Text Statistics Tool

This tool performs the following types of analysis on each text file in input:

- Distinct words count
- N-Grams statistics
- Total words count
- Word frequency distribution
- Word length distribution

Input

- You can give in input a set of specific N-Grams to look for and it will compute the frequency of the input word lists. For example, the routine can search in all the text files in a folder to show how many times the words: “the court”, “white men”, “sheriff” appear in the files. Regular expressions (not wildcards) are supported (for more information see: https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html).
- You can choose to multiply the frequency of each of the search terms by enclosing the term in “(” “)” and writing the scaling factor preceded by *, for example “(sheriff)*10”.
- You can choose the following grouping options:
  - By date (same day, month, quarter or year)
  - All together: the statistics are computed as if all the articles were one
  - No grouping: each file has its own statistics
- The analysis can be case-sensitive

Output

- A csv file that contains the statistics mentioned above regarding the text files you have chosen to analyze.
- If you choose to group the articles by date (day, month, quarter, year) then the results will be organized in groups. In each group there will be all the text files with the chosen characteristic (same day, same month, …). All types of analysis will be performed on each group and all the results for each group will be printed in the same output file.

Output file overview

The output file is organized in different sections:

- Groups composition (group label - group components)

  In this section there are two columns, one containing the labels of each group, the other containing the list of the filenames that belong to that group.
  The group labels can be a number (if you choose to group the text files by year for example, the label will be the year), or a string (if you choose not to group the files then the group label will be the filename).
- Searched N-Grams

  This section contains in one comma separated list the N-Grams that have been searched in each group
- Searched N-Grams statistics

  This section contains the statistics regarding each searched N-Gram. The results are organized in a table where the firs column identifies the group for which the results apply, the other columns contains the frequency of each of the searched N-Grams.
- Words count

  This section contains the count of the words considered in each group, specifying also how many distinct words were found.
- Word length distribution

This section contains the frequency distribution of the length of the words for each group. There are two columns, one for the word length, the other for its frequency.

- General N-Grams statistics, organized in:
    - 2-Grams frequencies
    - 3-Grams frequencies
    - 4-Grams frequencies
    - 5-Grams frequencies

  The results for each N-Gram are organized in two columns, one contains the N-Gram, the other its frequency. N-Grams with a frequency lower than 2 are not displayed.
- Word Frequencies

  This section contains the frequencies of all the words in each group. The results are organized in two columns, the firs one containing the word, the second one its frequency.

Each one of these sections has as many subsections as the groups are. In each of those subsections there are the global statistics regarding each group.

Pre-requisites for grouping by date

If your file name contains a date and it can be split into multiple parts, you can choose a separator and the position of the date part in the file name in order to aggregate the statistics of each text file by date.
For example, the file name "The Atlanta Constitution_01-12-1899_1.txt" could be a valid file name to be processed. In this case you would have to choose the separator that you used (in this case "_") and the position of the date object in the file name (in this case 1 because there is the newspaper name in the position 0, "The Atlanta Constitution").
The tool can group files by day, month, quarter (quarters are numbered from 0 to 3) or year.

You can also choose to get global statistics based on all files grouped together or to get separate statistics regarding each file.

Notes

- The tool removes the punctuation signs in each text file and considers all the words separated by one or more whitespace characters (space, tab, new line characters for example). If you choose to perform the analysis in a non-case-sensitive way, then all the text is converted to lowercase and then processed.
- Only plain text files are supported.
- All languages that separate words with one or more whitespace characters can run the analysis offered by this tool.