

NAME

as4b_getStats_prepare4R.pl

SYNOPSIS

```
% perl as4b_getStats_prepare4R.pl
```

DESCRIPTION

SUMMARY

Note: Assumes the script "as4a_mergeWordCounts.pl" has already been run (see README for that script to learn of assumed directory structure).

Creates a tab delimited file with the word counts across each text that was in the "as4a_mergeWordCounts.pl" "corpus_input" folder. It also finds the minimum, mean, median, maximum, and standard deviation of the word counts for each word across the entire input group, and puts those results into a second tab delimited file.

INPUT

Copy the "corpus_input" folder produced by the "as4a_mergeWordCounts.pl" script into the directory holding this script. That .txt file is read in, and used to create the tables and calculate the statistics.

OUTPUT

A new "STATISTICS_RESULTS" folder with two tab delimited.txt files. First "totalCounts.txt" with the total counts of the words across the input group (in the format a word per row, and a text per column). This is a rotated format (rows to columns) than that of the .txt file used as input. The second file, "statsPerWord.txt" contains the statistics (minimum, mean, median, maximum, and standard deviation), of the usage of the word over the group.

Our research group then passes the "totalCounts.txt" file to R for further processing. For example, clustering or classification analyses.

AUTHORS

Christina L. Nelson
Mark D. LeBlanc

MODIFICATION HISTORY

May 12, 2008 (cIn) --

create necessary data structures

May 13, 2008 (cIn) --

store the information in a hash of arrays

May 15, 2008 (cIn) --

change the standard deviation to be divided by (n-1) rather than n

May 20, 2008 (cIn) --

write dump_data subroutine

June 13, 2008 (cIn) --

change from one input file with all the information to two output files, one with statistics and the other with counts across the corpus

June 24, 2008 (cIn) --

change from .xls files to .txt (tab delimited)

May 13, 2009 (mdl, cIn) --

updated documentation

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/>>.