

NAME

as4a_mergeWordCounts.pl

SYNOPSIS

```
% perl as4a_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 as4a_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.

Note: If you want all the words, use a BFN.

DESCRIPTION**SUMMARY**

Note: Users of the Virtual Manuscript tool would use this script after building their virtual manuscript. If you have used the Virtual Manuscript tool, pay special attention to the directions in the INPUT section.

Assumes the script "as3_countWords.pl" has already been run or the Virtual Manuscript tool has been used. (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; merge these stats for each text across all subfolders within the input "word_counts" directory. A tab delimited text file is created for output with each word given its own column, and the rows corresponding to each input text. An additional statistics file is created with the word and its mean usage per text. A third output summary file is created with the number of texts, the number of texts kept (over the minimum requirement), the number of distinct words, number of true hapax legomena (appears only once in the entire input group), and total number of words that appear only in one text (may appear more than once in the text).

INPUT

The user must take special care to build the correct input directory structure. In the same directory as this script, there should be an input folder called "word_counts". Within "word_counts", the user should create subfolders for each genre. For example, if your input set contains both poetry and prose text files, then the user should create two subfolders within "word_counts" called "A_Poetry" and "B_Prose". Alternately, if your input files were only prose, only the "B_Prose" subfolder is necessary.

If you just ran as3_countWords.pl

The output of the as3_countWords.pl script is a folder containing the word counts for each text, for example, "analysis_word_counts_1_0_c_0". This folder contains subfolders named after the short title of the texts that were counted. An example of this is the subfolder named "Dan", which contains the word counts for the Daniel text from the A01 manuscript. These subfolders should be transferred to the appropriate subfolder of "word_counts" ("A_Poetry" or "B_Prose") in the "word_counts" folder just created. For example, the "Dan" folder from "analysis_word_counts_1_0_c_0" should be copied into the "A_Poetry" subfolder of "word_counts" in order for this script to run properly. In this case, the final input relative path expected by this script for the Daniel texts would be "word_counts/A_Poetry/Dan/".

If you just used Virtual Manuscript

The output of the Virtual Manuscript tool is a folder containing all of the .csv files. In order for this script to run properly, each of the .csv files must be put into a separate subfolder, based on their name. For example, if one of the files in your virtual manuscript was the file "A01.002_Ex_T0020_keepTagged_none_c.csv", a folder named "Ex" should be created, and this .csv file should be placed in the "Ex" folder. Then the "Ex" folder with the .csv file contained in it

should be moved into the "word_counts/A_Poetry/" folder. If the input text was prose, its folder with the .csv file should be moved into the "word_counts/B_Prose/" folder. Referring back to the previous example, the "Ex" folder should be placed in the "A_Poetry" subfolder of "word_counts". In this case, the final input relative path expected by this script for the Exodus text would be "word_counts/A_Poetry/Ex/".

OUTPUT

A new "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 new "corpus_input" folder is also created, which contains a tab delimited text file with the word counts per text across the entire input group. This is the folder that is necessary for input in the next program, as4b_getStats_prepare4R.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 (cln) --

do not remove hapax legomena, completely run on Linux

May 12, 2009 (mdl, cln) --

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/>>.