TextLink_Labnotes_02

Silvie Cinkova, TextLink Training School at UFAL, Prague, 2017-02-07

A shortened version of explor_pdt30, just code

```
library(dplyr)
library(tidyr)
library(stringr)
library(ggplot2)
library(ggthemes)
library(plotluck)
library(scales)
library(formatR)

pdt30 <- readRDS("edu/r/textlink/src_data/pdt_30.RDS")
```

How Many How Long Texts Are There in the Corpus?

```
doclen_set <- pdt30 %>% dplyr::distinct(document_id, .keep_all = TRUE) %>%
dplyr::select(-c(starts_with("discourse"), starts_with("sentence")))
set.seed(122)
dplyr::sample_n(doclen_set, 10)

## # A tibble: 10 × 3
##     document_id         genre number_of_sentences
##          <fctr>        <fctr>               <int>
## 1  mf920925_116          news                   7
## 2  mf920925_120 person_interv                  52
## 3    ln94203_75   description                  25
## 4  cmpr9415_018       comment                  26
## 5   ln95045_059          news                   6
## 6   ln95046_078          news                   5
## 7   ln95047_120          news                  13
## 8  cmpr9410_008        advice                  64
## 9  cmpr9413_052         essay                  58
## 10  ln95045_110          news                  11
```

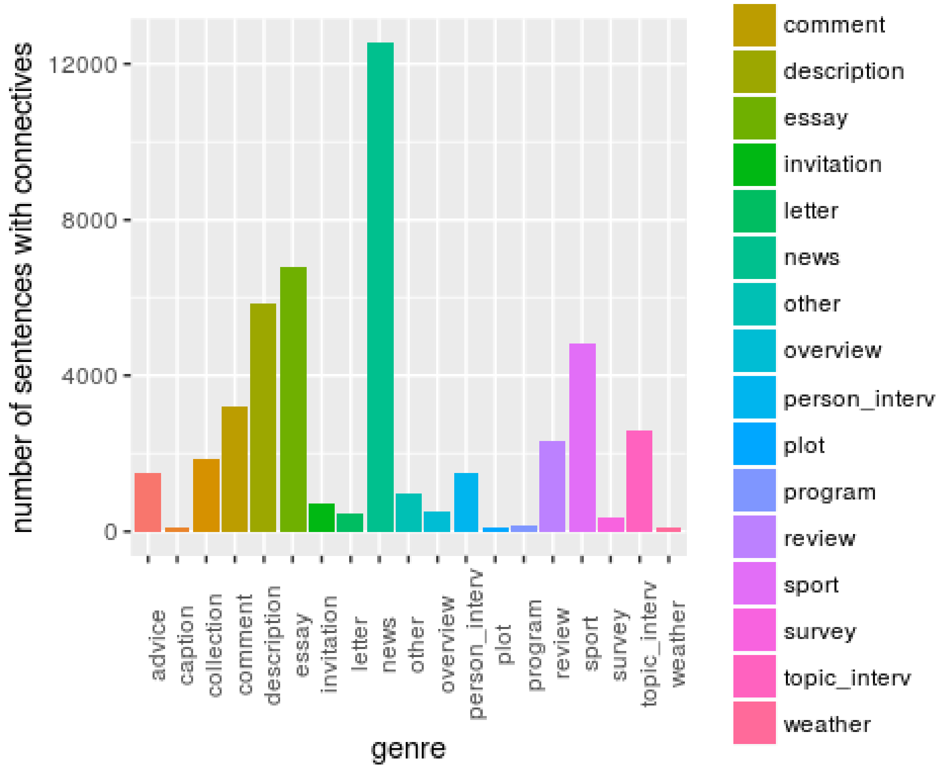
How Much Text Is There in the Corpus for Each Genre?

Text is calculated in length, i.e. number of sentences. This time we focus on the text bulk in each genre, not distinguishing individual documents. We add color distinction to genres for easier comparison with the following plots, although the colors add no information to the barplot.

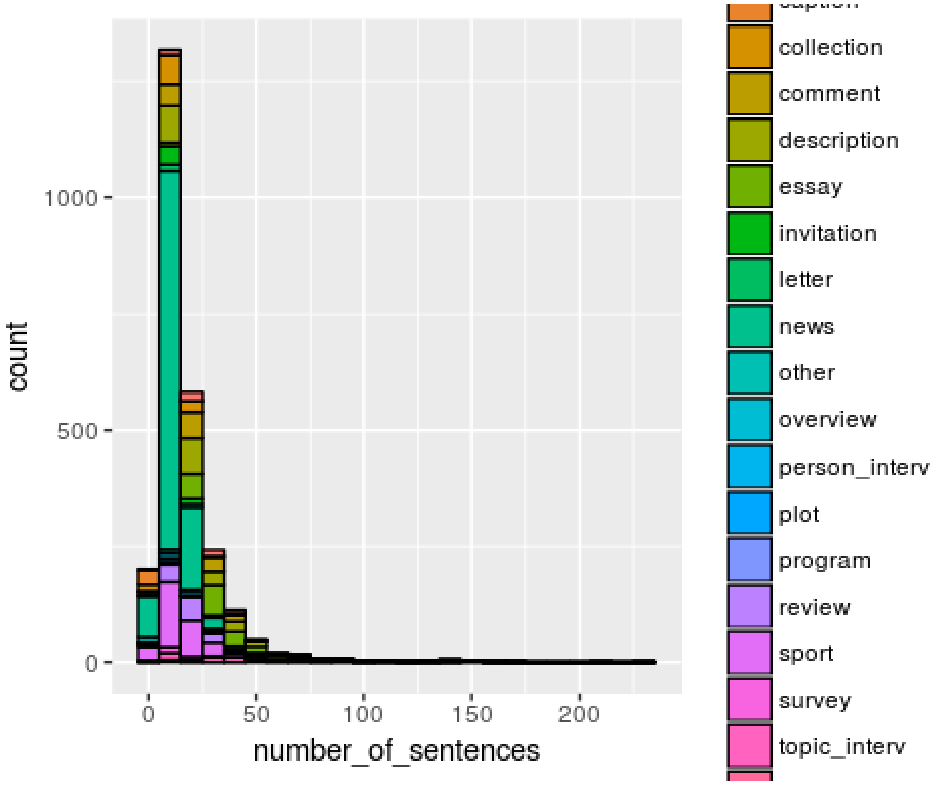
```
sentsums <- dplyr::summarise(group_by(doclen_set, genre),
sum(number_of_sentences))
colnames(sentsums)[2] <- "sumsentnumbers"
sentsums
```

```
## # A tibble: 19 × 2
##             genre sumsentnumbers
##            <fctr>          <int>
## 1          advice           1501
## 2         caption             90
## 3      collection           1833
## 4         comment           3203
## 5     description           5850
## 6           essay           6793
## 7      invitation            693
## 8          letter            434
## 9            news          12537
## 10          other            974
## 11       overview            511
## 12 person_interv            1471
## 13           plot             73
## 14        program            146
## 15         review           2314
## 16          sport           4817
## 17         survey            355
## 18  topic_interv            2602
## 19        weather            105
```

```
ggplot(sentsums, aes(y = sumsentnumbers, x = genre)) + geom_bar(stat =
"identity", aes( fill = genre)) + theme(axis.text.x = element_text(angle =
90)) + ylab("number of sentences with connectives")
```

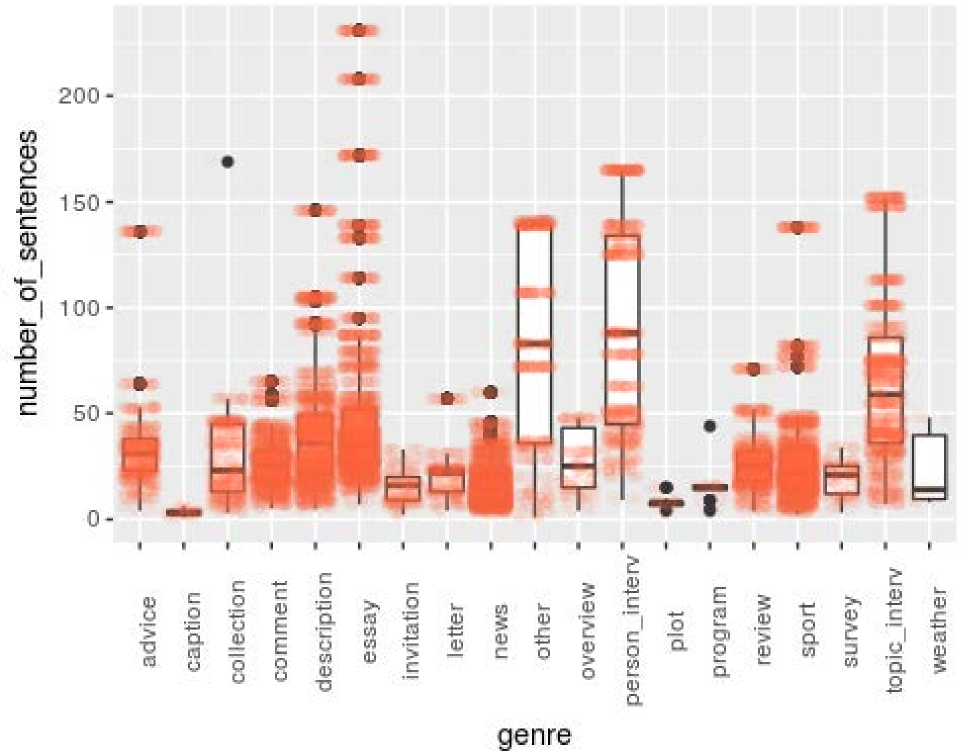


```
#doclen_set <- pdt30 %>% distinct(document_id, .keep_all = TRUE) %>% select(-
c(starts_with("discourse"), starts_with("sentence")))
ggplot(doclen_set, aes(x = number_of_sentences, fill = genre)) +
geom_histogram(binwidth = 10, col = "black")
```

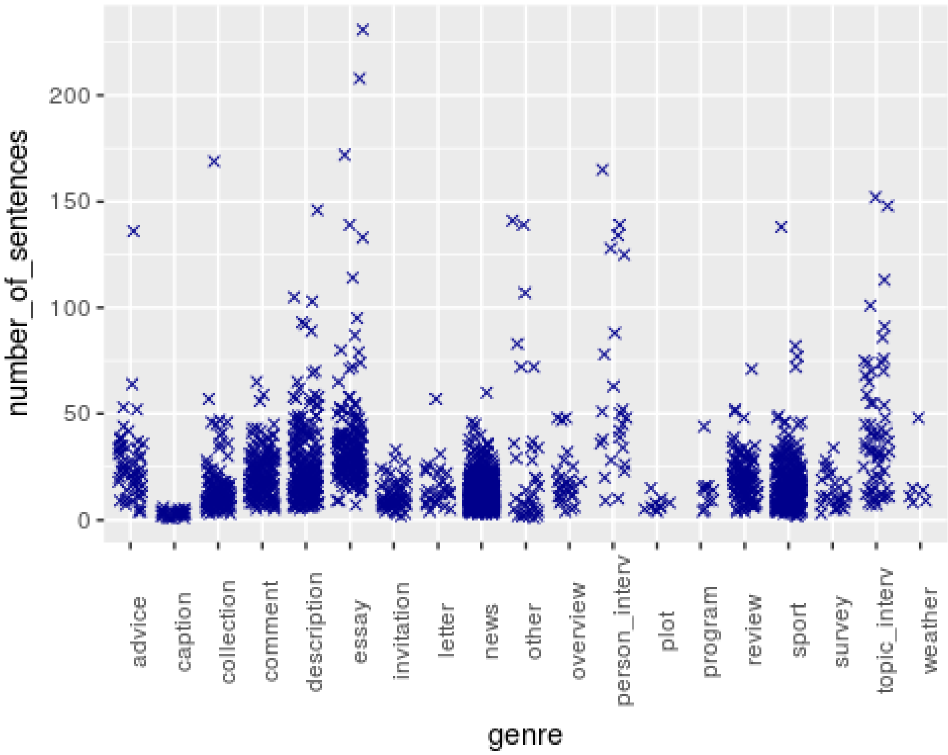


Identify Outliers

```
ggplot(pdt30, aes(x = genre, y = number_of_sentences)) +
  geom_boxplot()  +
  geom_jitter(alpha = 5/100, col = "tomato") +
  theme(axis.text.x =
          element_text(angle = 90))
```

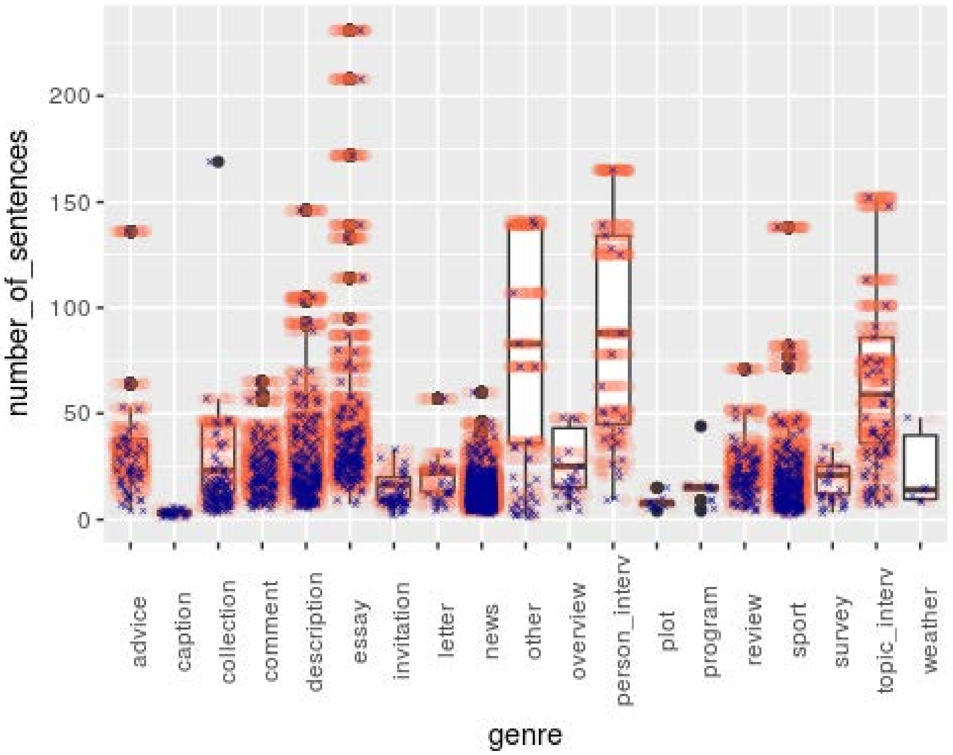


```
ggplot(doclen_set, aes(x = genre, y = number_of_sentences)) +
geom_point(color = "darkblue",  shape = 4, position = position_jitter(height
= 0, width = 0.3) ) +
  theme(axis.text.x = element_text(angle = 90))
```

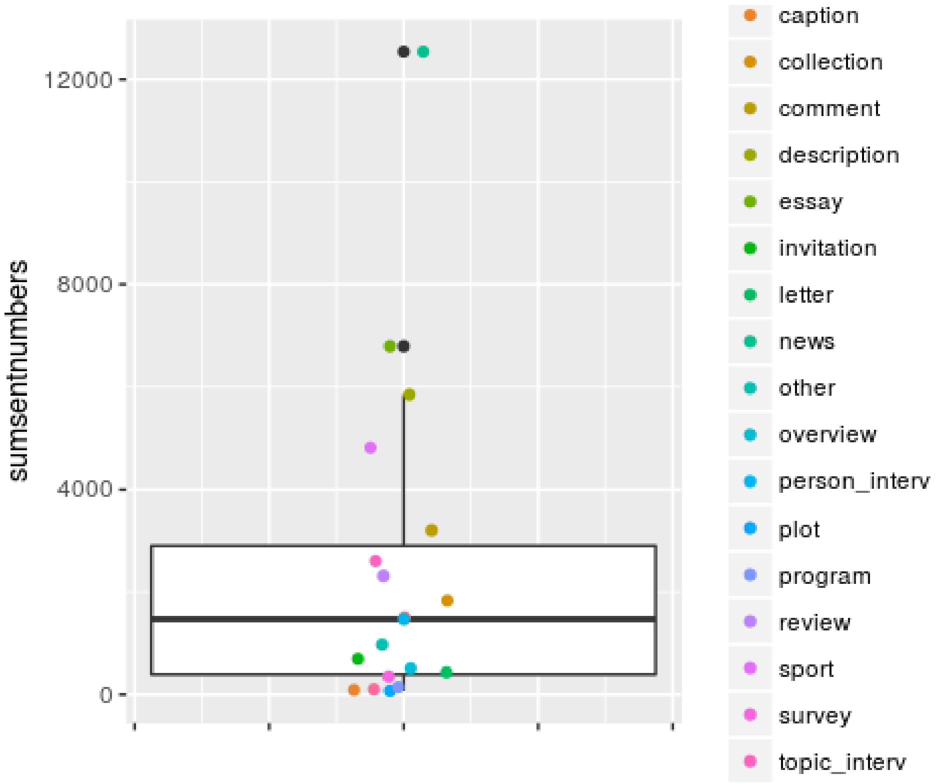


Layered Geoms in One Plot

```
indiv_docs <- select(doclen_set, -1)
ggplot(pdt30, aes(x = genre, y = number_of_sentences)) +
  geom_boxplot()  +
  geom_jitter(alpha = 5/100, col = "tomato") +
  theme(axis.text.x =
          element_text(angle = 90)) +
  geom_point(data = indiv_docs,
             color = "darkblue",
             shape = 4,
             position = position_jitter(height = 0,
                                        width = 0.3),
             alpha = 5/10, size = 2/3)
```

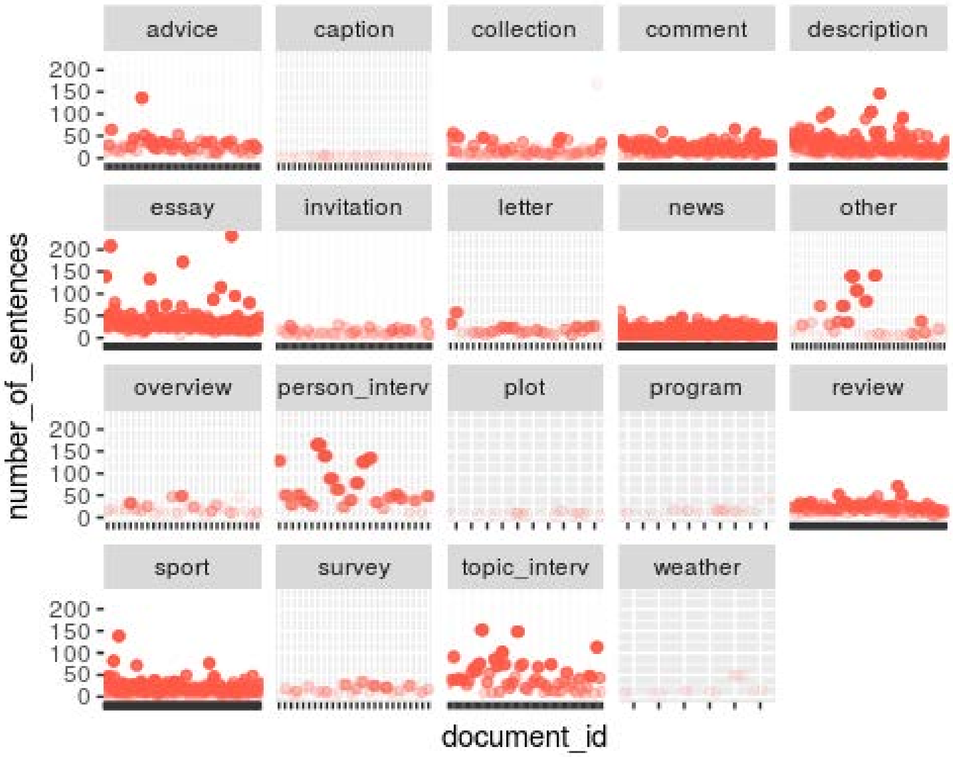


```
ggplot(sentsums, aes(x = 1, y = sumsentnumbers)) + geom_boxplot() +
theme(axis.text.x = element_blank()) + xlab("") + geom_point(aes(y =
sumsentnumbers, col = genre), position = position_jitter(height = 0, width =
0.1))
```



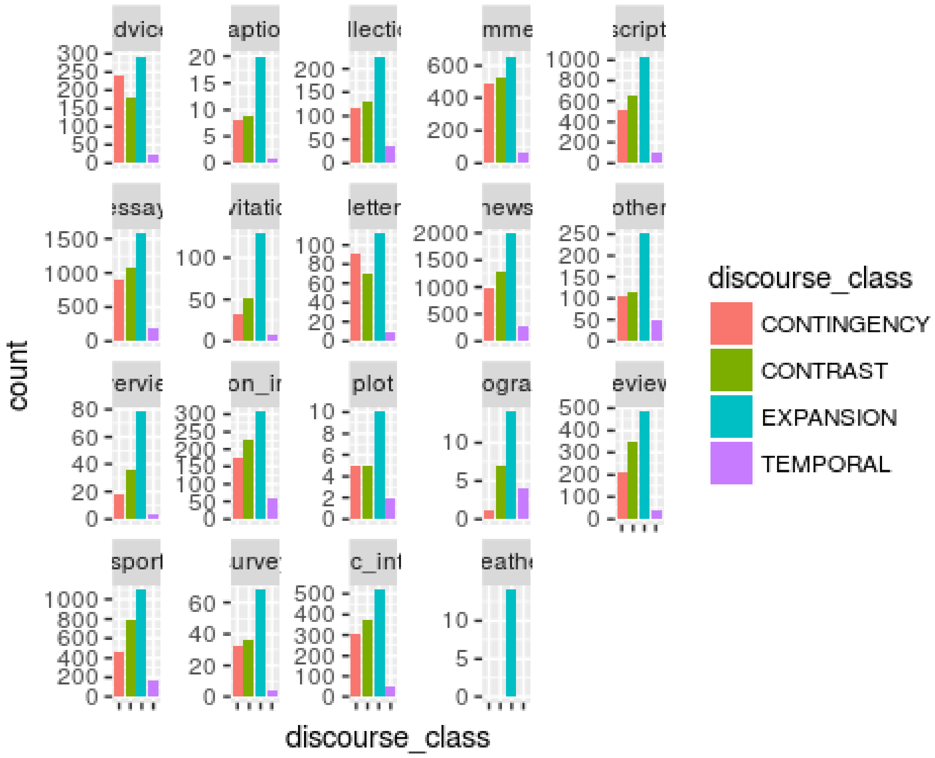
Faceted Plots

```
ggplot(pdt30, aes(x = document_id, y = number_of_sentences)) +
geom_jitter(alpha = 0.06, col = "tomato") + theme(axis.text.x =
element_blank()) + facet_wrap(~ genre, scales = "free_x")
```



```
mappings_01 <- ggplot(data = pdt30, aes(x = discourse_class, fill =
discourse_class))

mappings_01 + geom_bar(position = "dodge") +
  facet_wrap( ~ genre, scales = "free_y") +
theme(axis.text.x = element_blank()) +
  scale_y_continuous(breaks = scales::pretty_breaks())
```



Computing Expected Residuals Manually

Create a contingency table.

```
cont_matrix <- xtabs(formula = ~ genre  + discourse_class , data = pdt30) %>%
as.matrix()

(mat_cols <- rep(colSums(cont_matrix)/sum(cont_matrix), each =
nrow(cont_matrix)) %>%
   matrix(nrow = nrow(cont_matrix), ncol = ncol(cont_matrix)))

##             [,1]      [,2]      [,3]       [,4]
##  [1,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [2,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [3,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [4,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [5,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [6,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [7,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [8,] 0.2286924 0.2888694 0.4305799 0.05185834
##  [9,] 0.2286924 0.2888694 0.4305799 0.05185834
## [10,] 0.2286924 0.2888694 0.4305799 0.05185834
## [11,] 0.2286924 0.2888694 0.4305799 0.05185834
## [12,] 0.2286924 0.2888694 0.4305799 0.05185834
## [13,] 0.2286924 0.2888694 0.4305799 0.05185834
## [14,] 0.2286924 0.2888694 0.4305799 0.05185834
## [15,] 0.2286924 0.2888694 0.4305799 0.05185834
## [16,] 0.2286924 0.2888694 0.4305799 0.05185834
## [17,] 0.2286924 0.2888694 0.4305799 0.05185834
## [18,] 0.2286924 0.2888694 0.4305799 0.05185834
## [19,] 0.2286924 0.2888694 0.4305799 0.05185834
```

```
(mat_rows <- rep(rowSums(cont_matrix)/sum(cont_matrix), each =
ncol(cont_matrix)) %>%
   matrix(nrow = nrow(cont_matrix), ncol = ncol(cont_matrix), byrow = TRUE))
```

```
##                [,1]         [,2]         [,3]         [,4]
##  [1,] 0.0360478692 0.0360478692 0.0360478692 0.0360478692
##  [2,] 0.0018486087 0.0018486087 0.0018486087 0.0018486087
##  [3,] 0.0247129792 0.0247129792 0.0247129792 0.0247129792
##  [4,] 0.0842089901 0.0842089901 0.0842089901 0.0842089901
##  [5,] 0.1121327106 0.1121327106 0.1121327106 0.1121327106
##  [6,] 0.1827690212 0.1827690212 0.1827690212 0.1827690212
##  [7,] 0.0106051761 0.0106051761 0.0106051761 0.0106051761
##  [8,] 0.0135726795 0.0135726795 0.0135726795 0.0135726795
##  [9,] 0.2194006616 0.2194006616 0.2194006616 0.2194006616
## [10,] 0.0253453979 0.0253453979 0.0253453979 0.0253453979
## [11,] 0.0065187780 0.0065187780 0.0065187780 0.0065187780
## [12,] 0.0373127068 0.0373127068 0.0373127068 0.0373127068
## [13,] 0.0010702471 0.0010702471 0.0010702471 0.0010702471
## [14,] 0.0012648375 0.0012648375 0.0012648375 0.0012648375
## [15,] 0.0521502238 0.0521502238 0.0521502238 0.0521502238
## [16,] 0.1228351819 0.1228351819 0.1228351819 0.1228351819
## [17,] 0.0068593112 0.0068593112 0.0068593112 0.0068593112
## [18,] 0.0606635532 0.0606635532 0.0606635532 0.0606635532
## [19,] 0.0006810664 0.0006810664 0.0006810664 0.0006810664
```

```
(exp_matrix <- (mat_cols * mat_rows * sum(cont_matrix)) %>% round(1))
```

```
##         [,1]   [,2]   [,3]  [,4]
##  [1,]  169.5  214.1  319.1  38.4
##  [2,]    8.7   11.0   16.4   2.0
##  [3,]  116.2  146.7  218.7  26.3
##  [4,]  395.9  500.0  745.3  89.8
##  [5,]  527.1  665.8  992.5 119.5
##  [6,]  859.2 1085.3 1617.7 194.8
##  [7,]   49.9   63.0   93.9  11.3
##  [8,]   63.8   80.6  120.1  14.5
##  [9,] 1031.4 1302.8 1941.9 233.9
## [10,]  119.1  150.5  224.3  27.0
## [11,]   30.6   38.7   57.7   6.9
## [12,]  175.4  221.6  330.3  39.8
## [13,]    5.0    6.4    9.5   1.1
## [14,]    5.9    7.5   11.2   1.3
## [15,]  245.2  309.7  461.6  55.6
## [16,]  577.4  729.4 1087.2 130.9
## [17,]   32.2   40.7   60.7   7.3
## [18,]  285.2  360.2  536.9  64.7
## [19,]    3.2    4.0    6.0   0.7
```

```
row.names(exp_matrix) <- row.names(cont_matrix)
colnames(exp_matrix) <- colnames(cont_matrix)
exp_matrix
```

```
##               CONTINGENCY CONTRAST EXPANSION TEMPORAL
## advice              169.5    214.1     319.1     38.4
## caption               8.7     11.0      16.4      2.0
## collection          116.2    146.7     218.7     26.3
## comment             395.9    500.0     745.3     89.8
## description         527.1    665.8     992.5    119.5
## essay               859.2   1085.3    1617.7    194.8
## invitation           49.9     63.0      93.9     11.3
## letter               63.8     80.6     120.1     14.5
## news               1031.4   1302.8    1941.9    233.9
## other               119.1    150.5     224.3     27.0
## overview             30.6     38.7      57.7      6.9
## person_interv       175.4    221.6     330.3     39.8
## plot                  5.0      6.4       9.5      1.1
## program               5.9      7.5      11.2      1.3
## review              245.2    309.7     461.6     55.6
## sport               577.4    729.4    1087.2    130.9
## survey               32.2     40.7      60.7      7.3
## topic_interv        285.2    360.2     536.9     64.7
## weather               3.2      4.0       6.0      0.7
```

```
which(exp_matrix < 5, arr.ind = TRUE)
```

```
##         row col
## weather  19   1
## weather  19   2
## caption   2   4
## plot     13   4
## program  14   4
## weather  19   4
```

```
levels(pdt30$genre)[levels(pdt30$genre) %in% c("weather", "caption", "plot",
"program")] <- "other"
```