About the Test Data

Matt Mahoney
Last update: Sept. 1, 2011. History

The test data for the Large Text Compression Benchmark is the first 10^9 bytes of the English Wikipedia dump on Mar. 3, 2006. http://download.wikipedia.org/enwiki/20060303/enwiki-20060303-pages-articles.xml.bz2 (1.1 GB or 4.8 GB after decompressing with bzip2 - link no longer works). Results are also given for the first 10^8 bytes, which is also used for the Hutter Prize. These files have the following sizes and checksums:

File	Size (bytes)	MD5 (GNU md5sum 1.22)	SHA-1 (SlavaSoft fsum 2.51)
enwik8	100,000,000	a1fa5ffddb56f4953e226637dabbb36a	57b8363b814821dc9d47aa4d41f58733519076b2
enwik9	1,000,000,000	e206c3450ac99950df65bf70ef61a12d	2996e86fb978f93cca8f566cc56998923e7fe581

Download in PPMd var. J format (requires 256 MB free memory): enwik8.pmd (21,388,296 bytes) enwik9.pmd (183,964,915 bytes).

Download in zip format: enwik8.zip (36,445,475 bytes) enwik9.zip (322,592,222 bytes).

The data is UTF-8 encoded XML consisting primarily of English text. enwik9 contains 243,426 article titles, of which 85,560 are #REDIRECT to fix broken links, and the rest are regular articles. The example fragment below shows a redirection of "AdA" to "Ada programming language" and the start of a regular article with title "Anarchism".

The data is UTF-8 clean. All characters are in the range U'0000 to U'10FFFF with valid encodings of 1 to 4 bytes. The byte values 0xC0, 0xC1, and 0xF5-0xFF never occur. Also, in the Wikipedia dumps, there are no control characters in the range 0x00-0x1F except for 0x09 (tab) and 0x0A (linefeed). Linebreaks occur only on paragraph boundaries, so they always have a semantic purpose. In the example below, lines were broken at 80 characters, but in reality each paragraph is one long line.

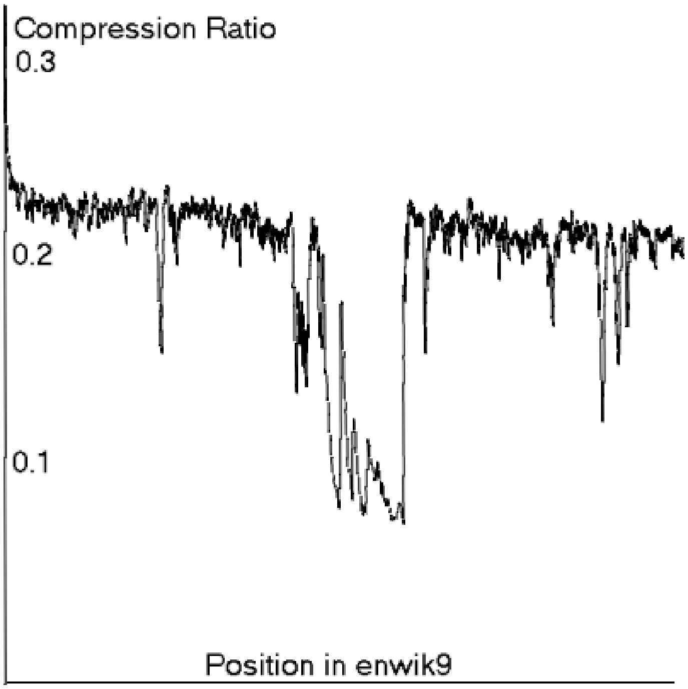
The data contains some URL encoded XHTML tags such as `&lt;ref&gt ... &lt;/ref&gt;` and `&lt;br /&gt;` which decode to `<ref> ... </ref>` (citation) and `<br />` (line break). However, hypertext links have their own encoding. External links are enclosed in square brackets in the form `[URL anchor text]`. Internal links are encoded as `[[Wikipedia title | anchor text]]`, omitting the title and vertical bar if the title and anchor text are identical. Non-English characters are sometimes URL encoded as `&amp;#945;`, meaning `&#945;` (Greek alpha, α), but are more often coded directly as a UTF-8 byte sequence.

```
  <page>
    <title>AdA</title>
    <id>11</id>
    <revision>
      <id>15898946</id>
      <timestamp>2002-09-22T16:02:58Z</timestamp>
      <contributor>
        <username>Andre Engels</username>
        <id>300</id>
      </contributor>
      <minor />
      <text xml:space="preserve">#REDIRECT [[Ada programming language]]</text>
    </revision>
  </page>
  <page>
    <title>Anarchism</title>
    <id>12</id>
    <revision>
      <id>42136831</id>
      <timestamp>2006-03-04T01:41:25Z</timestamp>
      <contributor>
        <username>CJames745</username>
        <id>832382</id>
      </contributor>
      <minor />
      <comment>/* Anarchist Communism */  too many brackets</comment>
      <text xml:space="preserve">{{Anarchism}}
'''Anarchism''' originated as a term of abuse first used against early [[working
 class]] [[radical]]s including the [[Diggers]] of the [[English Revolution]] an
d the [[sans-culotte|''sans-culottes'']] of the [[French Revolution]].[http://uk
.encarta.msn.com/encyclopedia_761568770/Anarchism.html] Whilst the term is still
 used in a pejorative way to describe ''&quot;any act that used violent means to
 destroy the organization of society&quot;''&lt;ref&gt;[http://www.cas.sc.edu/so
cy/faculty/deflem/zhistorintpolency.html History of International Police Coopera
tion], from the final protocols of the &quot;International Conference of Rome fo
```

```
r the Social Defense Against Anarchists&quot;, 1898&lt;/ref&gt;, it has also bee
n taken up as a positive label by self-defined anarchists.

The word '''anarchism''' is [[etymology|derived from]] the [[Greek language|Gree
k]] ''[[Wiktionary:&amp;#945;&amp;#957;&amp;#945;&amp;#961;&amp;#967;&amp;#943;&
amp;#945;|&amp;#945;&amp;#957;&amp;#945;&amp;#961;&amp;#967;&amp;#943;&amp;#945;
]]'' (&quot;without [[archon]]s (ruler, chief, king)&quot;). Anarchism as a [[po
litical philosophy]], is the belief that ''rulers'' are unnecessary and should b
```

The graph below shows the incremental compressed size of enwik9 over a sliding window of about 2-4 MB when compressed with ppmd var. J with options -o10 -m256 -r1 (maximum compression as in the main table). The horizontal axis is the position in the file, from 0 to 1 GB. The vertical axis is the compression ratio on a scale of 0 to 0.3. The graph was produced by modifying the source code for ppmd to print the graph coordinates, then smoothing the data for display.



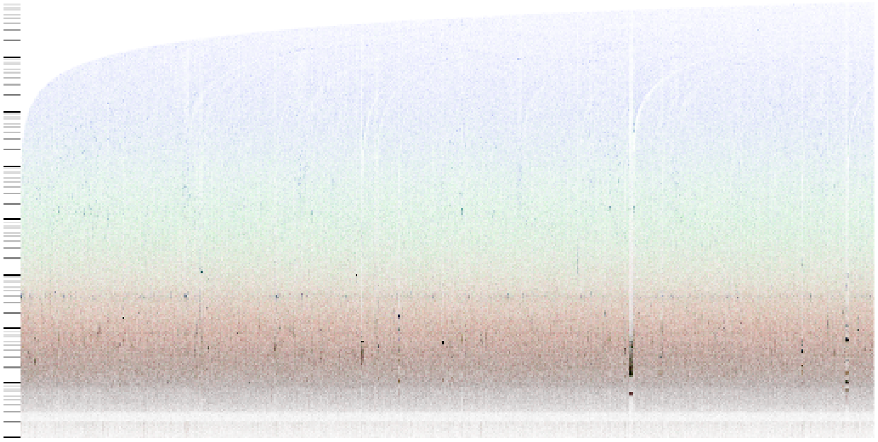
Incremental compression ratio of enwik9 with ppmd var. J set for maximum compression.

The dip in the middle of the graph is due to a large group of articles on towns in the US written in a similar style that appear to be generated automatically from a table of census data. A sample of titles from this region appears below. Unrelated articles are occasionally mixed in.

```
    Springboro, Pennsylvania
    Steuben Township, Pennsylvania
    Summerhill Township, Crawford County, Pennsylvania
    Summit Township, Crawford County, Pennsylvania
    Titusville, Pennsylvania
    Townville, Pennsylvania
    Troy Township, Crawford County, Pennsylvania
    Union Township, Crawford County, Pennsylvania
    Venango, Pennsylvania
    Venango Township, Crawford County, Pennsylvania
    Vernon Township, Pennsylvania
```

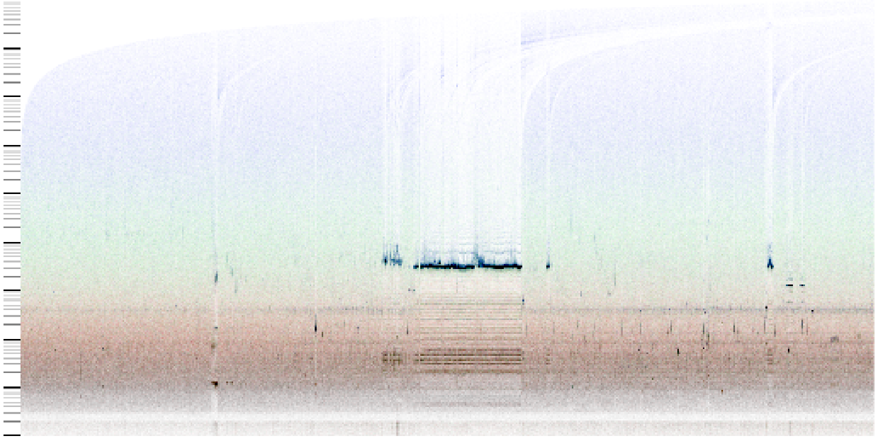
String repetition statistics

The following diagrams show the distribution of string matches of length 1 (black), 2 (red), 4 (green), and 8 (blue). The horizontal axis represents the position in the file. The vertical axis shows the distance backwards to the previous match on a logarithmic scale. The major tick marks reading upwards are 1, 10, 100, 1000, etc.



enwik8

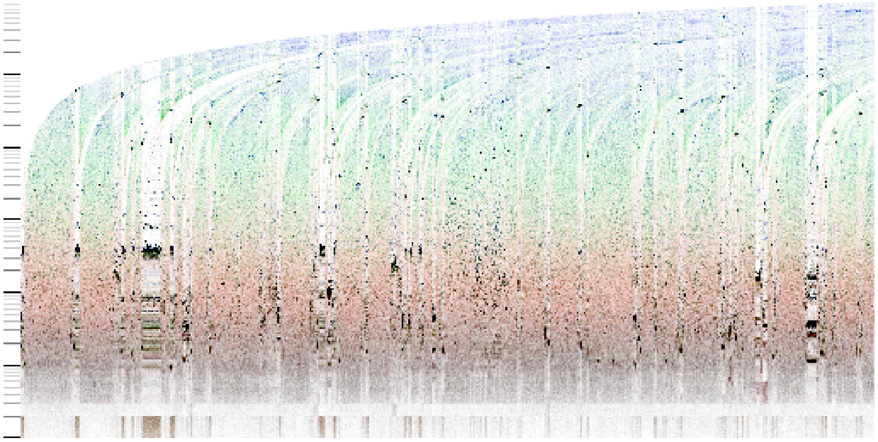
enwik8 is fairly uniform. The blue band at the top shows that matches of length 8 are most often separated by at least 10^5 bytes (5 major tick marks) up to the entire length of the file. The green band shows that matches of length 4 are most commonly separated by 10^3 to 10^4 bytes. The red band shows that matches of length 2 are separated by about 10 to 300 bytes. The gray band shows that single byte matches are usually separated by 1 or by 3 to about 10. The light gray band shows an absence of matches separated by 2, such as "aba".



enwik9

The highly compressible region in the center of enwik9 is clearly visible. The dark blue-green band shows that there are frequent matches of length 4-8 separated by about 3000 bytes, the length of one article. The articles are fairly uniform in length, but not exactly so. The dark red bands below it show a separation of around 20-80 bytes, typical of tables.

The blue region extends all the way to the top of the image, showing redundancy across the entire file. Thus, a compressor would benefit by using lots of memory. Breaking the file into smaller, separately compressed pieces would hurt compression.

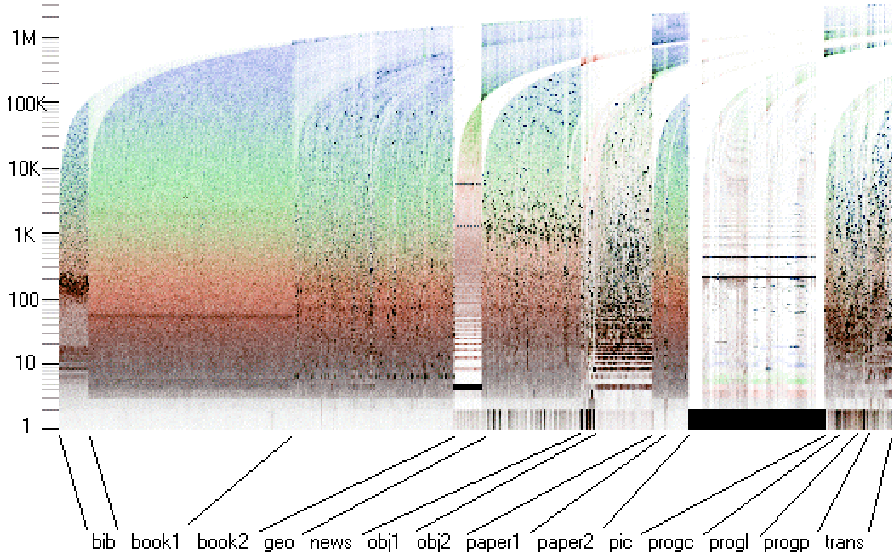


enwik6

This shows the first 10^6 bytes of data, essentially zooming into the first 1% of enwik8 or first 0.1% of enwik9. The clear, vertical bands show regions consisting primarily of XML to encode #REDIRECTs. These have a regular, repeating structure of length 300-400 bytes. The white vertical bands extending upwards show that there are no long distance matches backwards (they occur closer). The upward curving white bands show the absence of long distance forward matches.

These images were generated with the FV program. Download: fv.cpp (7KB, GPL, C++), fv.exe (25KB, 32-bit Windows). The program uses 512 MB memory. It took about 14 minutes to generate the image for enwik9 on a 2.2 GHz Athlon-64 under WinXP Home. The program outputs in .bmp format. You will need another program to convert to .jpg, .png, etc.

For comparison, the Calgary corpus (concatenated) is shown below. The image was hand edited to add the labels. The repetitive structure of book1 (70-80), geo (4), obj2 (4), and pic (216) is clearly visible. The image also shows that there is redundancy between text files but not the binary files. (More FV results by Leonardo Maffi).



Calgary corpus (image edited to add labels)

Lexical Analysis

The table below shows the frequency distribution for the 100 most common words in enwik9, as well as selected ranks of lower frequency. The file was parsed by considering all sequences of English letters (A-Z, a-z) as single words, and all other single characters as words. The most common token is the space, which occurs 139 million times. ^J is the linefeed character. Upper and lower case are considered distinct.

Freq	Rank	
139132610	1	
20216224	2	]
20214205	3	[
13147025	4	^J
9578832	5	.
8824782	6	'
7978922	7	,
6498257	8	the
6199271	9	;
6037325	10	1
5754062	11	0
5277489	12	/
5274001	14	< >
5084982	15	&
4907106	16	|
4848627	17	of
4721386	18	2
4252985	19	=
3542288	20	-
3536470	21	:
3239688	22	9
3050674	23	and
2736971	24	3
2596182	25	5
2579245	26	4
2523537	27	)

```
2521806      28 (
2490470      29 *
2443325      30 8
2372995      31 6
2288680      32 in
2130130      33 a
2122603      34 to
2005136      35 7
1875367      36 quot
1630357      37 is
1420384      38 id
1304048      39 The
1152218      40 lt
1151080      41 gt
 970827      42 %
 968882      43 s
 919535      44 amp
 901320      45 }
 901242      46 {
 861221      47 for
 852516      48 are
 815600      49 was
 741327      50 as
 715225      51 by
 707706      52 with
 638091      53 from
 606095      54 that
 579119      55 on
 541826      56 title
 532311      57 or
 530859      58 #
 522024      59 page
 513474      60 text
 494564      61 _
 489280      62 revision
 487656      63 contributor
 486934      64 timestamp
 486902      65 "
 442286      66 Ã
 426510      67 sup
 414409      68 at
 412849      69 http
 410993      70 username
 410788      71 S
 408867      72 it
 407150      73 Category
 382932      74 comment
 382925      75 an
 378934      76 U
 373978      77 his
 361588      78 have
 356881      79 which
 348190      80 be
 344554      81 In
 328703      82 www
 303270      83 Ð
 301709      84 Census
 281944      85 he
 276642      86 T
 276098      87 age
 273405      88 also
 267662      89 space
 259584      90 has
 257619      91 population
 253569      92 Z
 250192      93 td
 249502      94 American
 245321      95 preserve
 244574      96 xml
 244207      97 not
 243328      98 were
 236835      99 A
 226459     100 who
...
  92138     200 John
...
  34136     500 President
```

```
...
  15918    1000 instead
...
   7394    2000 Album
...
   2675    5000 Episode entrance perspective
...
   1088    9996 Ahmed Basil Chang Jakob Papua demanding
   1087   10003 Conservatives Hogan Vulcan fingers hospitals nautical ...
...
    422   19968 Atl Berks Billings Clovis Demons FAA Foundations Fujian G...
    421   20001 Armageddon Bella Champaign Cleese Comt Counsel DavidLevin...
...
    108   49714 AUC Abbe Accelerated Acupuncture Adria Agentsoo AirPort A...
    107   50021 ABN ALU Acacius Ach Aemilianus Agony Alfalfa Ardea Atoka A...
...
     35   99959 AAW ABR AIBO AICPA AIESEC AKS AUTf Abaddon Abomination Ad...
     34  101715 AFSC ARRL AXN Aare Aaronson Abai Abeokuta Abode Absolutis...
...
     10  212858 AAE AAEECC AAG AAMs AAWW AAbout AApre AApro ABAKO ABZ ACC...
      9  227146 AAAAA AACR AAFFAA AAPT AAV AAs ABCDEFGHIJKLMNOPQRSTUVWXYZ...
      8  244837 AABA AACO AADA AAO ABCDAB ABCNews ABSA ACAP ACTIVITY ACiD...
      7  266715 AAFP AAIB ABSTRACTS ABr ACADEMY ACATS ACCA ACCP ACs ADDIN...
      6  294832 AACS AAUP AArticles ABCDABD ABEND ABERR ABH ABVV ACBL ACI...
      5  332370 AAAHH AABB AABBA AAFSS AAJA AARON AASI AAZ ABCDABCDABDE A...
      4  389883 AAAAFF AAABA AABAB AABBFF AACM AADT AAFLA AAN AANR AAPG A...
      3  481186 AAAD AABAA AACA AACCA AAJ AAK AAMCO AAPA AAPBL AAPL AASB A...
      2  686619 AAAAAAA AAAAB AAABB AAABN AAADE AAALAC AABBB AABC AABEBWU...
      1 1418809 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA...
```

Note that the data deviates from a Zipf distribution (rank x frequency = constant). The vocabulary is 1,418,809 words. An order 0 model based on this distribution would have a compressed size of 400,889,188 bytes. In addition, an order 0 character encoding of the dictionary would require 7,044,509 bytes for a total of 407,933,697. This does not include a small amount of additional space that would be needed to encode the word frequencies, lengths of the dictionary and compressed data. For comparison, some other order 0 compressed sizes are given below for various parsing methods compared to this baseline parsing method.

Table (corrected Sep 2 2006) Order 0 compressed size of enwik9 using various parsing methds. Text, Dict, and Total are compressed sizes in bytes for an ideal coder with an order 0 model. Vocabulary is the number of words in the dictionary.

Text	Dict	Total	Vocabulary	Method
400889188	7044509	407933697	1418809	Words (baseline, described above)
644561309	198	644561507	206	1-gram
565716295	40632	565756927	21377	2-gram
503362325	987242	504349567	368046	3-gram
451610359	6136119	457746478	1851158	4-gram
408622694	20819133	429441827	5351189	5-gram
407112108	3186214	410298322	686619	Unique words are spelled
404716717	3186918	407903634	687011	Unique words spelled with 2-grams
403950277	3266186	407216463	719061	Unique words spelled with 3-grams
403378168	3725601	407103770	881331	Unique words spelled with 4-grams
409885058	2171224	412056282	417010	Words occurring once or twice are spelled
406488898	2172095	408660993	481672	Words occurring once or twice are spelled with 2-grams
370304509	7044513	377349022	1418810	Space modeling
414760783	6423232	421184015	1286437	Stemming
423866419	5003920	428870340	1094898	Words with capital encoding
406286891	6235043	412521935	1235411	Capital/lower encoding for less common type

In the 1-gram through 5-gram models, the text is divided into uniform blocks of 1 to 5 bytes. The 5-gram model compresses almost as well as the word model but the dictionary is almost 4 times as large.

Slight compression occurs when words that occur only once are spelled with 2, 3, or 4-grams rather than added to the dictionary. Spelling such words with letters results in a slight expansion.

Spelling words that occur once or twice does not compress as well as spelling words that occur only once. (That might not be true for higher order models).

In space modeling, words are assumed to be preceded by a space in certain contexts, and the space is removed from the encoded text. If a space is predicted but none occurs, then a special symbol is added to encode this fact. This results in an increase of 1 to the vocabulary size (compared to baseline, no words are spelled). A 7.5% improvement in compression over baseline was obtained by assuming that a space occurs before the word after any upper or lower case letter, or the characters "." (period), "," (comma), "]" (closing bracket), "}" (closing brace), or ")" (closing parenthesis). There are 2,196,539 no-space symbols in enwik9, making it the 33rd most frequent.

Stemming consists of replacing an inflected form of a word with its base form (stem) plus a symbol indicating the suffix. This was accomplished with a low error rate by trying a set of stemming rules on each incoming word and testing if the resulting stem occurs frequently enough in the baseline dictionary (collected in an earlier pass). Each word is tested to see whether it ends in one of the suffixes in the stemming table, and if so, the suffix is replaced and the frequency of the resulting word is tested from the baseline dictionary. If the stem occurs at least 1/16 as often and is at least 3 characters long, then the word is coded as a stem plus a suffix code. If more than one rule could apply, then the rule that produces the highest frequency stem is used.

The stemming table is below, sorted by "Freq", the number of times each rule was applied in enwik9.

Suffix	Replacement	Freq
"s"	""	7776340
"ed"	""	1740722
"ed"	"e"	1396317
"ing"	""	1128007
"er"	""	928728
"ly"	""	925579
"ing"	"e"	800321
"ies"	"y"	557109
"ion"	""	371964
"ion"	"e"	331435
"er"	"e"	302743
"ers"	""	144059
"ation"	""	121642
"ence"	"ent"	102084
"ation"	"e"	97988
"est"	""	92200
"ly"	"le"	69732
"ers"	"e"	56579
"est"	"e"	36774
"sses"	"ss"	34569
"ier"	"y"	32638
"nning"	"n"	23684
"mming"	"m"	19079

For example, the words "rotates", "rotated", "rotation", and "rotating" would all stem to "rotate". One problem is that the baseline dictionary is case sensitive, so that "Rotates" is stemmed only if "Rotate" occurs. There are occasional errors such as stemming "coming" to "com" + "ing" and "refer" to "ref" + "er". This happens because "com" is more common then "come" (in links) and "ref" occurs as an XHTML tag to encode references. The minimum stem length of 3 prevents "as" from being stemmed as the plural of "a" and similar errors.

Stemming makes order 0 compression worse due to the additional suffix tokens, but it reduces the dictionary size and might help in a model that uses syntactic or semantic modeling.

The simplest form of capital encoding is to replace each upper case letter with a special symbol followed by the lower case equivalent. This hurts compression over baseline but reduces the dictionary size and helps in some higher order models. enwik9 contains 41,507,612, making this the second most common symbol after space.

Some words such as proper nouns are always capitalized, so it is wasteful to use capital encoding. An improvement is to build a baseline dictionary and test whether the version with the first letter capitalized or lower case is more frequent (e.g. "Pat" or "pat"), and store that version in the dictionary. The less common form is encoded by preceding it with a special symbol to indicate the case of the first letter should be changed. This results in a larger dictionary but better compression than simple capital encoding. There are 15,065,442 change-case symbols in enwik9, making it the fourth most common symbol after space, [, and].

Relationship of Wikipedia Text to Clean Text

(June 11, 2006) Abstract: The entropy of "clean" written English, in a 27 character alphabet containing only the letters a-z and nonconsecutive spaces, has been estimated to be between 0.6 and 1.3 bits per character [3,8]. We find that most of the best compressors will compress Wikipedia text (enwik9, 1 GB) and equivalent cleaned text (fil9, 715 MB) to about the same ratio, usually within 3% of each other. Low end compressors will compress clean text about 5% smaller. Furthermore, a quick test on 100 MB of cleaned text (text8) will predict a compression ratio that is about 2% to 4% below the true ratio on fil9 for most compressors.

Introduction

Most data compression benchmarks, including the large text benchmark (enwik9), use data sets with unknown algorithmic complexity. For most benchmarks, this is not important because their purpose is to compare data compression algorithms to one another. However, this benchmark has the goal of encouraging research in natural language models, so we also wish to compare algorithms to human models. Shannon [3] and Cover and King [8] estimated the entropy of written English to be between 0.6 and 1.3 bits per character, based on the ability of humans to predict consecutive characters from a 27 character alphabet containing only the monocase letters a-z and nonconsecutive spaces. However, enwik9 is not in this form; it contains capitalization, punctuation, foreign text, tables, markup, formatting, hypertext links, and XML structure such as timestamps, authorship, and comments. In this paper we estimate the effects of these artifacts on compression ratio for 25 programs.

Experimental Procedure

We filter the 1 GB test file `enwik9` to produce a 715 MB file `fil9`, and compress this with 17 compressors. Furthermore, we produce the file `text8` by truncating fil9 to 100 MB, and test this on 25 compressors, including the 17 tested on fil9. The purpose of the smaller file is to allow quicker testing, and to establish the predictive value of this quick test on the larger data set.

The clean version of the Wikipedia was prepared with the goal of retaining only text that normally would be visible when displayed on a Wikipedia web page and read by a human. Only regular article text was retained. Image captions were retained, but tables and links to foreign language versions were removed. Citations, footnotes, and markup were removed. Hypertext links were converted to ordinary text, retaining only the (visible) anchor text. Numbers were spelled out ("20" becomes "two zero", a common practice in speech research). Upper case letters were converted to lower case. Finally, all sequences of characters not in the range a-z were converted to a single space. The effect of this filtering on enwik8 is to reduce the text to about 70% of its original size before spelling digits, then expand it to about 74%. The detailed effect of each step is shown in the table below for enwik8 (which would result in the first 74 MB of fil9 or text8. The effects of individual steps was not tested on enwik9, but the final result is a little smaller (71.5%)

```
  enwik8         Step
-----------   ------------
100,000,000   Original size
 96,829,911   Discard all outside <text...> ... </text>
 96,604,864   Discard #REDIRECT text
 96,210,439   Discard XML tags (<text...> and </text>)
 95,287,203   URL-decode &lt; &gt; and &amp; to < > and &
 95,087,290   Remove <ref> ... </ref> (citations)
 93,645,338   Remove other XHTML tags
 91,399,021   Replace [http:... anchor text] with [anchor text]
 90,868,662   Replace [[Image:...|thumb|left/right|NNNpx|caption]] with caption
 90,770,617   Replace [[category:text|title]] with [[text]]
 88,385,654   Remove  [[language:link]]  (links to same page in other languages)
 85,443,983   Replace [[Wiki link|anchor text]] with [[anchor text]]
 83,420,173   Remove  {{...}}  (icons and special symbols)
 80,943,967   Remove  { ... }  (tables)
 77,732,609   Remove  [ and ]
 75,053,443   Replace &...; with space (URL-encoded chars)
 70,007,945   Convert to lower case, replace all sequences not in a-z,0-9 with a single space
 74,090,640   Spell digits, leaving a-z and unrepeated spaces
```

The conversion was done by the Perl program given in Appendix A. The following example shows what the previous example looks like after conversion (although in reality there are no line breaks).

```
 anarchism originated as a term of abuse first used against early working class
radicals including the diggers of the english revolution and the sans culottes
of the french revolution whilst the term is still used in a pejorative way to
describe any act that used violent means to destroy the organization of society
it has also been taken up as a positive label by self defined anarchists the word
anarchism is derived from the greek without archons ruler chief king anarchism as
a political philosophy is the belief that rulers are unnecessary and should b
```

The two files have the following sizes and checksums. text8 is the first 10^8 bytes of fil9.

```
File      Size               MD5 checksum            Download
-----  -----------  --------------------------------  --------
fil9   713,069,767  2754e1cfcc34288745cd23272d976384  (use wikifil.pl to generate from enwik9)
text8  100,000,000  3bea1919949baf155f99411df5fada7e  text8.zip, 31,344,016 bytes (or truncate fil9)
```

Experimental Results

Compressed sizes of text8 and fil9 are given in the table below. For each compressor, the options are selected as in the main table (as of June 10, 2006), which were tuned for maximum compression on enwik9. (Note this may bias the results toward raw text). The column t8/e8 is the ratio of the compressed size of text8 to the compressed size of enwik8. It shows that the clean text usually compresses smaller by a few percent. The enwik8 results are from the main table, as is the algorithm and memory used (in MB). Decompression was not verified. Speed was not measured.

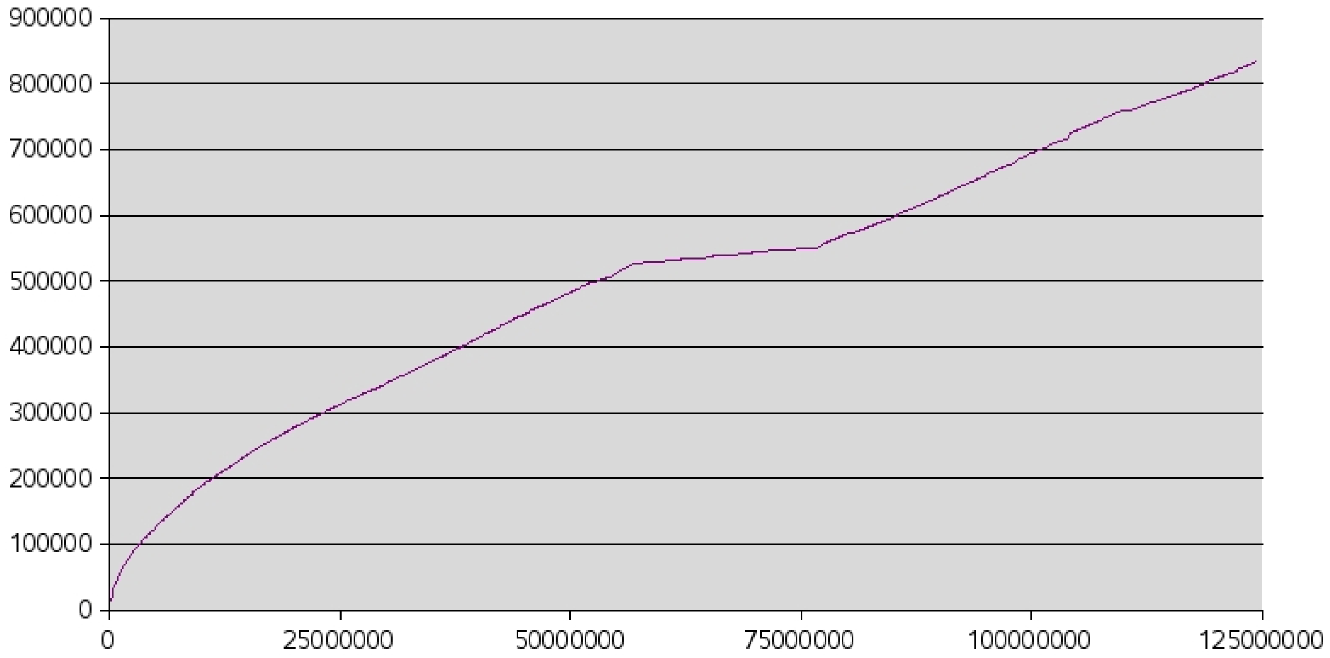
The fil9 results are shown as both the compressed size and compression ratio, since the uncompressed size is odd. The column f9/e9 is the compression ratio of fil9 divided by the compression ratio for enwik9, *not* including the size of the decompressor in either case. (The decompressor size has a very small effect). The compression ratio for enwik9 is from the main table. The error, t8/e8 - f9/e9 is the amount by which a test on text8 would underestimate the compressed size of fil9, usually about 2% to 4%.

Program, version, options	text8	enwik8	t8/e8	Alg	Mem	fil9 size	ratio	f9/e9	Error	Note
paq8h -7	17,461,782	17,674,700	.9879	CM	854	108,316,091	.1519	1.0320	-.0441	
durilca 05 -m800 -o12 -t2(3)	17,712,518	18,520,589	.9563	PPM	700					
xwrt|ppmonstr -f800 -m800 -o8	18,059,570	19,043,178	.9483	CM	800	108,740,992	.1525	.9854	-.0371	
ppmonstr J -m800 -o16	18,649,962	19,230,657	.9698	PPM	800	116,872,140	.1639	1.0149	-.0451	
slim23d -m700 -o10	18,421,266	19,264,094	.9562	PPM	700	115,433,736	.1619	.9960	-.0398	
ash 04a /m700 /o10	18,680,324	19,963,105	.9357	CM	700	121,001,817	.1697	.9389	-.0032	
WinUDA 2.91 mode 3	19,282,080	20,332,366	.9488	CM	194					
uhbc 1.0 -m3 -b800	19,724,021	20,930,838	.9424	BWT	800					
hipp 5819 /o8	20,026,417	20,555,951	.9743	CM	719					
ppmd J -m256 -o10 -r1	20,029,751	21,388,296	.9365	PPM	256	127,644,785	.1790	.9731	-.0366	
enc 0.15 aq	20,361,492	22,156,982	.9190	CM	50	132,364,111	.1856	.9490	-.0283	
M03exp 2005-02-15 (32 MB)	20,495,661	21,948,192	.9338	BWT	32					14
ocamyd LTCB 1.0 -s0 -m3	20,683,435	21,285,121	.9696	DMC	300					6
sbc 0.970r2 -ad -m3 -b63	20,723,754	22,470,539	.9222	BWT	224	133,110,739	.1867	.9473	-.0251	
bssc 0.95a -b16383	21,395,109	23,117,061	.8948	BWT	140					
ocamyd 1.65f -s0 -m8	21,419,608	21,456,536	.9983	DMC	800					
GRZipII 0.2.4 -b8m	22,019,644	23,846,878	.9233	BWT	58	141,150,532	.1979	.9471	-.0238	
uharc 0.6b -mx -md32768	22,841,858	23,911,123	.9552	PPM	50	147,933,009	.2075	.9977	-.0425	
px v1.0	23,846,604	24,971,871	.9549	CM	66					
cabarc 1.00.0601 -m lzx:21	25,662,446	28,465,607	.9015	LZ77	20	165,676,761	.2323	.9266	-.0251	
bzip2 1.0.2 -9	26,395,400	29,008,736	.9099	BWT	8	169,311,654	.2374	.9349	-.0250	
kzip 5/13/06 /b1024	31,344,016	35,016,649	.8951	LZ77	121					
gzip 1.3.5 -9	33,048,240	36,445,248	.9068	LZ77	1	213,697,635	.2997	.9290	-.0222	
pkzip 2.0.4	33,319,889	36,934,712	.9021	LZ77	1	215,527,700	.3023	.9226	-.0245	
lzop v1.01 -9	38,806,161	41,217,688	.9415	LZ77	1	251,384,828	.3525	.9623	-.0408	
compress 4.3d	39,179,237	45,763,941	.8561	LZW	1	259,977,297	.3645	.8587	-.0026	
fpaq0	51,551,380	63,391,013	.8132	o0	1	366,426,423	.5139	.8011	+.0121	
Uncompressed	100,000,000	100,000,000	1.0000			713,069,767	1.0000	1.0000	.0000	

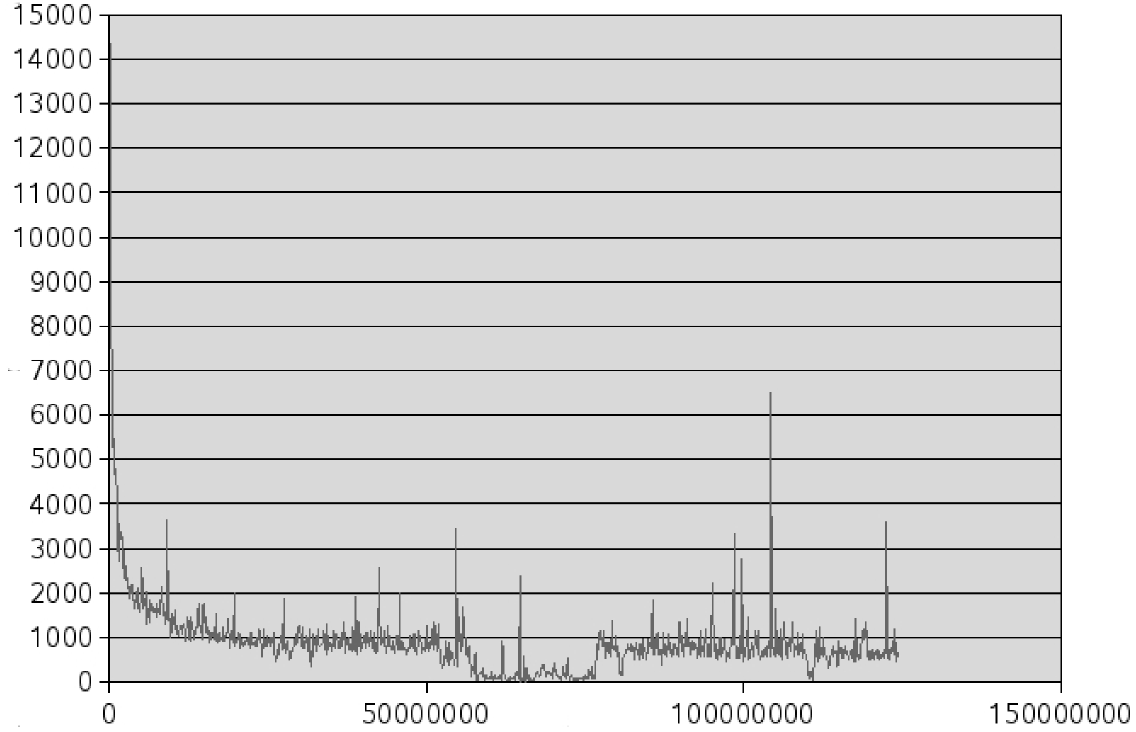
The results show that high end compressors have about the same compression ratio on clean text as on raw text, while faster compressors will compress clean text about 5% smaller. Using text8 as a fast test to predict the effect of cleaning enwik9 tends to underestimate the compressed size of clean text by about 2% to 3% on faster compressors and by about 4% on high end compressors.

Related Work

Alexandru Mosoi has produced some preprocessing tools to improve compression of fil9, discussed here, and has produced the graph below which shows the vocabulary size vs. word token count of fil9. The graph is consistent with a Zipf distribution: the n'th most frequent word has frequency proportional to 1/n [20].



Cumulative dictionary size vs. word count of fil9 (by Alexandru Mosoi).



New words added to dictionary (per 2^{17} word block) vs. word count of fil9. This is essentially the derivative of the above graph (by Alexandru Mosoi).

References

3. Shannon, Cluade E., “Prediction and Entropy of Printed English”, Bell Sys. Tech. J (3) p. 50-64, 1950.

8. Cover, T. M., and R. C. King, “A Convergent Gambling Estimate of the Entropy of English”, IEEE Transactions on Information Theory (24)4 (July) pp. 413-421, 1978.

20. Zipf, George Kingley, *The Psycho-Biology of Language, an Introduction to Dynamic Philology*, M.I.T. Press, 1935.

Appendix A

This Perl program filters Wikipedia text dumps to produce 27 character text (lowercase letters and spaces) as described in this article. To use:

```
perl wikifil.pl enwik9 > text
```

Then truncate the text to the desired length (e.g. 10^8 bytes).

You can cut and paste the program below. (Note it contains URL encoding to display properly).

```
#!/usr/bin/perl

# Program to filter Wikipedia XML dumps to "clean" text consisting only of lowercase
# letters (a-z, converted from A-Z), and spaces (never consecutive).
# All other characters are converted to spaces.  Only text which normally appears
# in the web browser is displayed.  Tables are removed.  Image captions are
# preserved.  Links are converted to normal text.  Digits are spelled out.

# Written by Matt Mahoney, June 10, 2006.  This program is released to the public domain.

$/=">";                     # input record separator
while (<>) {
  if (/<text /) {$text=1;}  # remove all but between <text> ... </text>
  if (/#redirect/i) {$text=0;}  # remove #REDIRECT
  if ($text) {

    # Remove any text not normally visible
    if (/<\/text>/) {$text=0;}
    s/<.*>//;               # remove xml tags
    s/&amp;/&/g;            # decode URL encoded chars
    s/&lt;/</g;
    s/&gt;/>/g;
    s/<ref[^<]*<\/ref>//g;  # remove references <ref...> ... </ref>
    s/<[^>]*>//g;           # remove xhtml tags
    s/\[http:[^] ]*/[/g;    # remove normal url, preserve visible text
    s/\|thumb//ig;          # remove images links, preserve caption
    s/\|left//ig;
    s/\|right//ig;
    s/\|\d+px//ig;
    s/\[\[image:[^\[\]]*\|//ig;
    s/\[\[category:([^|\]]*)[^]]*\]\]/[[$1]]/ig;  # show categories without markup
    s/\[\[[a-z\-]*:[^\]]*\]\]//g;  # remove links to other languages
    s/\[\[[^\|\]]*\|/[[/g;  # remove wiki url, preserve visible text
    s/{{[^}]*}}//g;         # remove {{icons}} and {tables}
    s/{[^}]*}//g;
    s/\[//g;                # remove [ and ]
    s/\]//g;
    s/&[^;]*;/ /g;          # remove URL encoded chars

    # convert to lowercase letters and spaces, spell digits
    $_=" $_ ";
    tr/A-Z/a-z/;
    s/0/ zero /g;
    s/1/ one /g;
    s/2/ two /g;
    s/3/ three /g;
    s/4/ four /g;
    s/5/ five /g;
    s/6/ six /g;
    s/7/ seven /g;
    s/8/ eight /g;
    s/9/ nine /g;
    tr/a-z/ /cs;
    chop;
    print $_;
  }
}
```