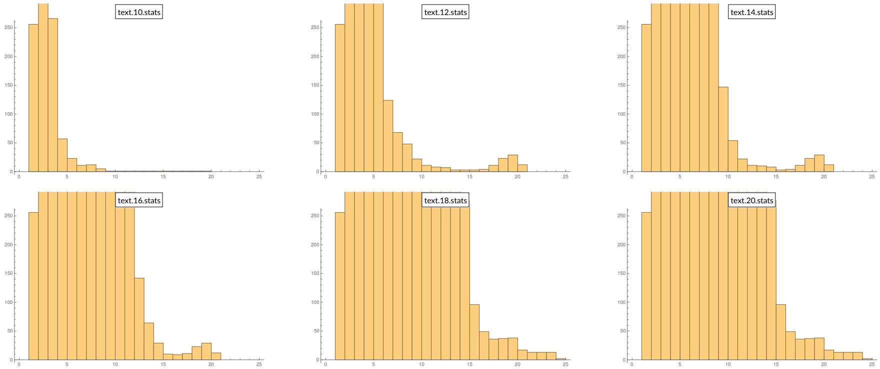
Analysis

Dictionary Entry Length, Various Code Widths



Analysis

Dictionary Entry Length, Various Code Widths



References

- Welch, T. “A Technique for High-Performance Data Compression”. *Computer* 17 (6): 8–19 (1984).
- Welch, T., inventor; High speed data compression and decompression apparatus and method. US Patent 4,558,302. December 10, 1985.
- “History of GIFLIB”. `http://giflib.sourceforge.net/history.html`
- “Lempel-Ziv-Welch (LZW) Compression”. `http://netghost.narod.ru/gff/graphics/book/ch09_04.htm`
- “History of Lossless Data Compression Algorithms”. `http://ethw.org/History_of_Lossless_Data_Compression_Algorithms`