COMPARISON OF CHARACTER COUNTING PROGRAMS

How Character Counting Software Reads and Interprets Text and MSWord Files

(test conducted 8/2/13 with 6 popular programs)

How character counting programs work

There are a number of programs designed to count characters in documents, whether they are MSWord doc files, text files, or other filetypes. Some word processors have a simple character counting function built in. Other programs are designed specifically to count characters, words, or phrases. However, the results seem to vary with almost every program. We ran tests using some of the most popular character counting programs to determine which programs were the most accurate. In determining their accuracy, we needed to examine exactly how characters were stored in the common file formats. This involved both visible characters, and hidden characters and control codes..

Sample files to determine what characters are counted

To help us analyze these character counting programs, we created three test files to use with our test programs. These files were each designed to test different abilities in these character counting software programs. The files are available for download by clicking on the filename in the list below. The files are:

File: `rawtext.txt` download file

This is a plain text file that has been saved in UTF-8 format to accommodate upper ASCII characters. It was created by entering text in a simple text editor and pasting in some special characters from a Word document. It is not very long, but is meant to test simple character counting and also counting of special characters that may be represented internally by more than one character. This analysis is for the UTF-8 format. Files in other formats will have different internal representations but we have found UTF-8 to be the most reliable way to display and edit text files. The actual file details and counts as verified by a binary editor are:

Item	# chars
Visible Characters	
Regular characters including: - punctuation - 40 upper ASCII accented characters - 70 special characters pasted from Word (en dash/em dash/curly quotes)	4324
Spaces	600
Tabs	100
Character Count	**5024**
Hidden Characters	
Hidden line endings (CR/LF) - 2 characters for each of 124 lines	248
Hidden extra characters for each of the 40 upper ASCII character since they are represented by 2 internal characters	40
Hidden extra characters for the 70 characters pasted from Word. Each character is represented internally by 3 characters, so we have 2 hidden characters for each one.	140
Actual File Size in Characters	**5452**

File: `PrincessOfMars.txt` download file

This document was downloaded as an HTML file from the Project Gutenberg website at http://www.gutenberg.org/ebooks/62. The file was then converted to text to ensure that sentences were not truncated on every physical line. It was also saved in UTF-8 format to ensure proper handling of special characters. The resulting file contains a few double and triple byte characters as shown in the chart below. This file was used to demonstrate the character counting programs' performance on a larger file.

Item	# chars
Visible Characters	
Regular characters including: - punctuation - 35 upper ASCII accented characters - 70 special triple-byte characters	318,458
Spaces and tabs	69,324
Character Count	**387,782**
Hidden Characters	
Hidden line endings (CR/LF) - 2 characters for each of 2721 lines	5,442
Hidden extra characters for each of the 35 upper ASCII character since they are represented by 2 internal characters	35
Hidden extra characters for each of the 70 triple byte characters. Each character is represented internally by 3 characters, so we have 2 hidden characters for each of these 70 characters	140
Actual File Size in Characters	**393,399**

File: `rawtextWithHeaders.doc` download file

The last file is a MS Word doc file that was created by opening the rawtext.txt file in Word and then adding a few Word-specific items. I added headers and footers, and a few footnotes and endnotes to see how the various counting software programs handle these items. Not all of the programs can handle Word files directly so we did not test those with this file. Also, the total file size calculation is not really meaningful for Word files since a lot of extra information is added internally in the file by Word. This also makes it impossible to accurately estimate the amount of internal storage for hidden characters. However, the main numbers of total characters should be consistent and accurate. It appears that none of the counting programs included the running headers and footers in their totals. Word reports the characters in footnotes and endnotes separately so we included those numbers on a separate line.

rawtextWithHeaders.doc

Item	# chars
Visible Characters	
Regular characters including: - punctuation - 40 upper ASCII accented characters - 70 special characters pasted from Word (en dash/em dash/curly quotes)	4,328
Headers and Footers - 29 characters on 3 pages - 87 total characters that were not included in any character counts	---
Spaces	600
Tabs	100
Initial Character Count	**5,028**
Footnotes and endnotes - 4 notes totaling 67 text characters and two extra characters per note (we didn't test having a large number of footnotes so using two per note may not scale)	75
Total Character Count	**5,103**
Probable Hidden Characters	
Hidden line endings (CR/LF) - 2 characters for each of 124 lines plus 8 lines for footnotes and endnotes	264
Hidden extra characters for each of the 40 upper ASCII character since they are represented by 2 internal characters	40
Hidden extra characters for each of the special 70 characters. Each character may be represented internally by 3 characters, so we have 2 hidden characters for each of these 70 characters	140
Actual File Size not relevant since Word adds many bytes	***

Character Counting Program Comparison

We tested a number of character counting programs by opening our three test files and running the count analysis and recording the results. Several of the programs had fairly consistent results and the *Word Count* feature on the Word *Tools* menu agreed exactly with our calculated character counts, as did *myWordCount* on all three tests and *Word Count Manager* on two of the tests. Some of the programs did not detail what they included in the total figure, so it was difficult to determine why their totals were different from the actual values.

Results of testing rawtext.txt

	Actual document	MS Word	my WordCount	SmartEdit	Count Anything	Word Count Manager	AnyCount
Version		*2003*	*3.07*	*3.001*	*2.1*	*2.5.2*	*8.0.7*
Total w/o spaces/tabs	4,324	4,324	4,324	n/a	4,324	4,324	n/a
Spaces/Tabs	700	700	700	n/a	n/a	700	n/a
Total characters	5,024	**5,024**	**5,024**	4,924	4,874	**5,024**	5,094

Indicates a program's character count exactly matches the actual numbers

Two of the programs undercounted characters, most likely due to missing some of the triple-byte special characters pasted in from Word. *AnyCount* actually overcounted the total, most likely due to counting what should be a hidden character as a visible character. MS Word, *myWordCount*, and *Word Count Manage*r all reported correct total counts.

Results of testing PrincessOfMars.txt

	Actual document	MS Word	my WordCount	SmartEdit	Count Anything	Word Count Manager	AnyCount
Version		*2003*	*3.07*	*3.001*	*2.1*	*2.5.2*	*8.0.7*
Total w/o spaces/tabs	318,458	318,458	318,458	n/a	n/a	318,458	318,316
Spaces/Tabs	69,324	69,324	69,324	n/a	n/a	69,324	n/a
Total characters	387,782	**387,782**	**387,782**	387,754	386,026	**387,782**	387,656

Indicates a program's character count exactly matches the actual numbers

The larger text file was mostly analyzed correctly or close to correctly. In this test, three of the programs undercounted characters, while the same three programs get the exact correct total.

Results of testing rawtextWithHeaders.doc

	Actual document	MS Word	my WordCount	SmartEdit	Count Anything	Word Count Manager	Total Assistant
Version		*2003*	*3.07*	*3.001*	*2.1*	*2.5.2*	*2.6.0.4*
Total w/o spaces/tabs	4,328	4,328	4,391	n/a	4,408	4,328	
Spaces/Tabs	700	700	712	n/a	n/a		n/a
Total characters	5,028	**5,028**	n/a	n/a			
Footnotes and endnotes (4 total)	75			n/a			
Total including footnotes and endnotes	5,103	**5,103**	**5,103**	5,139	4,970	5,028	5,124

Indicates a program's character count exactly matches the actual numbers

The last test of the MS Word document file had slightly more erratic results. Some of the character counting programs we looked at could not read MS Word files directly, so those programs could not be used. The doc file added several new variables that would affect character count, mainly headers, footers, footnotes, and endnotes. Since the programs that reported incorrect results had no indication, either on the screen or in their help files, of what was and was not counted in the Word document, we were unable to determine exactly what caused the discrepancies. For this test, only *MS Word* and *myWordCount* reported the correct total count.

Details of the Character Counting Programs Tested

Program	my WordCount	SmartEdit	Count Anything	Word Count Manager	AnyCount	Total Assistant
Version	*3.07*	*3.001*	*2.1*	2.5.2	*8.0.7*	*2.6.0.4*
Count listed for each character	**yes**	**no**	**no**	**no**	**no**	**no**
Special Word characters counts reported separately	**yes**	**yes**	**no**	**no**	**no**	**no**
Bar chart histogram showing count for each character	**yes**	**no**	**no**	**no**	**no**	**no**
Price	**$14.95**	**$59.95**	**freeware**	**$29.95**	**$49.95**	**$29.95**

Tests performed: 8/2/13