

TMG Version 6.0R7 Release Notes

1 Additions

1.1 Filtering Non-ASCII Documents

Version 6.0R7 improves the list of convertible to ascii texts documents. Current version supports parsing the non ascii documents of table 1

Type	ver.5.0R6	Filter ver.5.0R6	ver. 6.0R7	ver. 6.0R7
doc	×	×	✓	TIKA
docx	×	×	✓	TIKA
htm	✓	strip.html	✓	strip.html
html	✓	strip.html	✓	TIKA
odt	×	×	✓	TIKA
pdf	✓	ps2ascii	✓	ps2ascii
ps	✓	ps2ascii	✓	ps2ascii
rtf	×	×	✓	TIKA
tex	×	×	✓	Untex

Table 1: Supported non-ascii formats

TIKA™ toolkit: a Java API for detection and extraction of metadata and structured text from non-ascii documents¹

Untex: program to strip some Latex commands from the source file. Differs depending on operating system used; for Microsoft Windows the executable is `untex` package². Linux users should install the `untex` package to their systems³.

1.2 Management of Numeric and Alphanumeric Entries

Management of Numeric Entries: Version 6.0R7 supports the non-automatic removal of numeric entries; see Table 2

Management of Alphanumeric Entries: Version 6.0R7 allows users to specify whether alphanumeric strings are collected in the dictionary; see Table 3

¹<http://tika.apache.org/>

²<http://www.ctan.org/pkg/untex>

³<http://linuxappfinder.com/package/untex>

Version	Support
5.0R6	Automatic Removal of Numeric Entries
6.0R7	The user selects whether to include numeric strings in the dictionary by ticking the corresponding checkbox of the Indexing GUI

Table 2: Table of Numeric Entries Support

Version	Support
5.0R6	No support
6.0R7	The user selects whether to include alphanumeric strings in the dictionary by ticking the corresponding checkbox of the Indexing GUI

Table 3: Table of Alphanumeric Entries Support

1.3 Directories and Log Files

Directories and Log Files: New directories has been included in this version, for better manipulation of filtered datas and running results. Descriptions of these directories are summarized in Table 4

Log Files: Log Files are the files where statistics from different operations in TMG are saved. These files are summarized in Table 5

1.4 Management of Subdirectories

Subdirectories: Previous versions supported automatic parsing of subdirectories, current version supports selective parsing of subdirectories as specified by the user referred in Table 6

1.5 Check for Updates

Users of TMG are able to check for new versions released. This choice has been added to `about_tmog` GUI and requires connection to the internet. If a new version has been released then the user is notified and is directed to the download page. On the other hand, if there is no new version available, the user will be notified with an appropriate message.

2 Important Indications

2.1 Running `init_tmog`

This specific function should run in the two following two cases.

Directory	Description
Directory [TEXT_RESULTS]	Is created during the parsing of a single file, directory or directories when a term-by-document matrix is created. Each ascii file which can be parsed, can be found inside this specific directory. This directory can exist in the root directory and to include parsed texts from previous runs. In this case [TEXT_RESULTS] is not deleted (though all its previous contents are deleted). This particular job is implemented by the function <code>cleanup</code> .
Directory [TEXT_RESULTS_U]	Is created during the parsing of a single file, directory or directories when an existing term-by-document matrix is updated. Each ascii file which can be parsed, can be found inside this specific directory. This directory can exist in the root directory and to include parsed texts from previous runs. In this case [TEXT_RESULTS_U] is not deleted (though all its previous contents are deleted). This particular job is implemented by the function <code>cleanup</code> .
Directory [TEXT_RESULTS_Q]	Is created during the parsing of a single file, directory or directories when a Query matrix is created. Each ascii file which can be parsed, can be found inside this specific directory. This directory can exist in the root directory and to include parsed texts from previous runs. In this case [TEXT_RESULTS_Q] is not deleted (though all its previous contents are deleted). This particular job is implemented by the function <code>cleanup</code> .
Directory <code>log_files</code>	Exists in the TMG root directory; it is where the log files <code>tmg.log</code> , <code>history.log</code> as well as the new log file <code>filters.result.log</code> are saved.

Table 4: Table of ³New Directories

Version	Files	Description
5.0R6	tmg.log	tmg.log is used to store a history of the operations performed at different times
	history.log	history.log stores all different paths followed earlier when opening files
6.0R7	tmg.log history.log filters_results.log	filters_results.log is used to store a history of converting into ascci,non ascci documents

Table 5: Table of log files

Version	Support
5.0R6	automatic parsing of all subdirectories
6.0R7	The user can select whether to parse a detected subdirectory. A message will appear at the command prompt. Recommended for small collections because it requires respective user interaction
	The user can select to parse all subdirectories without being questioned by ticking the corresponding checkbox of the Indexing GUI. It is recommended for large collections so that they can be run in batch mode.

Table 6: Table of Subdirectories parsing Support

- In the very first use of TMG the user must run function `init_tmg`. This will cause the addition of the necessary directories required by TMG into the current Matlab path, for users of MySql it would also establish the connection between the database and TMG.
- Users of MySql should run function `init_tmg` every time they start TMG, in order to establish the connection with the database.

2.2 Creating A New Stopwords List

Any user of TMG has the ability to create his own `stopwords list`. In that case the user should take to eliminate any trailing spaces at the end of each stopword because, currently, trailing spaces are considered part of the word which can be misleading.

2.3 Parsing Unsupported File Types

When an unknown file type is going to be parsed, TMG ignores it, indicates it as unsupported type and continues parsing procedure to the next file. Though the automatic recognition of unsupported files, it is recommended the collections to consist of supported file types in its majority. This will help to avoid long duration of parsing procedure.