First exploratory analysis of BRD

1. Publisher-level analysis.

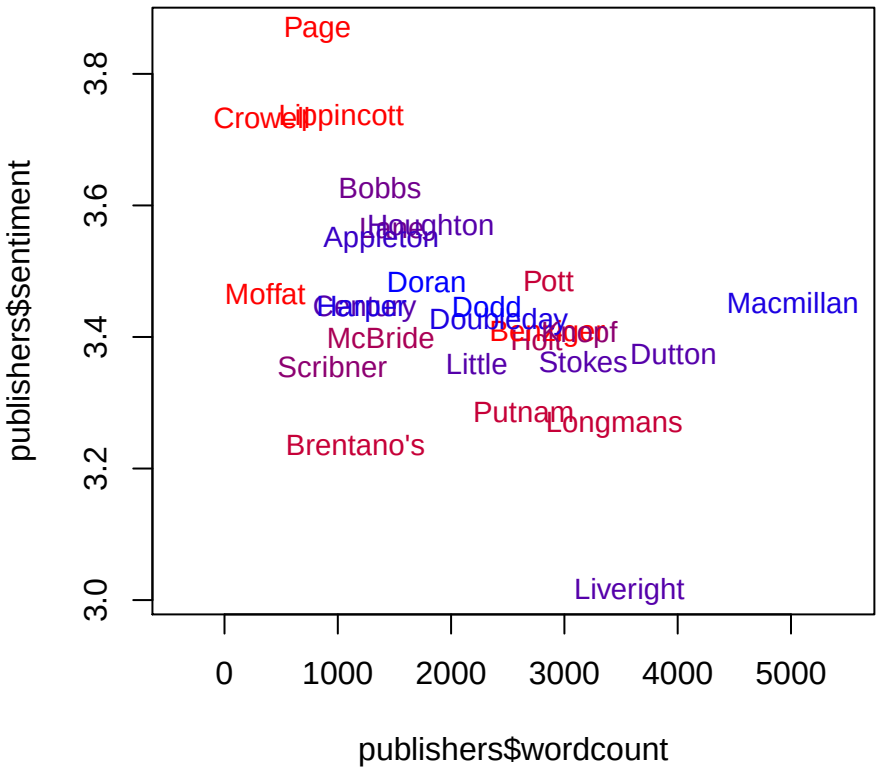
The script `aggregate_publishers.py` goes through `pairedvoloutfile.tsv` (or `pairedficvol14.tsv`) to aggregate price, sentiment, and total wordcount data at the publisher level.

Below we plot the average sentiment (+/-) for a publishers' books against the average *total* wordcount of the reviews (those reported in BRD, of course.)

```
publishers <- read.csv('publishers.tsv', sep = '\t')
str(publishers)
```

```
## 'data.frame':    27 obs. of  6 variables:
##  $ publisher : Factor w/ 27 levels "Appleton","Benziger",..: 12 21 15 24 2 26 9 8 5 20 ...
##  $ wordcount : num  2753 1380 1477 2860 2850 ...
##  $ price     : num  1.5 1.32 1.48 0.35 1.86 ...
##  $ sentiment : num  3.4 3.4 3.57 3.49 3.41 ...
##  $ numreviews: num  8.33 6.33 7 7 12 ...
##  $ numbooks  : int  3 3 6 2 1 5 16 29 10 17 ...
```

```
plot(publishers$wordcount, publishers$sentiment, type = 'n', xlim = c(-400, 5500))
rbPal <- colorRampPalette(c('red','blue'))
publishers$Col <- rbPal(10)[as.numeric(cut(log(publishers$numbooks),breaks = 10))]
text(publishers$wordcount, publishers$sentiment,
     labels = publishers$publisher, col = publishers$Col, cex = 0.9)
```



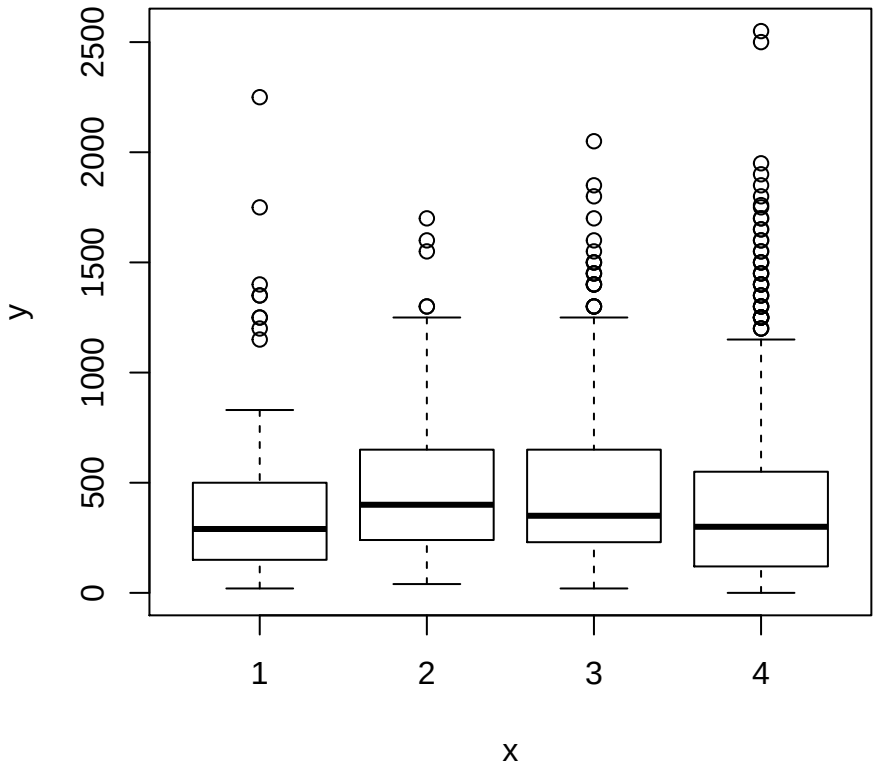

```
cor.test(publishers$wordcount, publishers$sentiment)
```

```
##
##  Pearson's product-moment correlation
##
## data:  publishers$wordcount and publishers$sentiment
## t = -3.023, df = 25, p-value = 0.005713
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7499512 -0.1709950
## sample estimates:
##        cor
## -0.5173888
```

There's a very clear pattern and significant negative correlation. I think it would remain significant even if we weighted the test to focus on the largest publishers (which are the blue and purple ones). I infer that getting a lot of attention ≠ good reviews.

The negative correlation holds across both elements of total wordcount–the number of reviews per book, and the average wordcount per review. It is esp strong for the latter component. We can perhaps understand contributing factors to this by examining the distribution of review wordcounts across sentiment levels, from 1 to 4: -, -/+, +/-, and +.

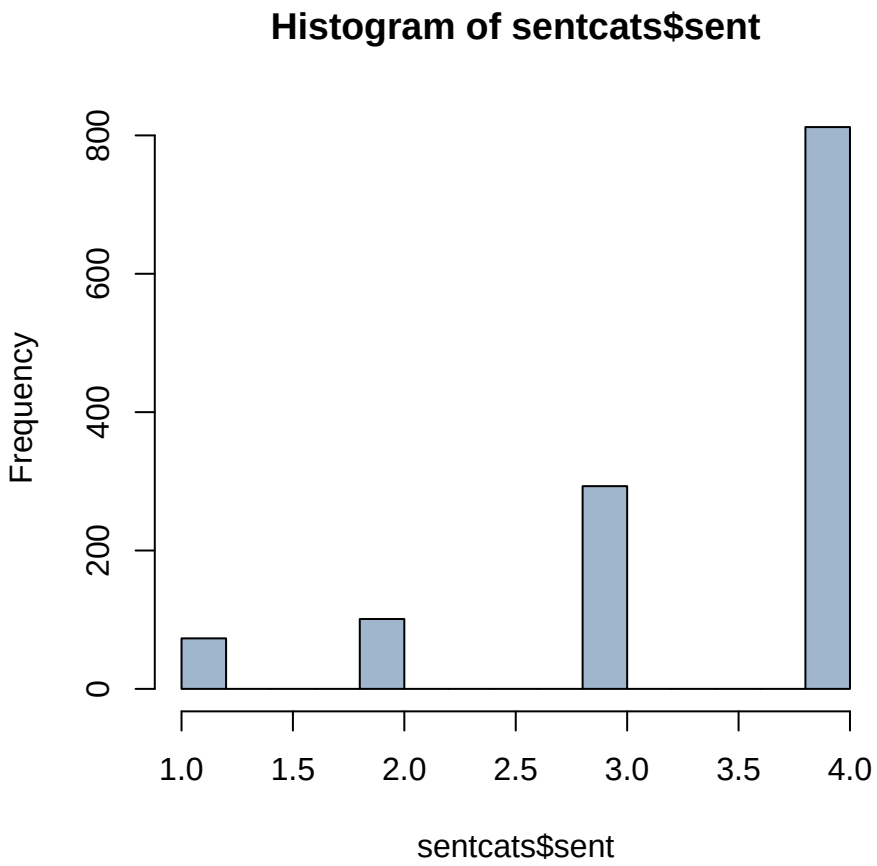
```
sentcats <- read.csv('sentcats.tsv', sep = '\t')
sentcats$sfactor <- as.factor(as.character(sentcats$sent))
plot(sentcats$sfactor, sentcats$wcount)
```



Basically, the median wordcount is significantly higher for ambiguous reviews than for straightforwardly positive ones. This might well have as much to do with the BRD editors' ability to discern shadings of sentiment as it does with the original sentiment. It would be interesting to re-run the publisher correlation using automated sentiment analysis.

And note that while wordcount is higher for category 2 than 3, the trend is probably more shaped by category 3.

```
hist(sentcats$sent, col = 'slategray3')
```

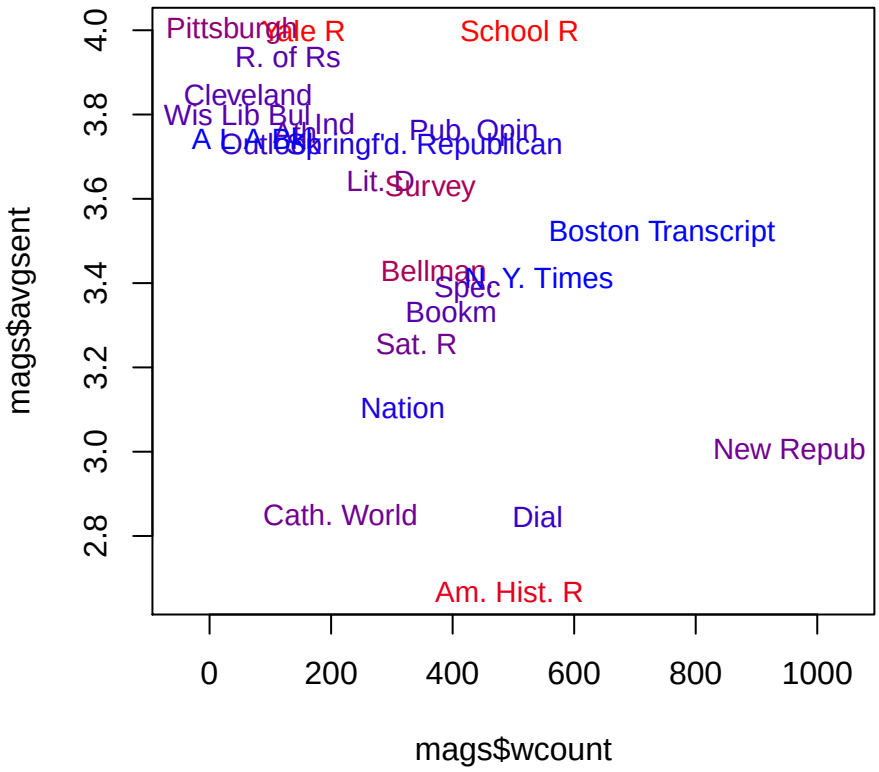


On the other hand, there are also reasons to suspect that this is not just an artefact. One part of what's going on is that more ambitious books get held to higher standards, partly because they're reviewed in periodicals that apply higher standards.

```
mags <- read.csv('publicationstats.tsv', sep = '\t')
str(mags)
```

```
## 'data.frame':    26 obs. of  4 variables:
##  $ pubname: Factor w/ 26 levels "A L A Bkl","Am. Hist. R",..: 2 3 4 5 7 9 10 11 13 12 ...
##  $ wcount : num  493 141 369 398 216 ...
##  $ avgsent: num  2.67 3.76 3.43 3.33 2.85 ...
##  $ numrevs: int  4 43 14 58 33 73 42 31 142 285 ...
```

```
plot(mags$wcount, mags$avgsent, type = 'n', xlim = c(-50, 1050))
rbPal <- colorRampPalette(c('red','blue'))
mags$Col <- rbPal(10)[as.numeric(cut(log(mags$numrevs), breaks = 10))]
text(mags$wcount, mags$avgsent, labels = mags$pubname, col = mags$Col, cex = 0.9)
```




```
cor.test(mags$wcount, mags$avgsent)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mags$wcount and mags$avgsent
## t = -3.0256, df = 23, p-value = 0.006018
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7669940 -0.1754207
## sample estimates:
##        cor
## -0.5335678
```

The strong negative correlation there would get even stronger if we ignored publications with a very low number of reviews (the red outliers).

Bassically, the *New Republic* writes long reviews, and they're going to rag on you. Also, e.g. *Dial*, compared to say *A L A Booklist* in the upper corner, which is going to provide a short descriptive/positive account of a book.

But that doesn't mean that it's bad news to get covered in *The New Republic* or *Dial*!

The statistics for this are produced by `analyze_publications.py`, and require some data from the original full review data, with +/- sentiment at the review level.

To push this a little further I started to attempt factor analysis of the amount of attention books receive from different publications. Data for this is more complete than sentiment data; it's easier to export; and I have an intuition that it is going to be more revealing than sentiment, actually.

So I construct a matrix where columns are publications (newspapers or magazines), rows are books, and each cell is filled with the number of words publication X devoted to book Y. I do also include price and book-level avg sentiment as columns, but they're not crucial to the analysis.

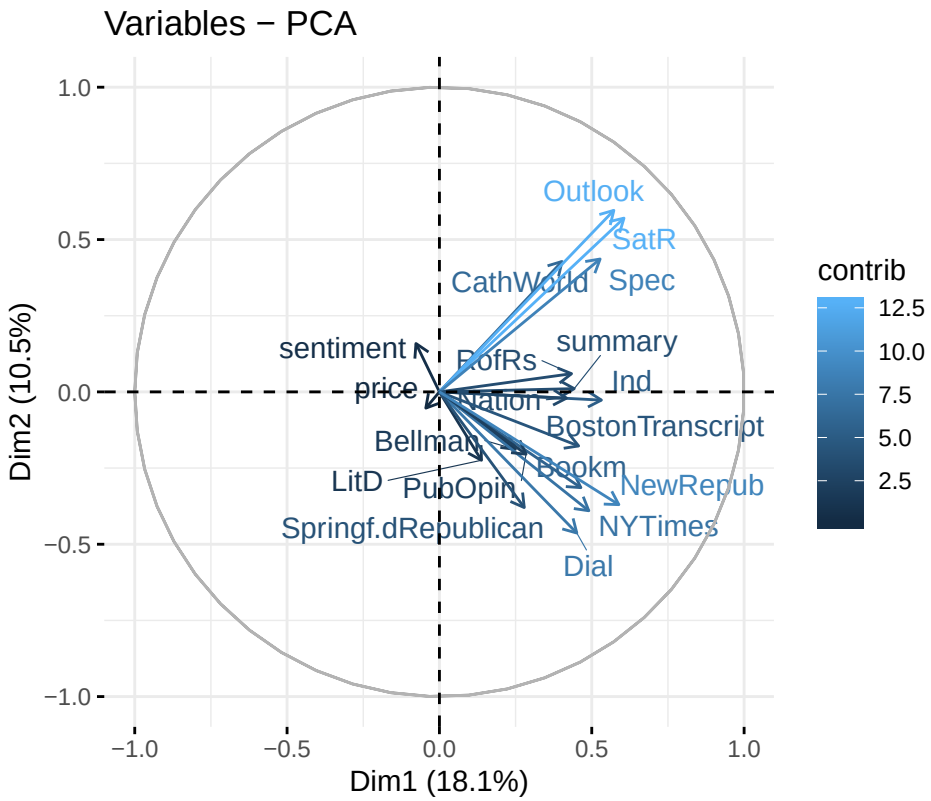
Then I perform PCA. No rotation is required, the PCA in this case is super-interpretable.

```
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
bookmatrix <- read.csv('bookmatrix.tsv', sep = '\t')
pub.pca <- prcomp(bookmatrix, scale = TRUE)
fviz_pca_var(pub.pca,col.var = "contrib", repel = TRUE, axes = c(1,2))
```

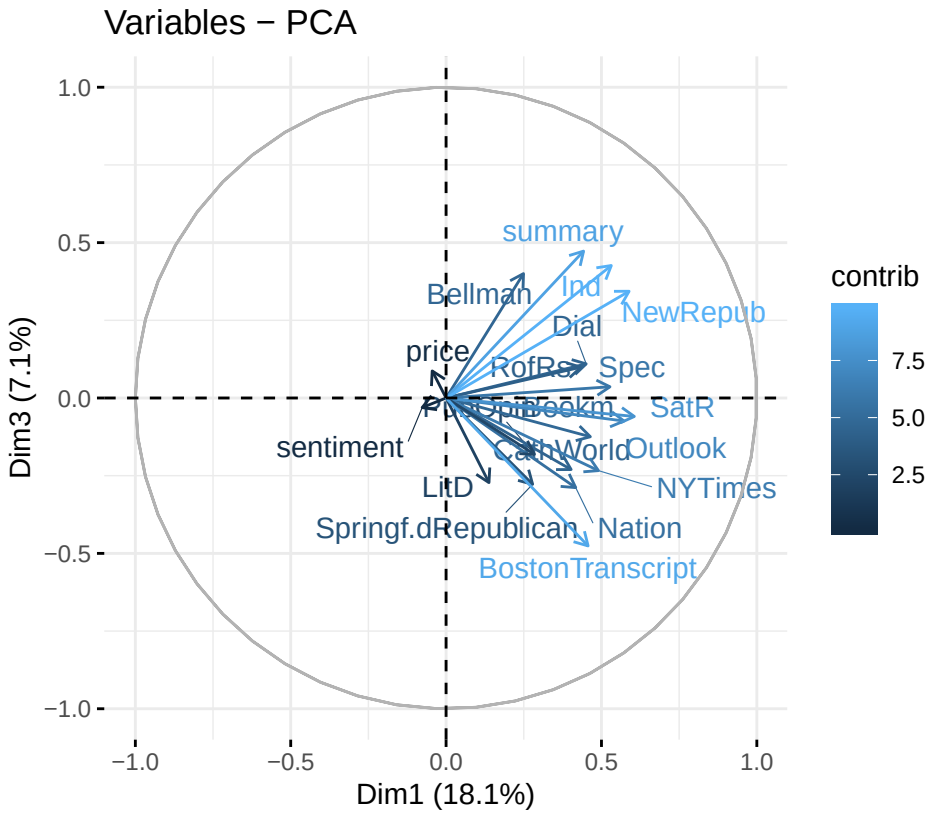


So, the first dimension is very simple. Some books get more attention than others. Getting more attention from publication A does not, in general, reduce the amount of attention you get from publication B; in general, they all correlate positively. For the most part. That's why all the arrows in the most important dimension point to the right.

But there's also a second dimension, which is basically, British/American. Here there actually is a zero-sum relationship. Getting more from one direction means less from the other.

More interesting to me is the third dimension.

```
fviz_pca_var(pub.pca,col.var = "contrib", repel = TRUE, axes = c(1,3))
```



This is basically, literary or political magazines with intellectual pretensions (*TNR, Bellman, Dial, The Independent*) versus newspapers and general-interest publications. This is our first, promising sign that factor analysis can by itself reveal market segmentation!

For more on PCA in R and the useful `factoextra` packages see: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

Things I want to explore further:

1. Obviously, how all of the above changes across time.
2. Confirm my intuition that wordcount/attention is more important than sentiment; I might do this by testing
   - how stable both metrics are from one book by author A to the next
   - or how well both metrics predict external measures of sales/prestige
3. What do career arcs look like? To answer this we might need to manually code books with an integer sequence-in-authors-career number. We could do this for a subset of authors across 12 years; start by selecting authors who appear > 2 times? or distribute an equal number of authors who appear 1,2,3 ... n times. Ask whether e.g. sentiment for book i affects the amount of attention received at $i+1$. Or, which publications do most to boost a career from i to $i+1$?