Sentiment analysis is typically applied to connected text such as product reviews. However, it can also be extended to names, potentially delivering rich insights into psychology and culture. Globally and historically, names hold important familial, cultural, and religious significance. The foundation for much of this significance is a concept called *nominal realism*, which holds that the name imbues characteristics into the named. For instance, personal names in many cultures are based on totemic animals so that valued traits of the totem are transferred to the human namesake. We see nominal realism in our own culture from our tendency to name sports teams such as the Detroit Lions and Chicago Bears after predatory animals with stereotypically aggressive dispositions rather than, say, the Cincinnati Sloths, Chicago Sheep, or Green Bay Guinea Pigs. My own research has examined nominal realism in names by documenting biases toward positive versus negative sentiment in names. For example, in cultures around the world, people emphasize the positive more than the negative in everyday speech. But I found a much more pronounced focus on the positive in a sentiment analysis of US place names. The positivity bias is especially large in names of cities and towns – which are closely connected to the self – than names of natural features. More recently, I've shown that business names also show a strong bias toward positive over negative words – with consequences for business performance. Specifically, revenues of businesses containing negative words are significantly lower than those for businesses containing positive or neutral words. In this post, I will extend sentiment analysis to surnames such as "Smith" and "Jones". Surnames are interesting since technically they have no meaning, although they may at one time. Today's "Shoemakers" for example are probably no more likely to be in that profession than those with other surnames (though I suppose this is an assumption that warrants testing). That said, sentiment analysis would code surnames like "Grief" and "Coward" as negative while "Hardy" and "Courage" would be coded positive. Nominal realism would predict that negative surnames would be less common than positive surnames given fears that negative characteristics of the name would carry over to the named. I tested this hypothesis using a data set of surnames occurring at least 100 times in the 2010 US Census. We'll start the analysis by first reading the downloaded csv file into a data frame, and then streamlining to just the two key variables used in the analysis, the name and count of occurrences:

```
surnames <- read.csv("/Users/mike/R/Names/Names_2010Census.csv",header=
TRUE)
surnames <- select(surnames, name, count)
head(surnames)
      name   count
1    SMITH 2442977
2  JOHNSON 1932812
3 WILLIAMS 1625252
4    BROWN 1437026
5    JONES 1425470
6   GARCIA 1166120
```

Next, we'll convert the surnames to lower case for matching to a sentiment dictionary:

```
surnames$name <- tolower(surnames$name)
```

We'll identify surnames with positive or negative sentiment using the AFINN sentiment lexicon, specifically the 2011 version. Each of the 2477 words in this lexicon is coded with an integer score ranging from -5 to +5 with negative/positive values reflecting sentiment valence and magnitude. I downloaded this lexicon, saving in Excel which we'll load and merge with the surnames data frame. We'll then remove all non-matching surnames (i.e., the vast majority of names like "Baker" and "Smith" with neutral sentiment).

```
affin <- read_excel("/Users/mike/R/AFFIN_Sentiment_Lexicon.xlsx",
sheet="AFINN-111")
surnames <- left_join(surnames,affin,by=c("name"="Word"))
surname_sent <- filter(surnames,!is.na(Sentiment))
```

This leaves us with 332 surnames with a coded sentiment score, representing 13% of the words in the AFINN sentiment lexicon. We can look at a few surnames randomly selected from those with positive and negative sentiment to get a sense for them:

```
filter(surname_sent, Sentiment > 0) %>% slice_sample(n=10)
        name count Sentiment
1       free  9923         1
2      mercy   575         2
3    freedom   138         2
4       gift  1490         2
5   straight  4307         1
6       fair 18609         2
7      spark   472         1
8     heaven   625         2
9      hardy 80252         2
10 brilliant   491         4

filter(surname_sent, Sentiment < 0) %>% slice_sample(n=10)
      name count Sentiment
1     fail   754        -2
2  failing   717        -2
3   sullen   401        -2
4    angry   154        -3
5     bias  6518        -1
6     sore   115        -1
7    moody 64429        -1
8    blind   835        -1
9     glum   118        -2
10    lack  2661        -2
```

Next, we'll test whether surnames with positive sentiment occur more frequently than those with negative sentiment, as nominal realism would predict. Consistent with other word frequency analyses that include words with a huge frequency range (100-2,442,977), we'll first convert the frequency counts to logs and use those values in a t-test:

```
t.test(log10(surname_sent$count[surname_sent$Sentiment
>0]),log10(surname_sent$count[surname_sent$Sentiment < 0]))
    Welch Two Sample t-test

data:  log10(surname_sent$count[surname_sent$Sentiment > 0]) and
log10(surname_sent$count[surname_sent$Sentiment < 0])
t = 3.5516, df = 322.6, p-value = 0.0004399
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1283081 0.4469758
sample estimates:
mean of x mean of y
 3.130305  2.842663
```

Results were in the predicted direction, with mean frequency of positive surnames ~1350 and negative surnames ~700, or almost 2:1. I replicated the results using another sentiment lexicon. In sum, surname usage in the US shows a bias toward positive sentiment/avoidance of negative sentiment similar to those seen in US place and business names. It would be interesting to test whether there are significant consequences of having a negative surname (e.g., like the analysis of negative business names described above).