
NAME

generateStats.pl

SYNOPSIS

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

```
x = <1=keep_tagged_words | 0=discard_tagged_words>
```

```
y = <0=no_consolidations | 10=consolidate_and | 20 =consolidate_thorn | 30 =consolidate_both>
```

```
z = <w=sortByWords | c=sortByCounts>
```

```
i.e. % perl generateStats.pl 1 0 c
```

means keep_tagged_words, no_consolidations, and sortByCounts

DESCRIPTION**SUMMARY**

```
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; gather stats on all files in one manuscript (for each manuscript), then all stats on all manuscripts in one genre (for each genre), then all stats on all genres in the entire corpus.

INPUT

The word_counts folder created by the "wordCounts.pl" that has already been run. We assume the word counts to be input are stored up one directory level, that is, in "../2_countWords/word_counts". See the variable \$WORD_COUNTS in code below. If the files are in a different directory, you should change this variable.

OUTPUT

A "statistics" folder with subfolders for each genre (ex. A_Poetry), as well as each manuscript (ex. A01). Each folder will contain a .csv file with word counts and frequencies for its respective information. For example, if the user decides to keep tagged words, not do any consolidations, and sort by count, the statistics (corpus) folder has a "corpus_keepTagged_none_c_STATS.csv" file, which is the word counts/relative frequencies for the entire corpus. The A_Poetry (genre) folder has an "A_Poetry_keepTagged_none_c_STATS.csv" file, which is the word counts/relative frequencies for the A_Poetry genre. Lastly, the A01 (manuscript) folder has an "A01_keepTagged_none_c_STATS.csv" file, which is the word counts/relative frequencies for the A01 manuscript.

```
Example (Partial) output for entire corpus:
corpus 144293 81548 2339694
RANK HTML WORD RAW WORD COUNT RELATIVE FREQUENCY
1 & <code>&&</code> 75801 0.0323978
2 on <code>on</code> 61911 0.0264612
3 and <code>and</code> 51865 0.0221674
```

AUTHORS

Mark D. LeBlanc
Christina L. Nelson

MODIFICATION HISTORY

May 23, 2008 (cIn) --

getting started

May 27, 2008 (cIn) --

create hashes, count, create directories, write
print function for manuscript.

May 29, 2008 (mdl) --

make one printStat() function to handle
all four hashes; handle new proportions

May 30, 2008 (cIn) --

change output to .csv, write pod

June 4, 2008 (cIn) --

add second header line, subsort alphabetically
before printing sorted by counts

June 11, 2008 (cIn) --

add rank to output

June 16, 2008 (cIn) --

change filenames to include command line information,
i.e. A01_keepTagged_none_c_STATS.csv

May 11, 2009 (mdl, cIn) --

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/>>.