

**NAME**

4a\_mergeWordCounts.pl

**SYNOPSIS**

```
% perl 4a_mergeWordCounts.pl x y z
```

x = <keep prose if >= N words>

y = <keep poetry if >= N words>

z = <stats for top N words>

i.e. % perl 4a\_mergeWordCounts.pl 250 40 100

means keep prose if greater than or equal to 250 words, keep poetry if greater than or equal to 40 words, and get the stats for the top 100 words.

**DESCRIPTION****SUMMARY**

Note: This is the last in a series of four scripts that we used to generate the data that is available via our website tools. (If you want to set up your own experiments, see the Analysis Software scripts).

Assumes the script "countWords.pl" has already been run (see README for that script to learn of assumed directory structure).

Process the stats (word counts, relative frequency) of the texts one at a time; combine these stats for each text across the entire "corpus". A tab delimited text file is created with each word given its own column, and the rows corresponding to each text in the "corpus". This shows the word counts for each text in the input. A statistic file is created with the word and its mean usage per text. A summary file is created with the number of files, the number of files kept (over the minimum requirement), the number of distinct words, number of true hapax legomena, and number of words that appear once in each text.

**INPUT**

We are assuming you are running this script on the entire collection of poetry and prose. (See Analysis Software to run your own subset). Make a copy of the word\_counts folder (the output of 2\_countWords.pl) in this current directory.

WARNING: running the entire corpus on a Linux box (2.4 GHz) took about 8 hours. If you want to run a smaller subset of your own files, see Analysis Software.

**OUTPUT**

A "RESULTS" folder containing two .xls files--one stats file, with each top word and its usage per text, and one summary file, with the number of files, number of files kept, distinct words, and hapax legomena. A "corpus\_input" folder is also created, which contains a tab delimited text file with the word counts per text across the entire input corpus. This is the folder that is necessary for input in the next program, 4b\_minMaxMean.pl.

**AUTHORS**

Mark D. LeBlanc  
Christina L. Nelson

**MODIFICATION HISTORY**

June 14, 2007 (mdl) --

just getting started

**June 18, 2007 (mdl) --**

takes 36 minutes to dump all 2444 texts (size: 780M)  
with 165,821 distinct words over all the texts :(  
added command-line arg for the minimum number of words  
in order to keep a file; new stats in this case:  
total texts: 1493 distinct words: 161,224 size: 460M

**June 18, 2007 (mdl) --**

reserved the zeroth row in the hash of arrays to hold  
the AVERAGE PERCENTAGE use of each word, e.g.,  
\$wordCount{"and"}[0] is the average percentage for 'and'  
over all the texts with total words >= \$MIN\_WORDS

**Aug. 3, 2007 (mdl) --**

reserved the first \$FIRST\_TEXT\_COL's for AVG,MEDIAN,STDdev  
NOTE: only AVG is implemented currently in zeroth column

new command-line arg for top-N-words to print and accompanying subroutine

**Aug. 7, 2007 (mdl) --**

drat, running out of memory; try reducing size of word\_count hash by  
not keeping hapax words

**Aug. 7, 2007 (mdl) --**

computing avg. and stdev relative freq. for each word across all texts

**Aug. 10, 2007 (mdl) --**

now using Statistics::Lite for mean and stddev (rather than me doing it)

**Sept 14, 2007 (mdl) --**

add separate cmd-line args for minimum words, poetry vs. prose  
keepProse >= arg1 keepPoetry >= arg2  
arg3 is now the topNwords as before

**June 11, 2008 (cIn) --**

do not remove hapax legomena, completely run on Linux

**May 12, 2009 (mdl, cIn) --**

update documentation with the assumption that this script is run  
on the entire collection of poetry and prose

**COPYRIGHT INFORMATION**

=====  
Copyright (C) 2009 Wheaton Lexomics Research Group, Norton MA

This program is free software: you can redistribute it and/or modify  
it under the terms of the GNU General Public License as published by  
the Free Software Foundation, either version 3 of the License, or  
(at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <<http://www.gnu.org/licenses/>>.