Rare Word Filter

S. Matthew English

February 2015

1 Introduction

This document comprises a brief summary of a program which applies the post-processing step of eliminating rare characters from the raw data compiled by the system described in 'Chinese Character Component (部首) Scraper'. The software described herein can be found at the following GitHub page or in the directory:

`\Dropbox\Work\Code\character_component_table\rare_word_filter`

2 System Architecture

The initial step is following all the links associated with the characters represented on the menu page of commonly used ideographs and scraping the character associated with it. This functionality can be found in the file:

`\spiders\filter_creator.py`.

Subsequently this data must be "cleaned", or rather, rendered in a format which makes it amenable to direct comparison with the data set generated by 'Chinese Character Component (部首) Scraper', hereafter referred to as the 'raw data'. This functionality is found in: `extract_ideograph.java`, which outputs a file entitled `commonly_used_characters_extracted.txt` comprised of all commonly used characters, one per line, each separated by a blank line.

The raw data is found in the file
`character_component_document_including_rare.txt`
The Java program `rare_word_filter.java` uses
`commonly_used_characters_extracted.txt` as a metric, whereby if a character is listed there and does not appear in
`character_component_document_including_rare.txt` then that instance of `character_component_document_including_rare.txt` is not transmitted to the data structure comprising the output of `rare_word_filter.java`, after this procedure is complete, the resulting table of commonly used words and their component characters is output to the file `refined_table.txt`.