
NAME

as2_cutter.pl

SYNOPSIS

```
% perl as2_cutter.pl x y z
```

x = <# of words in a chunk>

y = <# of words to shift over>

z = <last chunk size>

i.e. % perl as2_cutter.pl 3000 1000 0.50

means each chunk will have 3000 words, the starting point for the subsequent chunk will be 1000 words beyond the current starting point (for example, first chunk starts at word 1, second chunk would start at word 1001), and the percentage of the last chunk, that is, 0.50 means the last chunk must be 50% or greater than all the other chunk sizes in order to be its own chunk. Otherwise, the remaining text will be added onto the previous chunk, thus the last chunk could be larger than the others.

The first number should be larger or at least equal to the second number. If the numbers are equal, you will get chunks that do not overlap. If the second number is larger, you will get chunks that miss words.

Another Example:

```
% perl as2_cutter.pl 2000 2000 0.50
cut texts into 2000 non-overlapping word chunks
```

DESCRIPTION**SUMMARY**

Script to split texts into smaller chunks with or without overlapping chunks. Assumes the script "as1_sortIntoDirectories.pl" has already been run (see README for that script to learn of assumed directory structure).

INPUT

This script snags one file at a time from the "texts_to_split" directory in the same folder as this script. For this script to work properly, the filenames must match the file names as created by the "as1_sortIntoDirectories.pl" script. For example, the Daniel text from manuscript A01 would be named A01.003_Dan_T00030.txt. It takes three command line arguments. See above.

OUTPUT

One new directory is created called "split_texts" in the directory with this script. Within this folder is any number of subdirectories that share the name with each short title contained in the input filename. For example, the Daniel text has a short title of "Dan", so this is the name of the new subdirectory within "split_texts". These files are to be used in "as3_analysis_countWords_disEm.pl".

AUTHORS

Amos C. Jones
Christina L. Nelson
Mark D. LeBlanc

MODIFICATION HISTORY

June 12, 2008 (acj) --

wrote and debugged script

June 16, 2008 (acj) --

finished commenting and wrote pod
fixed the last chunk problem

June 17, 2008 (acj) --

changed output file names to include genre and manuscript and text numbers

reworked algorithm to work work with chunk sizes and shift sizes that are
not multiples of each other

Feb 10, 2009 (cln) --

fixed bug (found by Scott Kleinman's bug): now handles case when
CHUNK SIZE is greater than overall size of file

May 12, 2009 (mdl, cln) --

update 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/>>.