

Corpus Sources

Corpora	Description	Source
Corpus	A general purpose wrapper	Installed with tm
VCorpus	<i>A volatile</i> corpus	Installed with tm
SimpleCorpus	A corpus optimized for text documents	Installed with tm
PCorpus	<i>A persistent</i> corpus	Installed with tm
DCorpus	A corpus distributed across Hadoop or LFS	Installed with tm.plugin.dc

Corpus Sources

Source Type	Description	Source	Compatible with VCorpus and PCorpus	Compatible with SimpleCorpus
DataframeSource	Read a data.frame where each row represents a document.	Part of the tm package	yes	yes
DirSource	Read a directory containing one or more files.	Part of the tm package	yes	yes
URISource	Read a list of URLs pointing to text documents.	Part of the tm package	yes	no
VectorSource	Read a vector of document contents.	Part of the tm package	yes	yes
XMLSource	Read a list of XML documents.	Part of the tm package	yes	no
ZipSource	Read a zip file with a collection of documents.	Part of the tm package	yes	no

Corpus Sources

Source Type	Description	Source	Compatible with VCorpus and PCorpus	Compatible with SimpleCorpus
MBoxSource	Read a mailbox.	tm.plugin.mail	yes	no
AlcesteSource	Read files used by the Alceste application	tm.plugin.alceste	yes	no
FactivaSource	Read files used by Dow Jones 'Factiva'	tm.plugin.factiva	yes	no
LexisNexisSource	Import articles from LexisNexis	tm.plugin.lexisnexis	yes	no
EuropresseSource	Import articles from Europresse	tm.plugin.europresse	yes	no