qdap-tm Package Compatibility

Tyler W. Rinker

May 15, 2018

The qdap package (Rinker, 2013) is an R package designed to assist in quantitative discourse analysis. The package stands as a bridge between qualitative transcripts of dialogue and statistical analysis and visualization. The tm package (Feinerer and Hornik, 2014) is a major R (R Core Team, 2013) package used for a variety of text mining tasks. Many text analysis packages have been built around the tm package’s infrastructure (see CRAN Task View: Natural Language Processing). As qdap aims to act as a bridge to other R text mining analyses it is important that qdap provides a means of moving between the various qdap and tm data types.

This vignette serves as a guide towards navigating between the qdap and tm packages. Specifically, the two goals of this vignette are to (1) describe the various data formats of the two packages and (2) demonstrate the use of qdap functions that enable the user to move seamlessly between the two packages.
1 Data Formats

The `qdap` and `tm` packages each have two basic data formats. `qdap` stores raw text data in the form of a `data.frame` augmented with columns of demographic variables whereas `tm` stores raw text as a Corpus and annotates demographic information with Meta Data attributes. The structures are both lists and are comparable.

The second format both packages use is a matrix structure of word frequency counts. The `qdap` package utilizes the `Word Frequency Matrix` (wfm function) whereas the `tm` package utilizes the `Term Document Matrix` or `Document Term Matrix` (`TermDocumentMatrix` and `DocumentTermMatrix` functions). Again the structure is similar between these two data forms. Table 1 lays out the data forms of the two packages.

<table>
<thead>
<tr>
<th>Package</th>
<th>Raw Text</th>
<th>Word Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>qdap</td>
<td>Dataframe</td>
<td>Word Frequency Matrix</td>
</tr>
<tr>
<td>tm</td>
<td>Corpus</td>
<td>Term Document Matrix/Document Term matrix</td>
</tr>
</tbody>
</table>

Table 1: qdap-tm Data forms

Figure 1 provides a visual overview of the `qdap` functions used to convert between data structures. Many of these conversion could be achieved via the `tm` package as well.

*Note: as.tdm & as.dtm are short hand for as.TermDocumentMatrix & as.DocumentTermMatrix*

Figure 1: Converting Data between qdap and tm
One of the most visible differences between qdap-tm data forms is that qdap enables the user to readily view the data while the tm utilizes a print method that provides a summary of the data. The tm::inspect function enables the user to view tm data forms. The qdap package provides qdap::qview and qdap::htruncdf functions to view more digestible amounts of the data. Let’s have a look at the different data types. We’ll start by loading both packages:

```r
library(qdap); library(tm)
```

Now let us have a look at the raw text storage of both packages.

### 1.1 Raw Text

#### 1.1.1 qdap’s Raw Text

**DATA**

**qview**

**htruncdf**

```r
# > DATA
##
## # > qview(DATA)
##
## # > htruncdf(DATA)
```
## 1 sam m 0 Computer i K1
## 2 greg m 0 No it's no K2
## .
## .
## .
## 8 sam m 0 I distrust K8
## 9 sally f 0 What are y K9
## 10 researcher f 1 Shall we m K10

```r
> htruncdf(DATA)
```

```r
## person sex adult state code
## 1 sam m 0 Computer i K1
## 2 greg m 0 No it's no K2
## .
## .
## .
## 8 sam m 0 I distrust K8
## 9 sally f 0 What are y K9
## 10 researcher f 1 Shall we m K10
```

### 1.1.2 tm’s Raw Text

data("crude")
crude
inspect(crude)

```r
> crude
> # A corpus with 20 text documents
> #
> # > crude[[1]]
> # Diamond Shamrock Corp said that
> # effective today it had cut its contract prices for crude oil by
> # 1.50 dlrs a barrel.
> # The reduction brings its posted price for West Texas
```
" Intermediate to 16.00 dlrs a barrel, the company said.
" The price reduction today was made in the light of falling
" .
" .
" .
" Diamond is the latest in a line of U.S. oil companies that
" have cut its contract, or posted, prices over the last two days
" citing weak oil markets.
" Reuter

### 1.2 Word/Term Frequency Counts

Now we’ll look at how the two packages handle word frequency counts. We’ll start by setting up the raw text forms the two packages expect.

```r
tm_dat <- qdap_dat <- DATA[1:4, c(1, 4)]
tm_dat[['doc_id']] <- paste("docs", 1:nrow(tm_dat))
colnames(tm_dat)[2] <- c('text')
tm_dat <- Corpus(DataframeSource(tm_dat))
```

Both `qdap_dat` and `tm_dat` are storing this basic information:

```r
## person state
## 1 sam Computer is fun. Not too fun.
## 2 greg No it's not, it's dumb.
## 3 teacher What should we do?
## 4 sam You liar, it stinks!
```
### 1.2.1 qdap's Frequency Counts

```r
code

with(qdap_dat, wfm(state, person))

# greg sam teacher
# dumb 1 0 0
# it 2 1 0
# no 1 0 0
# not 1 1 0
# s 2 0 0
# computer 0 1 0
# fun 0 2 0
# is 0 1 0
# liar 0 1 0
# stinks 0 1 0
# too 0 1 0
# you 0 1 0
# do 0 0 1
# should 0 0 1
# we 0 0 1
# what 0 0 1
```

### 1.2.2 tm’s Frequency Counts

```r
code

TermDocumentMatrix(tm_dat, control = list(
  removePunctuation = TRUE,
  wordLengths=c(0, Inf)
))

# TermDocumentMatrix (terms: 16, documents: 4)
# Non-/sparse entries: 18/46
# Sparsity : 72%
# Maximal term length: 8
# Weighting : term frequency (tf)
```
Now we’ll look at the `tm` output using `inspect`.

```r
inspect(TermDocumentMatrix(tm_dat,
    control = list(
        removePunctuation = TRUE,
        wordLengths=c(0, Inf)
    )
))
```

```
# Docs
# Terms 1 2 3 4
# computer 1 0 0 0
# do 0 0 1 0
# dumb 0 1 0 0
# fun 2 0 0 0
# is 1 0 0 0
# it 0 0 0 1
# its 0 2 0 0
# liar 0 0 0 1
# no 0 1 0 0
# not 1 1 0 0
# should 0 0 1 0
# stinks 0 0 0 1
# too 1 0 0 0
# we 0 0 1 0
# what 0 0 1 0
# you 0 0 0 1
```

The two matrices are essentially the same, with the exception of column order and names. Notice that by default `tm` removes words with fewer characters (word length) and does not discard punctuation (we made the matrices equal by specifying `removePunctuation = TRUE` and `wordLengths=c(0, Inf)` for `tm`’s control argument). `qdap` takes the opposite approach, removing punctuation and utilizing all words, by default. Likewise, the `tm` package stores demographic information as meta data within the `Corpus`, whereas, `qdap` incorporates the demographics with the text into a single `data.frame` structure. These differences arise out of the intended uses, audiences, and philosophies of the package authors. Each has strengths in particular situations. The `qdap` output is an ordinary `matrix` whereas the `tm` output is a more compact
simple_triplet_matrix. While the storage is different, both packages can be made to mimic the default of the other.

Also note that the \texttt{qdap} \texttt{summary} method for \texttt{wfm} provides the user with information similar to the \texttt{TermDocumentMatrix}/\texttt{DocumentTermMatrix} functions' default print method.

\begin{verbatim}
summary(with(qdap_dat, wfm(state, person)))
\end{verbatim}

\begin{verbatim}
## <<A word-frequency matrix (16 terms, 3 groups)>>
##
## Non-/sparse entries : 18/30
## Sparsity : 62%
## Maximal term length : 8
## Less than four characters : 62%
## Hapax legomenon : 12(75%)
## Dis legomenon : 3(19%)
## Shannon's diversity index : 2.69
\end{verbatim}

Now we'll look at some \texttt{qdap} functions that enable the user to move between packages, gaining the flexibility and benefits of both packages.

## 2 Converting Data Forms

We'll again use the following preset data:

\begin{verbatim}
    tm_dat <- qdap_dat <- DATA[1:4, c(1, 4)]
    tm_dat[,'doc_id'] <- paste("docs", 1:nrow(tm_dat))
    colnames(tm_dat)[2] <- c('text')
    tm_dat <- Corpus(DataframeSource(tm_dat))

    qdap_wfm <- with(qdap_dat, wfm(state, person))
    tm_tdm <- TermDocumentMatrix(tm_dat,
        control = list(
            removePunctuation = TRUE,
            wordLengths= c(0, Inf)
        )
    )
\end{verbatim}
1. `qdap_dat` – is a **qdap** raw text form

2. `tm_dat` – is a **tm** raw text format

3. `qdap_wfm` – is a **qdap** word frequencies count

4. `tm_tdm` – is a **tm** word frequencies count

The reader is encouraged to view each of the data formats:

```r
qdap_dat; qview(qdap_dat)
tm_dat; inspect(tm_dat)
qdap_wfm; summary(qdap_wfm)
tm_tdm; inspect(tm_tdm)
```

### 2.1 Corpus to data.frame

To move from a Corpus to a `data.frame` the `as.data.frame` function is used as follows:

```r
as.data.frame(tm_dat)
```

```r
doc_id text person
1  docs 1 Computer is fun. Not too fun. sam
2  docs 2 No it’s not, it’s dumb. greg
3  docs 3 What should we do? teacher
4  docs 4 You liar, it stinks! sam
```

### 2.2 data.frame to Corpus

To move from a `data.frame` to a Corpus the `as.Corpus` function is used as follows:

```r
with(qdap_dat, as.Corpus(state, person))
```

```r
<<SimpleCorpus>>
# Metadata: corpus specific: 1, document level (indexed): 3
# Content: documents: 3
```

*Note the 3 text documents; one for each grouping variable. To get one for each row use:

```r
with(qdap_dat, as.Corpus(state, id(person)))
```
2.3 TermDocumentMatrix/DocumentTermMatrix to wfm

To move from a TermDocumentMatrix to a wfm the as.wfm function is used as follows:

```r
as.wfm(tm_tdm)
```

<table>
<thead>
<tr>
<th></th>
<th>docs 1</th>
<th>docs 2</th>
<th>docs 3</th>
<th>docs 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fun</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>is</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>too</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dumb</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>it</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>no</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>do</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>should</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>we</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>what</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>liar</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>stinks</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>you</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

2.4 wfm to TermDocumentMatrix/DocumentTermMatrix

To move from a wfm to a TermDocumentMatrix or DocumentTermMatrix the as.tdm and as.dtm functions can be used as follows:

```r
as.tdm(qdap_wfm)
```

```r
as.dtm(qdap_wfm)
```

```
<<TermDocumentMatrix (terms: 15, documents: 3)>>
Non-/sparse entries: 17/28
Sparsity : 62%
Maximal term length: 8
Weighting : term frequency (tf)
```
2.5 Corpus to wfm

One can also move directly from a \texttt{tm} Corpus to a \texttt{qdap} \texttt{wfm} with the \texttt{as.wfm} function.

\begin{verbatim}
as.wfm(tm_dat)
\end{verbatim}

\begin{verbatim}
## all
## computer 1
## do 1
## dumb 1
## fun 2
## is 1
## it 3
## liar 1
## no 1
## not 2
## s 2
## should 1
## stinks 1
## too 1
## we 1
## what 1
## you 1
\end{verbatim}

3 Stemming, Stopwords, and Choosing n-Character Words/Terms from a wfm

Many of the \texttt{qdap} and \texttt{tm} functions have means of stemming, removing stopwords, and bounding, that is filtering rows (greater than, equal to or less than) meeting min/max criteria. \texttt{qdap} also offers two external functions to address these issues directly.
### 3.1 stemming

*qdap* takes the approach that the user stems the dataframe upon creation (using `sentSplit(..., stem = TRUE)`) or after (using the `stem2df` function), maintaining a column of stemmed and un-stemmed text for various analyses.

```r
sentSplit(qdap_dat, "state", stem = TRUE)
```

<table>
<thead>
<tr>
<th>#</th>
<th>person</th>
<th>tot</th>
<th>state</th>
<th>stem.text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sam</td>
<td>1.1</td>
<td>Computer is fun.</td>
<td>Comput is fun.</td>
</tr>
<tr>
<td>2</td>
<td>sam</td>
<td>1.2</td>
<td>Not too fun.</td>
<td>Not too fun.</td>
</tr>
<tr>
<td>3</td>
<td>greg</td>
<td>2.1</td>
<td>No it's not, it's dumb.</td>
<td>No it not it dumb.</td>
</tr>
<tr>
<td>4</td>
<td>teacher</td>
<td>3.1</td>
<td>What should we do?</td>
<td>What should we do?</td>
</tr>
<tr>
<td>5</td>
<td>sam</td>
<td>4.1</td>
<td>You liar, it stinks!</td>
<td>You liar it stink!</td>
</tr>
</tbody>
</table>

### 3.2 Filtering: Stopwords and Bounding

*qdap*'s `Filter` function allows the user to remove stopwords and bound a Word Frequency Matrix (*wfm*). First we'll construct a minimal Word Frequency Matrix:

```r
qdap_wfm <- with(qdap_dat, wfm(state, person))
```

<table>
<thead>
<tr>
<th>#</th>
<th>greg</th>
<th>sam</th>
<th>teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>dumb</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>it</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>no</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>s</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>computer</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>fun</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>is</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>liar</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>stinks</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>too</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>you</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>do</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Now we’ll move through a series of examples demonstrating the usage of `Filter` on a `wfm` object.

```r
Filter(qdap_wfm, min = 5)
```

```
## greg sam teacher
## computer  0 1 0
## stinks    0 1 0
## should    0 0 1
```

```r
Filter(qdap_wfm, min = 5, max = 7)
```

```
## greg sam teacher
## stinks    0 1 0
## should    0 0 1
```

```r
Filter(qdap_wfm, 4, 4)
```

```
## greg sam teacher
## dumb      1 0 0
## liar      0 1 0
## what      0 0 1
```

```r
Filter(qdap_wfm, 4, 4, count.apostrophe = FALSE)
```

```
## greg sam teacher
## dumb      1 0 0
## liar      0 1 0
## what      0 0 1
```
4 Apply Functions Intended for TermDocumentMatrix to wfm Object

At times it is convenient to apply a function intended for a \texttt{tm} TermDocumentMatrix or DocumentTermMatrix directly to a \texttt{qdap} wfm object. \texttt{qdap}'s apply\_as\_tm function enables these functions to be used directly on a wfm.

4.1 A Minimal wfm Object

Let us begin with a slightly larger wfm minimal example:

\begin{verbatim}
a <- with(DATA, wfm(state, list(sex, adult)))
\end{verbatim}

\begin{verbatim}
## <<A word-frequency matrix (41 terms, 4 groups)>>
##
## Non-/sparse entries : 45/119
\end{verbatim}
4.2 A Small Demonstration

Here we will use the `tm` package’s `removeSparseTerms` to remove sparse terms from a `wfm` object and return a Word Frequency Matrix object (`wfm` class).

```r
out <- apply_as_tm(a, tm::removeSparseTerms, sparse=0.6)
```

```r
summary(out)
```

```r
## <<A word-frequency matrix (3 terms, 4 groups)>>
##
## Non-/sparse entries : 7/5
## Sparsity : 42%
## Maximal term length : 4
## Less than four characters : 67%
## Hapax legomenon : 0(0%)
## Dis legomenon : 1(33%)
## Shannon's diversity index : 1.06
```

```r
class(out)
```

```r
## [1] "wfm" "true.matrix" "matrix"
```

4.3 Further Examples to Try

Here are some further examples to try:
5 Apply Functions Intended for qdap Dataframes to tm Corpus

While the tm package (and other packages used on tm objects) tends to conduct analysis by feeding functions a TermDocumentMatrix or DocumentTermMatrix qdap generally feeds functions raw text directly. There are advantages to both approaches (e.g., the matrix is a mathematical structure while raw text maintains word order). Many qdap functions can be used on the Corpus structure via the apply_as_df function.

5.1 A Small Demonstration

Here we will use the qdap package’s trans_cloud function, on our minimal tm Corpus, to produce a word cloud with particular words highlighted:

```r
matches <- list(
  good = "fun",
  bad = c("dumb", "stinks", "liar")
)
apply_as_df(tm_dat, trans_cloud, grouping.var=NULL, 
  target.words=matches, cloud.colors = c("red", "blue", "grey75"))
```
5.2 Further Examples to Try

Here are some further examples to try:

```r
library(tm)
reut21578 <- system.file("texts", "crude", package = "tm")
reuters <- Corpus(DirSource(reut21578),
                   readerControl = list(reader = readReut21578XML))
apply_as_df(reuters, word_stats)
apply_as_df(reuters, formality)
apply_as_df(reuters, word_list)
apply_as_df(reuters, polarity)
apply_as_df(reuters, Dissimilarity)
apply_as_df(reuters, diversity)
apply_as_df(tm_dat, pos_by)
```
apply_as_df(reuters, flesch_kincaid)
apply_as_df(tm_dat, trans_venn)
apply_as_df(reuters, gantt_plot)
apply_as_df(reuters, rank_freq_mplot)
apply_as_df(reuters, character_table)
apply_as_df(reuters, trans_cloud)

matches2 <- list(
  oil = qcv(oil, crude),
  money = c("economic", "money")
)

(termco_out <- apply_as_df(reuters, termco, match.list = matches2))
plot(termco_out, values = TRUE, high="red")

(wordcor_out <- apply_as_df(reuters, word_cor, word = unlist(matches2)))
plot(wordcor_out)

(f_terms <- apply_as_df(reuters, freq_terms, at.least = 3))
plot(f_terms)

finds <- apply_as_df(reuters, freq_terms, at.least = 5,
  top = 5, stopwords = Top100Words)
apply_as_df(reuters, dispersion_plot, match.terms = finds[, 1],
  total.color = NULL)

6 Conclusion

This vignette described the various data formats for the qdap and tm packages. It also demonstrated some of the basic functionality of the qdap functions designed to navigate between the two packages. For more information on the tm package (Feinerer et al., 2008) use:

browseVignettes(package = "tm")

Likewise, the user may view additional information about the qdap package (Rinker, 2013):
Acknowledgments

qdap relies heavily on the tm package. The tm package has extended text analysis to the R platform. Thank you to Ingo Feinerer and Kurt Hornik for their work on this and many other R packages.

This document was produced with knitr (Xie [2013]). Thank you to Yihui Xie for the knitr package and his many other contributions to the R community.

References


