Pre-Processor application can scan all reports in a directory and pre-processes all reports, then generate input files for SVM-Light, SVM-Rank or LDA.

Instructions

Main Window

- **Ignore words**
  The words in this list will be removed from the stemmed results.
- **Ignored Number Suffixes**
  Usually this list contains units. Any word in the format of num***suffix***, num-***suffix***, num1-num2***suffix***, num1-num2-***suffix*** will be removed from the stemmed results, where num, num1 and num2 are numbers.
- **Save->Reports**
  Save a list of all reports scanned, based on the 3 options in the bottom left.
- **Save->Words**
  Save a list of all stemmed words from all reports, based on the 3 options in the bottom left. This is list is necessary for exporting to SVM files.
- **Export->To Text**
  Export all reports with stemmed words into a text file. The text file format is "report_name word1 word2 …".
- **Export->To SVM**
  Export all reports with stemmed words to SVM files, based on the 3 options in the bottom left. This requires a word list.
- **Export->To LDA**
  Export all reports with stemmed words to LDA file, based on the 3 options in the bottom left. This requires a word list.
- **Import->As Negative (-1)**
  Set a list of reports to be negative, with target value -1 for SVM. Reports will be unchecked in Reports list. You can also manually change the check states of any reports.
- **Import->As Neutral (0)**
  Set a list of reports to be neutral, with target value 0 for SVM. Reports will be unchecked in Reports list. You can also manually change the check states of any reports.
- **Import->As Positive (1)**
  Set a list of reports to be positive, with target value +1 for SVM. Reports will be unchecked in Reports list. You can also manually change the check states of any reports.
- **Sorted Words**
  Stemmed words will be sorted for saving or exporting.
- **Unique Words**
  Stemmed words will be kept unique for saving and exporting. That is, a word will occur at most once in a report.
- **Unique Reports**
  This will remove duplicate reports for saving and exporting.

Export to SVM

You need to select a word list file which contains all word in ascending order from the whole corpus. This file can be generated from the main window's save words function.

- **Term Frequency** and **Invert Document Frequency**
  These two options define the value for every feature (a stemmed word). More detailed can be found at https://en.wikipedia.org/wiki/Tf%E2%80%93idf#Definition
- **Decimals**
  Maximum decimal places for any real value saved in the output files.
- **Include 0 feature value**
  If checked, a feature with value 0 will be saved in the output file, otherwise it is ignored. Unchecking this option can significantly reduce the output file sizes.

Other libraries and tools

- **LDA**
  http://gibbslda.sourceforge.net/

- **SVM Light**
  http://svmlight.joachims.org/

- **SVM Rank**
  http://www.cs.cornell.edu/People/tj/svm_light/svm_rank.html

Attached Files

Options: Sorted Words **unchecked**, Unique Words **unchecked**, Unique Reports **checked**.

- **filter_number_suffix.txt**
  This contains all number suffixes used for stemming.
- **filter_words.txt**
  This contains all words to be removed from stemming.
- **report_list.txt**
  A list containing all **unique** reports, in ascending order.
- **stemmed_report_list.txt**
  A list containing all unique reports, with their stemmed words, in ascending order.
- **stemmed_words.txt**
  All stemmed words in ascending order. This list is generated from the above two filters.
- **svm_light_2_1.txt**
  A SVM Light source file generated with Raw Frequency TF and Unary IDF. 4 decimals places and 0 value features are skipped.
- **svm_rank_2_1.txt**
  A SVM Rank source file generated with Raw Frequency TF and Unary IDF. 4 decimals places and 0 value features are skipped.