TextDB

This is sample text for TextDB. This is sample text for TextDB. This is sample text for TextDB. This is sample text for TextDB.

Unstructured text

0	this
1	is
2	sample
3	text
4	for
5	textdb

Word Index

\+

0 1 2 3 4 5. 0 1 2 3
4 5. 0 1 2 3 4 5. 0 1
2 3 4 5. 0 1 2 3 4 5.

Indexed Docs

10101010001
01111001010
11010101010
1011111

Byte Encoded

10101010001
01111001010

Compressed

Word Index

This is a sample doc

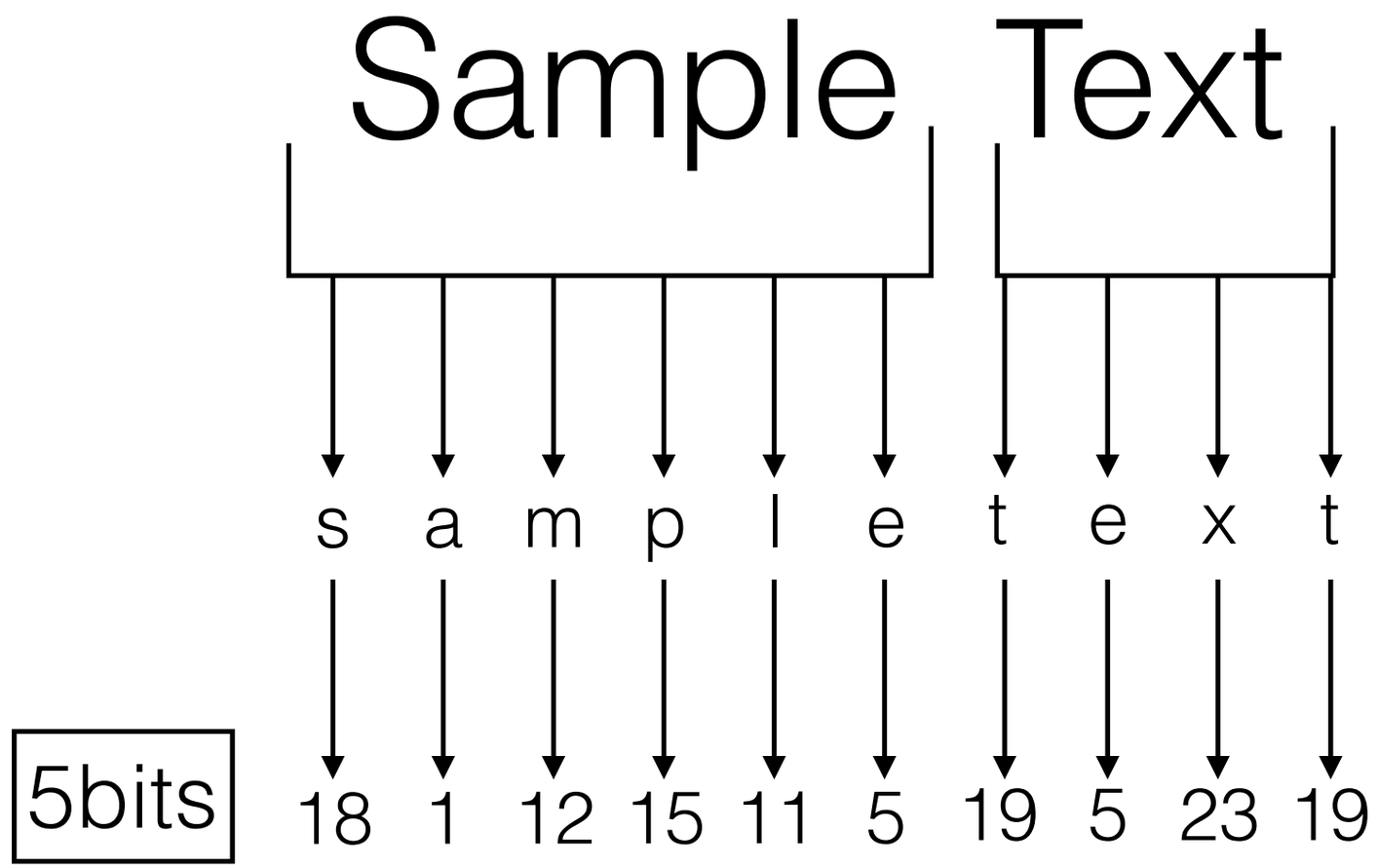
Create map from unique word stems

0	this
1	is
2	a
3	sample
4	doc

Recreate doc using indices

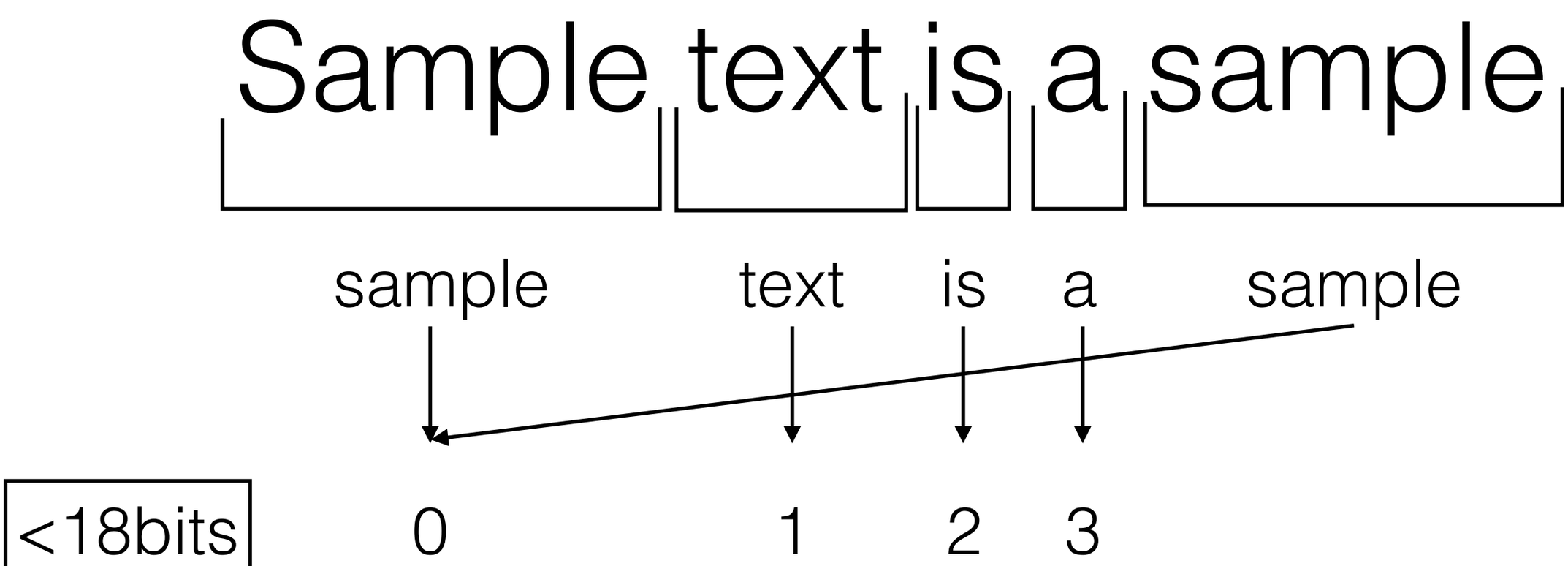
0 1 2 3 4

Character Encoding



Any lowercase english character has a value < 32

Word Index



Max idx val = 2^18. This number is chosen because number of words in the english language is < 2^18

The index grows dynamically as the number increases to save memory

Compare

Step 1: Compress characters

The standard way

Size of a character: 1 byte

Number of bits to store a character: 8 bits

The TextDB way

Number of bits to store a (lower case) character: 5 bits

37.5% compressed after step 1

Compare

Step 2: Compress docs to indices (per word)

The standard way

① Average length of word: 8

② Number of bits to store a character: 8

Average number of bits to store a word: ① x ② = 64 bits

The TextDB way

① Average number of unique words in a book: 2,500 - 10,000

Number of bits to store a word index: $\log_2(①) < 14$ bits

Since the index grows with ①, smaller docs will require far less bits

> 80% compressed after step 2

Compare

Document with 10,000 unique words (Worst Case)

The standard way

① Average length of word: 8

② Number of bits to store a character: 8

Average number of bits to store a word: ① x ② = 64 bits

Total number of bits: 10,000 x ③ = 64,000

The TextDB way

① Number of bits to store a word index: $\log_2(10{,}000) = 14$ bits

② Total number of bits in doc: 10,000 x ① = 14,000

③ Total number of bits in word index: 10,000 x 8 x 5 = 40,000

④ Total number of bits: ② + ③ = 64,000

Equal storage cost

Compare

Step 3: Compress using Google Snappy

Additional 20% - 95% depending on data