Wörter zählen in LaTeX-Dokumenten

Uwe Ziegenhagen

Die Information über die Anzahl der Wörter und Absätze ist in gängigen Textverarbeitungen meist nur einen Mausklick entfernt, einige LaTeX-Editoren wie Kile bieten diese Funktion ebenso. Doch abseits von Kile & Co ist man auf die Hilfe externer Tools angewiesen, von denen ich einige in diesem Artikel vorstellen möchte.

wordcount

Das Testdokument besteht aus 100 Worten des bekannten Lorem Ipsum Textes, zwei section Überschriften mit insgesamt drei Worten und drei Formeln, die jeweils ein Wort in einem text Befehl ausgeben, also insgesamt 106 Worte.

Unter Unix/Linux ist wc die Standardlösung, wenn es um das Ermitteln der Wörter eines Textes geht. wc testdokument.tex gibt die die Anzahl der Zeilen, Wörter und Zeichen aus, im Beispiel 21 130 997. An der Ausgabe wird klar, dass wc nicht für die explizite Verwendung mit LaTeX ausgelegt ist, da alle LaTeX-spezifischen Kommandos als Wörter gezählt werden.

Das das Bereinigen der Quelldatei mittels untex ist vom Ergebnis her noch schlechter, da zu wenige LaTeX-Befehle entfernt werden. Ein untex testdokument| wc gibt deutlich vom richtigen Ergebnis abweichende 21 134 880 aus. TeX Live bietet zusätzlich das Kommando detex, das ebenso versucht, LaTeX-Kommandos aus dem Dokument zu entfernen. Angewandt auf unser Dokument ergeben sich gute 108 Wörter. Dabei entfernt detex zwar Konstrukte wie documentclass, section und usepackage, lässt aber die Paketnamen im Text stehen. Dafür entfernt es rigoros mathematischen Input, das text{Pythagoras} in den Formeln wird daher auch unterschlagen.

Als Alternative zum „Entrümpeln“der LaTeX-Datei kann auch die PDF-Datei per pdftotext umgewandelt werden. Dessen Ausgabe unterdrückt naturgemäß alle LaTeX-Befehle, wandelt jedoch auch die Formeln um. Das Ergebnis von wc, angewandt auf die Ausgabe von pdftotext, ergibt 125 Wörter. Verwandte Resultate erbringt auch ps2ascii, das 123 Wörter zählt.

LaTeX word count von E. Rødland

Sehr gute Ergebnisse erzielt LaTeX word count von Einar Andreas Rødland, das unter `http://folk.uio.no/einarro/Comp/texwordcount.html` zum Download bereitsteht

und in seiner neuesten Version 2.2.beta auch mit UTF-8 und Chinesisch/Japanisch umgehen kann. Das in Perl geschriebene Programm steht online zur Verfügung und kann auch heruntergeladen werden. Die Resultate, verglichen mit den bisher vorgestellten Lösungen, sind deutlich besser: Es werden 100 Worte im Text gefunden und drei in den Überschriften. Ignoriert werden nur die Worte, die innerhalb der text Umgebungen stehen.

texWordCount

Ebenfalls in Perl geschrieben ist das aus Singapur stammende texWordCount, das unter `http://wing.comp.nus.edu.sg/~min/texWordCount/` zum Download und online verfügbar ist. Rein qualitativ ist es deutlich schlechter als LaTeX word count, für das Testdokument errechnet es 122 Wörter.

LaTeX Word Counter

LaTeX Word Counter ist ein Java-Programm und über `http://sourceforge.net/projects/lwc/` erhältlich. Das Programm startet eine kleine grafische Nutzeroberflöche, über die die durchzuzählende Datei geladen wird.

Zusammenfassung

Von allen vorgestellten Lösungen gefällt LaTeX word count am besten. Von allen Programmen kam es am nächsten an die tatsächliche Wortzahl heran und besticht durch die gut gemachte Webseite und die Tatsache, dass es vom Autor aktiv gepflegt wird. Im Einzelfall hängt es jedoch ab, welche Struktur das Dokument hat, in dem man die Wörter zählen möchte. Bei großen text-lastigen Dokumenten spielt es sicherlich eine geringere Rolle, ob Formeln bzw. der Text in Formeln gezählt werden als naturwissenschaftlichen Veröffentlichungen.

LaTeX word count

Format/colour codes of verbose output:

Text which is counted .... counted as text words
Header and title text .... counted as header words
Caption text and footnotes .... counted as caption words
Ignored text or code .... excluded or ignored
\documentclass .... document start, beginning of preamble
\macro .... macro not counted, but parameters may be
\macro .... macro in excluded region
[Macro options] .... not counted
\begin{group} \end{group} .... begin/end group
\begin{group} \end{group} .... begin/end group in excluded region
$ $ .... counted as one equation
$ $ .... equation in excluded region
% Comments .... not counted
%TC:TeXcount instructions .... not counted
File to include .... not counted but file may be counted later
ERROR .... TeXcount error message

```
\documentclass { scrartcl }
\usepackage [latin1]{ inputenc }
\usepackage [T1]{ fontenc }
\usepackage [ngerman]{ babel }
\usepackage []{ amsmath }

\begin {document}

\section { Einleitung }

Lorem ipsum dolor sit amet , consetetur sadipscing elitr , sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat , sed diam voluptua . At vero eos et accusam et justo duo dolores et ea rebum . Stet clita kasd gubergren , no sea takimata sanctus est Lorem ipsum dolor sit amet . Lorem ipsum dolor sit amet , consetetur sadipscing elitr , sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat , sed diam voluptua . At vero eos et accusam et justo duo dolores et ea rebum . Stet clita kasd gubergren , no sea takimata sanctus est Lorem ipsum dolor sit amet . $ \text { Phythagoras } \quad a ^ 2 + b ^ 2 = c ^ 2 $

\section { Abgesetzte Mathematik }

\begin {equation}
\text { Phythagoras } \quad a ^ 2 + b ^ 2 = c ^ 2
\end {equation}

$$ \text { Phythagoras } \quad a ^ 2 + b ^ 2 = c ^ 2 $$

\end {document}
```

Word count
Words in text: **100**
Words in headers: **3**
Words in float captions: **0**
Number of headers: **2**
Number of floats: **0**
Number of math inlines: **1**
Number of math displayed: **2**

Abb. 1: Ausgabe von LaTeX word count für das Testdokument