String Master 0.47

User Guide

Purpose and functionality of the program

String Master Utility is designed to split one string data source file into two or more files in accordance with the chosen mask or substrings. Main application areas of the software are the search of the required entries in the database, forum and web log dumps, text files of large volume, automation of the reference database sorting by domains and names.

(Figure 1. Reference window)

String Master reads the source file information string by string, verifies its compliance with the specified substrings or mask before putting in buffer. As the buffer fills the sorted data is stored in files, the names of which duplicate the name of the initial file with the addition of the sorting numbers of masks in the program list. Strings that do not match any of the user-defined masks are recorded in the file with a zero.

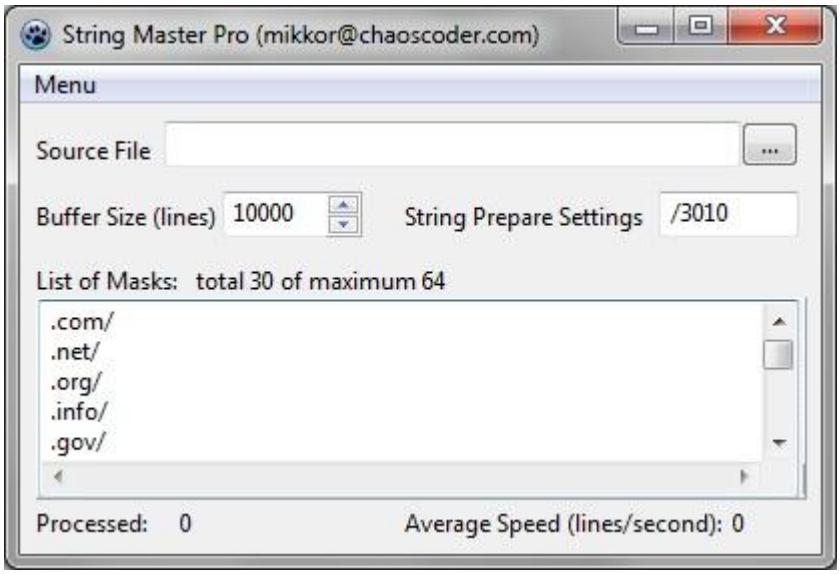
Basic and Professional versions of String Master

Starting from the version 0.46 the features of the Lite and Pro builds of String Master are almost identical. The main difference with the professional version is the possibility to work simultaneously with 256 masks, while the basic version is limited to 32. Also, on each start of the free version a window displays, a so-called Nag screen, where the user has to press one of five buttons to access and work with the program. If you purchase the professional version, the Nag screen will not appear. Otherwise the features of these versions are identical.

(Figure 2. Nag screen of the free version)

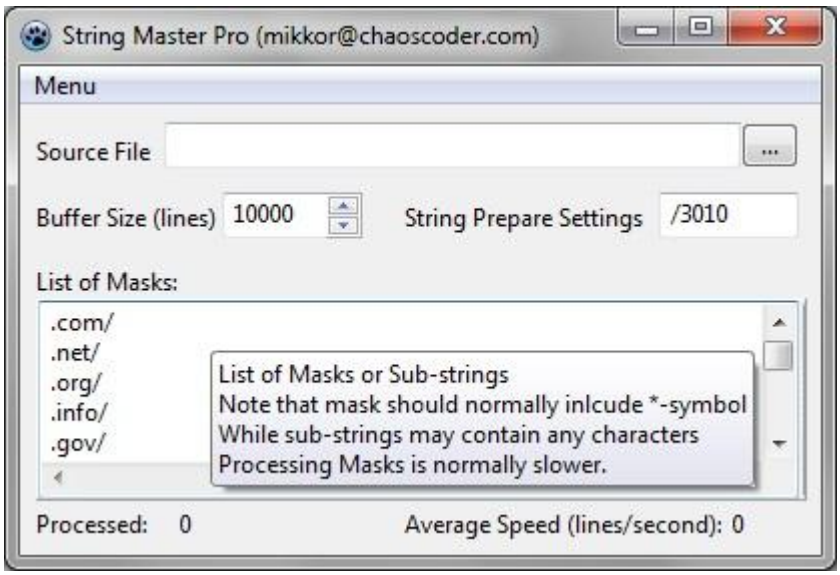
Software interface and controls

The program interface is compact and convenient. From top to bottom the String Master window contains: the menu bar; the source file string selection field; the buffer size and statistics update frequency setting field, reference setup field; Combo box with a list of masks or substring; the information line.



(Figure 3. Main window)

When you mouse over any of the controls appears a tooltip with a brief explanation.

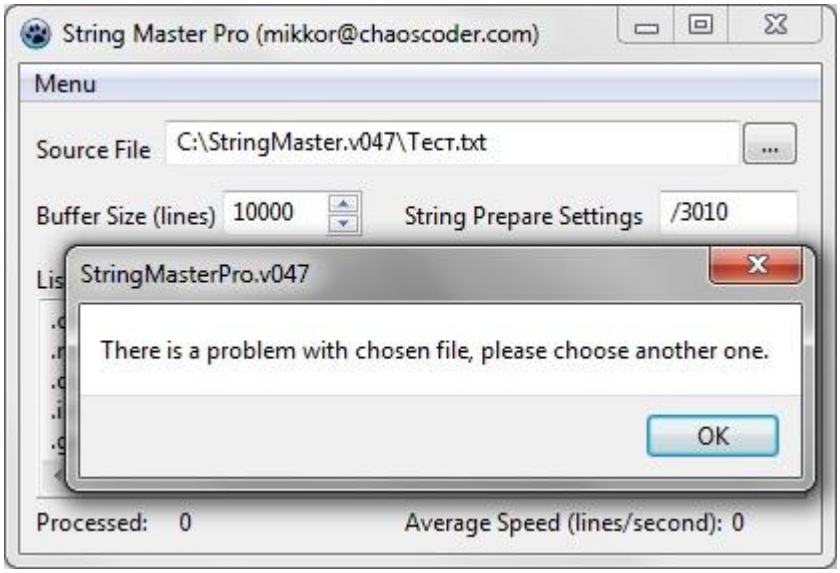


(Figure 4. Flyover)

By using **the menu bar (Menu)**, you can call a help window, select one of the file processing or statistical data collecting modes, command the use of the substring instead of a mask, use the option

of working with Unicode-coding, enable / disable the protocol of the software, and immediately close the application window . More information about the menu items can be found in the section "Software menu and file-processing modes."

In **the source file selection field (Source File)**, you can insert a link to the source file in the operating system clipboard, or select it manually, using the "..." button on the right. The file name should not contain Cyrillic characters, otherwise the program will report an error.



(Figure 5. Error while opening a file with Cyrillic characters in the name)

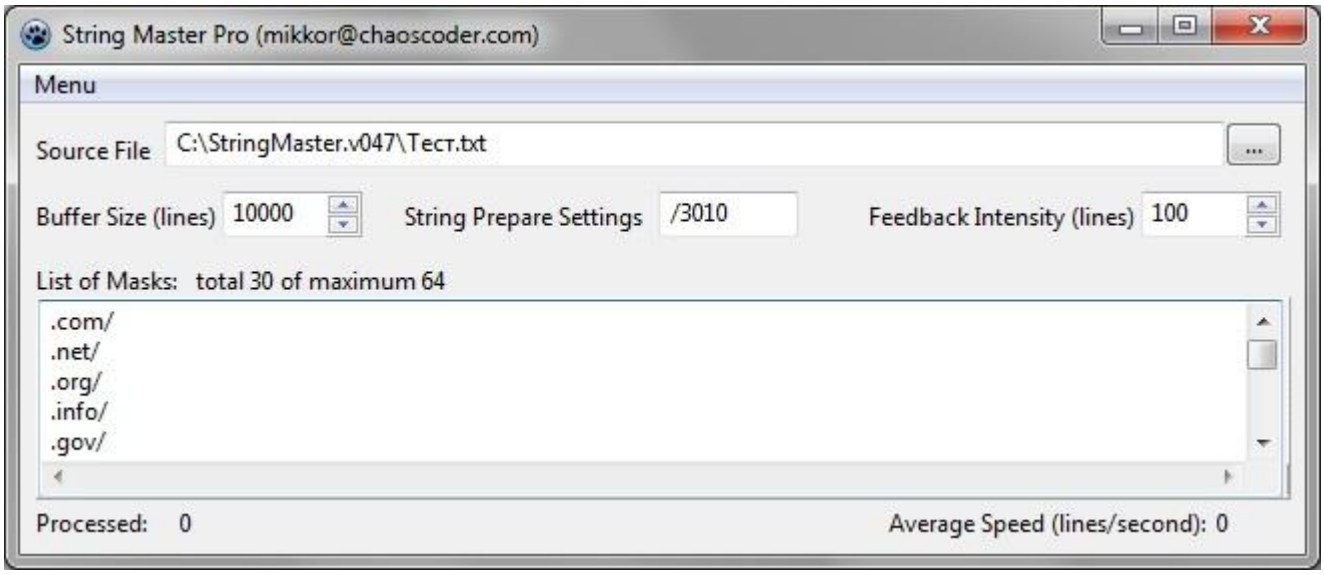
Buffer size (lines) setting field contains the number of lines which will be stored in sting-buffers during the sorting before actually being written to the files for each mask. The use of buffering helps to speed up the program, depending on your computer's performance the optimal buffer size can vary from 1,000 to 10,000 lines, the default settings are set at 10000. A further increase of this number will not provide a significant increase of processing speed. The buffer size can be changed directly during the processing of the file.

Link preparing field is used to preprocess the links (removal of transmissible parameters, the script name and paths, reducing the reference to the first folder or to the domain). More about conversion logic links - in the "Program menu and modes of processing files."

The Combo box with a list of masks or substring allows you to enter the required filters for the processing of the source file, their total number cannot exceed 32 for the free version and 256 for the paid one. By default, this list contains 30 domains (com /,. Net /,. Org /,. Info /,. Gov /,. Edu /,. Mil /,. Biz /,. Co /,. Xxx /,. tv /,. ru /,. ua /,. by /,. cz /,. bg /,. sk /,. kz /,. es /,. de /,. fr /,. uk /,. eu / ,. pl /,. ca /,. ch /,. cn /,. jp /,. it /,. kr /), the user can enter any other values. Each new mask must begin on a new line.

Statistics refresh frequency setting field in the information line. To access it the main window needs to be enlarged horizontally until the **Feedback Intensity (lines)** appears. The default value for this field is 100, in other words, the program updates the statistics every 100 processed lines of the

source file. If you reduce the value the statistics will be updated more frequently if you increase it then less frequently, but this will allow increasing the performance when working with large databases. The value of Feedback Intensity can be changed at any time as well at the start of the file processing, as during the sorting process.



(Figure 6. Field change the refresh rate statistics)

The Information lines contain the total number of processed lines, and the average speed of the program i.e. the number of lines processed per second (Average Speed (links / second)).

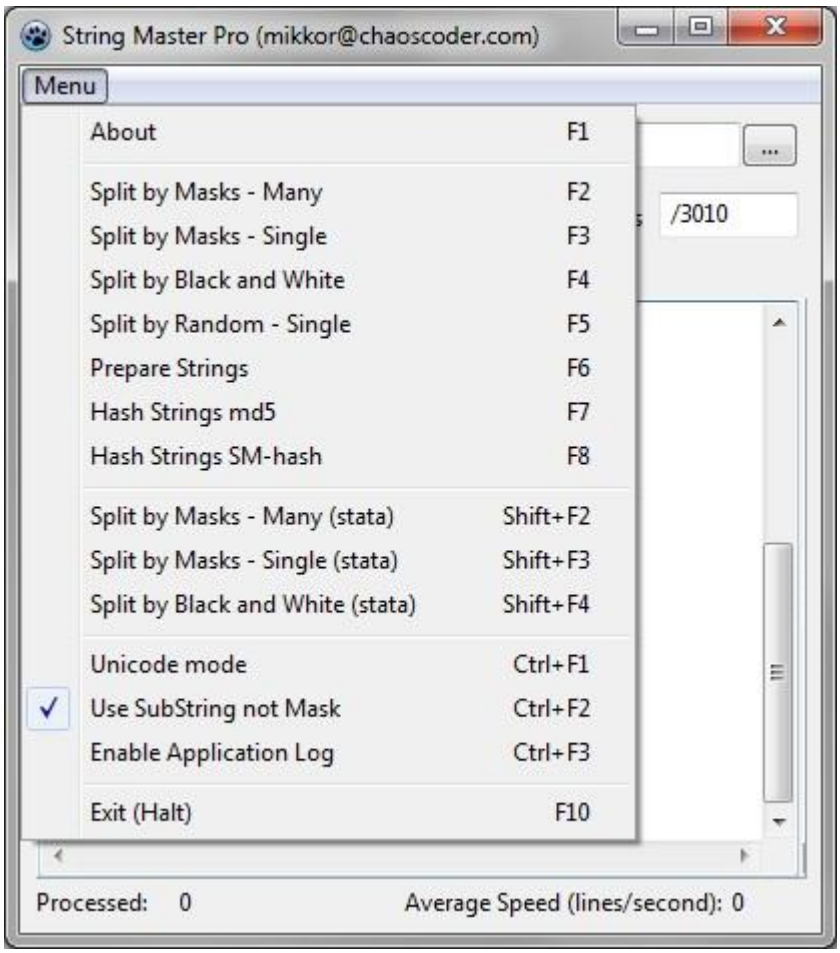
When the application closes the usual way (with the close button in the operating system or by pressing the appropriate key combination), the program creates a configuration file (ini-file), which saves most of the information entered by user : the name of the file chosen for the processing and the full path to it, the contents of the link-preparing line with the field of the list of masks, the paths and the names of the creating files with their sorting results and the value of the intensity of the survey. This file also records whether the Unicode-encoding and the use of substrings instead of masks are enabled, as well as user-defined window size. Next time you'll start String Master it will automatically read this data from the configuration file.

The option of log entering and, if necessary the size of string buffer must be set manually each time the program restarts: when you start the program the possibility of entering the log is disabled by default, the buffer size is 10,000 lines.

The configuration file can be edited manually using any text editor. To reset all settings simply delete ini-file, in this case, when you'll start String Master it will be launched with all the parameters set to default values.

Menu and file processing modes

String Master's menu is divided into several thematic categories. In addition to selecting the Help and quick closing of the application window, there are three other groups of items. The first group allows selecting one of four sorting modes, line preparation mode and one of the two modes of encryption. The second group contains three points for the collection of statistical information, and allows us to estimate the size of the resulting files and the number of occurrences of the masks, useful when working with large amounts of input data. The third group includes the Unicode-encoding options, the use of substrings instead of masks and entering the processing log by the program. Each of the menu items can be activated by a hot key or combination of keys, which makes the program simple and intuitive.



(Figure 7. String Master's menu)

About (F1) – opens the information window that contains information about the version, the release date and the e-mail of the author of the program.

After the user selects a file to process, sets all the required substrings or masks in the appropriate field, and selects one of the modes of splitting (Split by Masks - Many, Split by Masks - Single, Split by Random - Single, Split by Black and White), string preparing or encryption (Hash Strings

md5, Hash Strings SM-hash) the program starts to analyze the file line by line, checking the conformity of each component of the lines with the defined parameters and displays the information in the resulting files.

When **Split by Masks - Many (F2)** is selected, files are created according to the number of masks / substrings; each file is given a name composed of the source file's name and the sorting number of the mask or substring. When naming the base.txt source file and all the links displayed in the default substrings containing «. Com /», will be saved in a file named base.txt.1.txt; containing «. Net /» - to file base.txt .2. txt; containing «. org /» - to file base.txt.3.txt, and so on. In the case of a match the string source file will be copied to all files containing the matching masks / substrings. If in addition, if there are lines not suiting any mask or substring, then will be they will be placed in a file bearing the source name and the index 0.

Example: If two occurrences will be found in one of the lines - «. Com /» and «. Net /», then it gets in base.txt.1.txt, and base.txt.2.txt.

When using **Split by Masks - Single (F3)** the number and the names of the created files are similar, but if a line corresponding to several substrings or masks is found, then it will be placed only in the file with a mask placed higher in the list.

Example: If two occurrences are found in one of the lines - «. Com /» and «. Net /», then it gets only base.txt.1.txt, but will not get into base.txt.2.txt.

With **Split by Black and White (F4)** the program will only create two files - with index 0 and index 1. In the index 1 file will be placed the source data lines, which will coincide with at least one of the specified masks or substrings; the index 0 file will get all those for which there was no match.

Mode **Split by Random - Single (F5)** is designed to split large files into parts with automatic mixing. In this case, all the masks are ignored, and the program creates 32 files: each line of the source file will have an equal probability to get into one of them.

By selecting **Prepare Strings (F6)**, the user can give the single required form to a list of links (the so-called pre-processing). In earlier software releases, this option was used in conjunction with the sorting modes allowing to save a lot of time, but it also could lead to the loss of links (for example, when reducing the links to the sign "?" in the base, all the lines that do not contain this character get in a file with a zero index, and can be lost when combined with other preparing base with other processing modes). Pre-processing used as a separate procedure allows avoiding such errors. Also it's good to see an unfinished version before creating an encrypted data file.

All masks and substrings in this mode are ignored by the program, a predefined sequence of characters is used in the field **URL Prepare Settings**, where:

- First character - necessary to find the characters in the file, which will be the string splitting place;
- Second character - counter of entries of the specified character;

- Third character - is 0 or 1. If it is 0, then the entry of the character is in the beginning of the string. If 1, the end;
- Fourth mark – can be 0 or 1. If it is 0, then all the data in the string from the beginning to the entry character is deleted. If 1, then from the character to the end;
- Fifth character - is 0 or 1. If it is 0, then the entry character is included in the line of the resulting file. If 1, then excluded.

Example 1: If in the URL Prepare Settings the value is / 3010, then links likehttp://ru.someforum.com/index.php?showtopic=10358&st=40&#entry136670 will be reduced to the form http://ru.someforum.com, because:

- The character «/» is searched;
- The third entry of the character is searched;
- The line is checked from the beginning;
- all data after the entry character is deleted;
- The entry character itself is not included in the output file.

Example 2: If in the URL Prepare Settings the value is .3110, then links like http://music.bla.bla.ru.someforum.com will be converted to ru.someforum.com, because:

- The character «.» is searched;
- The first entry of the character is searched;
- The line is checked from the end;
- all data after the entry character is deleted (When checking from right to left, this is the same part of the link «http://music.bla.bla»)
- The entry character itself is not included in the output file.

In this case, there is a clearing of links from sub-domains.

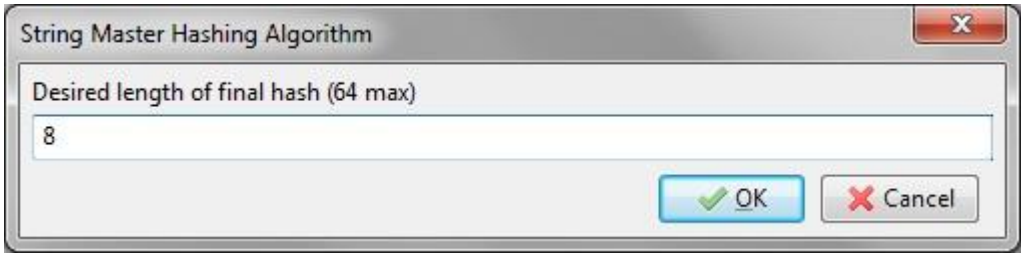
Example 3: If in the URL Prepare Settings billed value? 1010, a link type http://ru.someforum.com/index.php?showtopic=10358&st=40&#entry136670 take the form http://ru.someforum.com/index.php, because:

- The character «?» is searched;
- The first entry of the character is searched;
- The line is checked from the beginning;
- all data after the entry character is deleted
- The entry character itself is not included in the output file.

This treatment removes the additional parameters of the script.

With the **Hash Strings md5 (F7)** the user can encrypt each line of the algorithm md5. This program creates a single file, where all the hashes of the source file lines will be placed. This is useful when multiple users want to compare their line bases, without revealing to each other their direct content, as well as for selling the bases.

Encryption mode **Hash Strings SM-hash (F8)**, in contrast to previous mode it allows to specify the hash length. Since the using the full length MD5 in the processing of small files is impractical, reducing the length of the hash will in many cases increase the speed of the program and the size of the output file containing the encrypted string values. Using a longer hash is recommended when working with large amounts of information, as this will help to avoid potential conflicts.

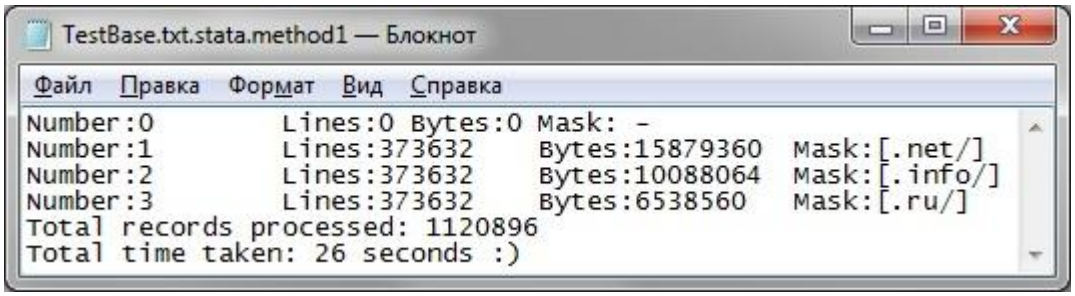


(Figure 8. Specifying the length of the hash for encryption)

Menu. Modes to obtain statistics and options

Besides the basic ways of working with a file in String Master the user can also enjoy several additional statistical modes. To select them, the user can use the menu items **Split by Masks - Many (stata) (Shift+F2), Split by Masks - Single (stata) (Shift+F3), Split by Black and White (stata) (Shift+F4)**.

These modes are designed to determine the size of the output files and the number of entries according to the user defined parameters. The principle of operation is similar to sorting algorithms, but the output files contain statistics instead of the appropriate to the masks strings, which makes things much faster, as the performance in this case is only limited by the speed of line by line reading. Name of the output file is composed of the name of the original file with the addition of «stata» and the selected mode of statistics. For example, if the file's name is base.txt under Split by Masks mode - Many (stata) the resulting file will be named base.txt.stata.method1. In the statistics file apart the counted masks are entered the data on the total number of lines in the source file and the time spent on analysis. The ability to collect statistics is extremely useful when working with large databases of links.



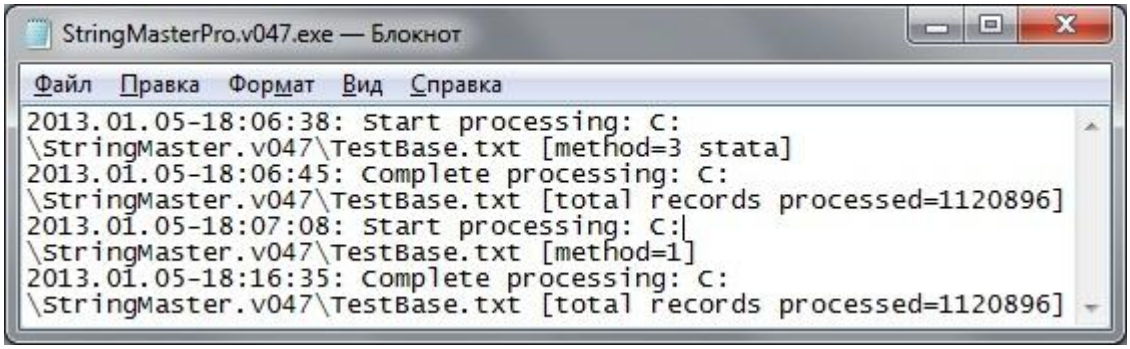
(Figure 9. Example of output file in stata mode)

Regardless of the mode selected to work with the files the user can choose additional options: the use of Unicode-encoding, working with substrings instead of masks (enabled by default), enabling the protocol work by the program.

Unicode mode (Ctrl + F1) option can process files with UTF-8 coding (Unicode Transformation Format, compatible with 8-bit encoding of the text) under an enabled flag and ANSI coding if disabled. Thus, with String Master there is no need to use a third-party program to save the text in a proper encoding.

With the option **Use Substring not Mask (Ctrl + F2)**, the user can use only the substrings for the sorting, in this case the use of «*» is relevant only if it can occur in a source file text. Use Substring not Mask option is enabled by default, if it's disabled the masks mode is activated, in which the character «*» has a wider application.

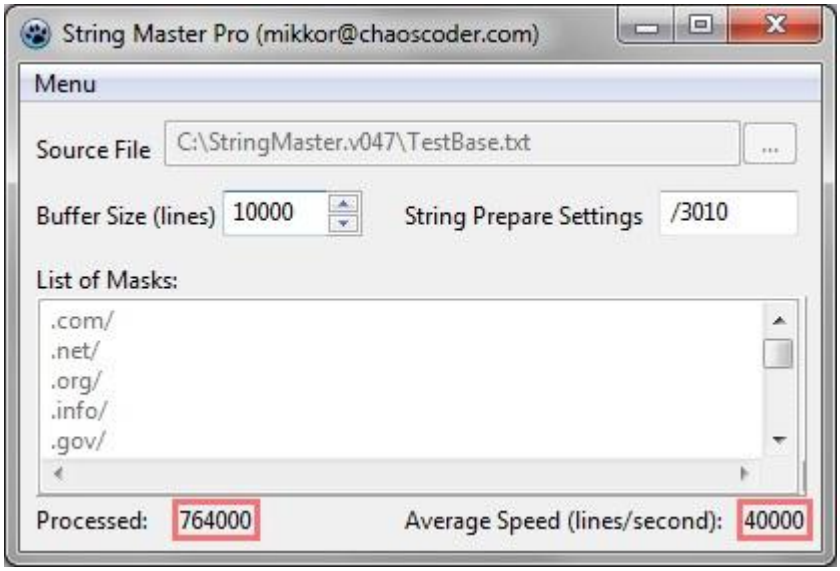
Enable Application Log option allows the application to keep a record protocol of work, which will contain information about the date and time of commencement and completion of the selected files, the sorting and encryption methods, as well as the file name and its full path.



(Figure 10. program working Minutes)

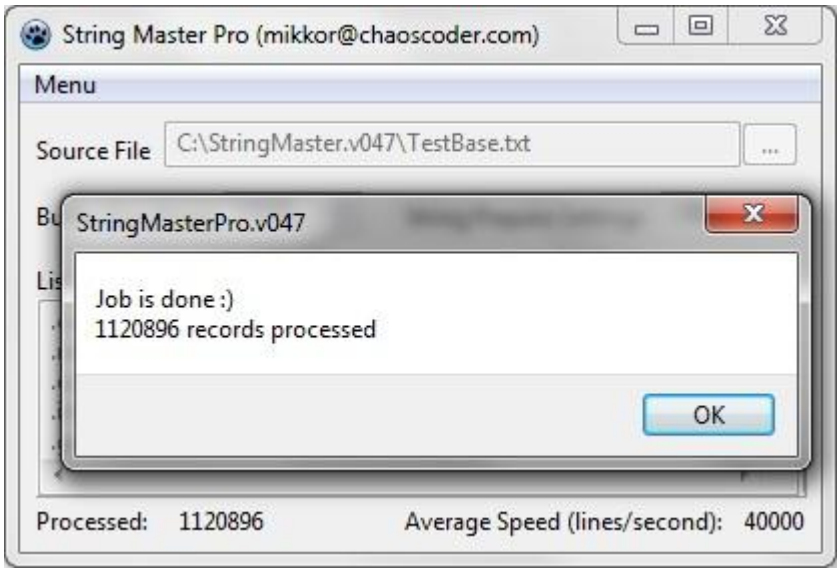
Work programs and advanced options

When selecting the desired processing mode from the menu or pressing the corresponding hotkey (shortcut key), the program will start and the status bar will display the progress of the job and line processing speed.



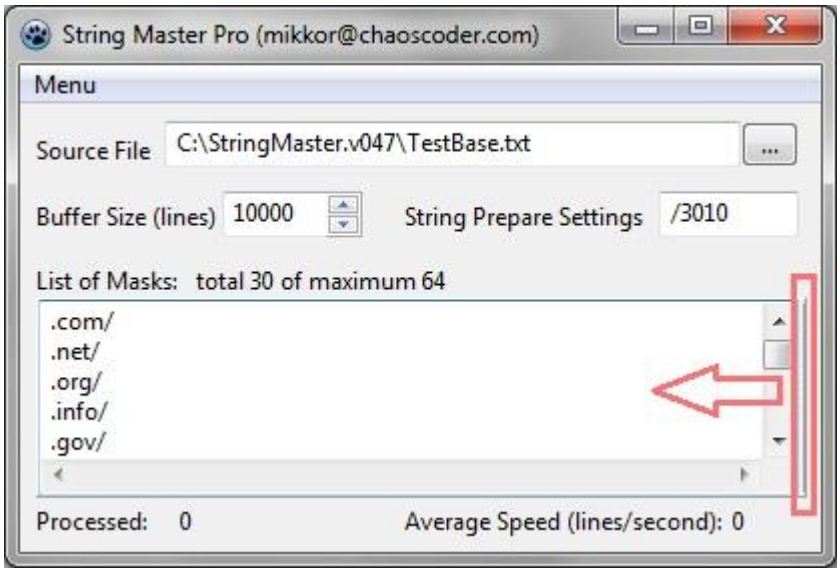
(Figure 11. Progress of a task)

After the sorting, the user will be shown a report window, indicating that the work has been done.



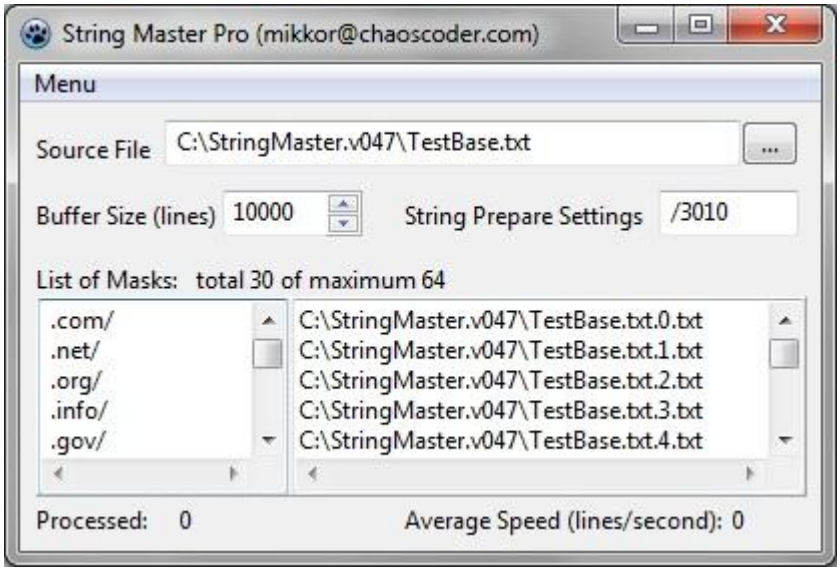
(Figure 12. Completion report of the program)

If the user thinks that file names composed of the source file name and the digital index, are inconvenient, he can configure all the names manually. To do this, he must move the right side of the scroll area of the combo box with the list of masks to the left.



(Figure 13. Display of the file names configuration field)

Changing names is useful when working with large files, because by changing the address you can record files on a different physical drive. When you first start using a filename field, it is empty; the values in it appear when you select a source file in the Source File field. When closing the program the standard way, the file names specified by the user are stored in the configuration file, but when a new database is chosen for the sorting with the «…» button they are reset to the default values.



(Figure 14. Kind field settings file names)

Important features of the program

Among the important features of the program are the following: the secondary processing of the same file does not overwrite the output files but appends them. Therefore, when working with large files, the processing of which is time consuming, do not forget about it.

At each start of the program the buffer value is set to 10,000 lines, the logging option is off. If the user needs to keep a record or use the value of the string buffer different from the default, each time at the start of the program, the appropriate settings will have to be reset. All other changes are saved in the ini-file are read automatically at the start of String Master.

Menu item **Exit (Halt)** or pressing **F10** closes the program immediately, herewith the recording of the current data from the buffer to the output file and the configuration file settings are not performed. In the event of any failure or the need to close the program the value of the buffer should be reduced. The possibility of an emergency exit can be applied when the program "freezes", in the rest of the time String Master should be closed in a standard way.

Conclusion

Of course, String Master is not only suitable for working with lists of links, and a variety of databases of forums and websites. Through a stable and fast work with large files this program can have lots of other uses.