Frequently Asked Questions

This document contains answers to some of the most frequently asked questions about tm.

---

1. *How should I cite tm?*

   Please have a look at the output of `citation("tm")` in R. A BibTeX representation can be obtained via `toBibtex(citation("tm"))`.

   The preferred way for journal and conference papers is to cite the JSS article.

2. *I want to read in a PDF file using the `readPDF` reader. However, the manual says I need the tool `pdftotext` installed and accessable on my system. Where can I find and how can I install this tool?*

   Many linux distributions provide pre-built packages: `poppler-utils`, `xpdf-utils`, or similar. Windows users need to download and install Xpdf. Ensure that the program is included in your PATH variable.

   Windows users might find a R-help thread on this topic useful.

3. *My documents are stored in file format XYZ. How do I get the material into tm and construct a corpus from it?*

   Please have a look at the vignette Extensions: How to Handle Custom File Formats.

4. *What about error messages indicating invalid multibyte strings?*

   Ensure that all your datasets and documents are encoded in UTF-8. If you still have problems `tm_map(yourCorpus, content_transformer(function(x) iconv(enc2utf8(x), sub = "byte")))` will replace non-convertible bytes in `yourCorpus` with strings showing their hex codes.

5. *Can I use bigrams instead of single tokens in a term-document matrix?*

   Yes. Package NLP provides functionality to compute n-grams which can be used to construct a corresponding tokenizer. E.g.:

```
library("tm")
data("crude")

BigramTokenizer <-
function(x)
  unlist(lapply(ngrams(words(x), 2), paste, collapse = " "), use.names = FALSE)

tdm <- TermDocumentMatrix(crude, control = list(tokenize = BigramTokenizer))
inspect(removeSparseTerms(tdm[, 1:10], 0.7))
```

6. *How can I plot a term-document matrix like Figure 6 in the JSS article on tm?*

   Please check the manual accessible via `?plot.TermDocumentMatrix` for available arguments to the plot function. A plot similar to Figure 6 can be produced e.g. with:

```
library("tm")

data("crude")

tdm <- TermDocumentMatrix(crude, control = list(removePunctuation = TRUE,
                                                removeNumbers = TRUE,
                                                stopwords = TRUE))

plot(tdm, terms = findFreqTerms(tdm, lowfreq = 6)[1:25], corThreshold = 0.5)
```