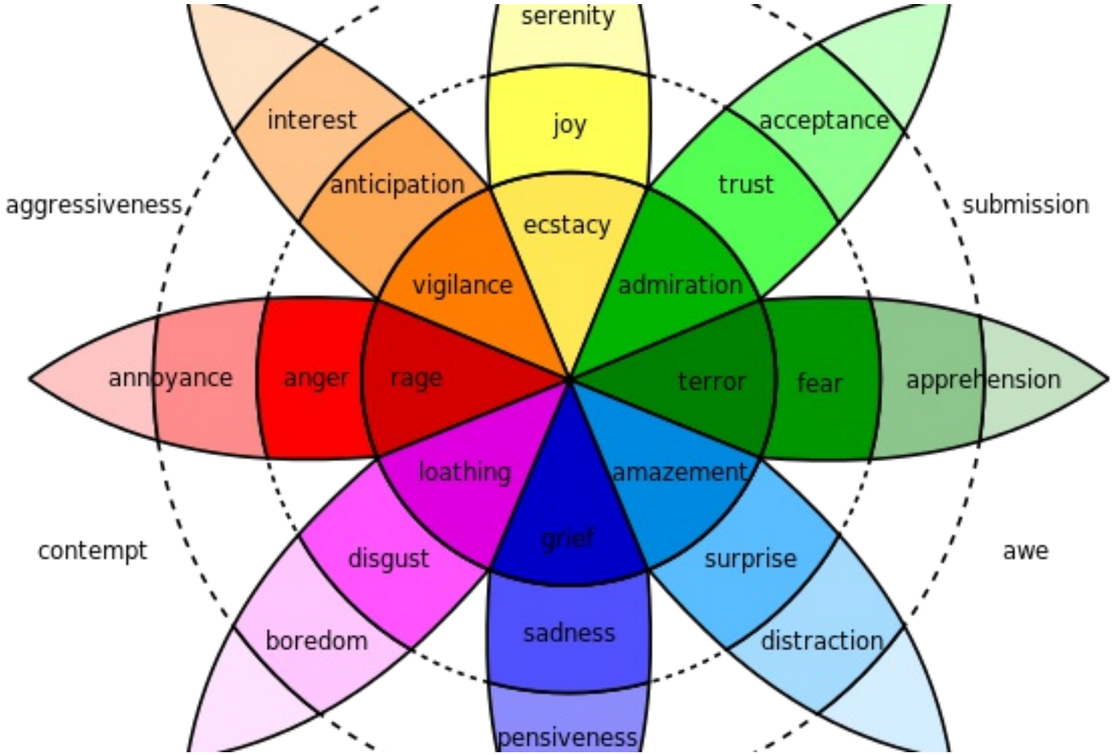
SENTIMENT ANALYSIS IN R

Plutchik's wheel of emotion, polarity vs. sentiment

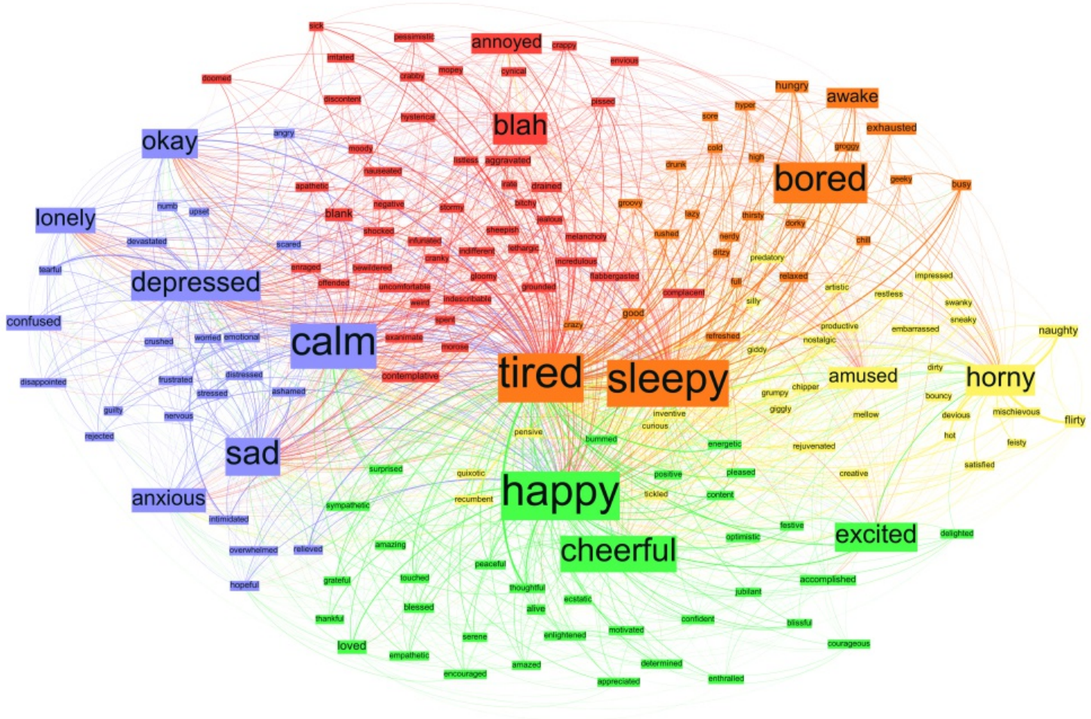
Ted Kwartler
Data Dude

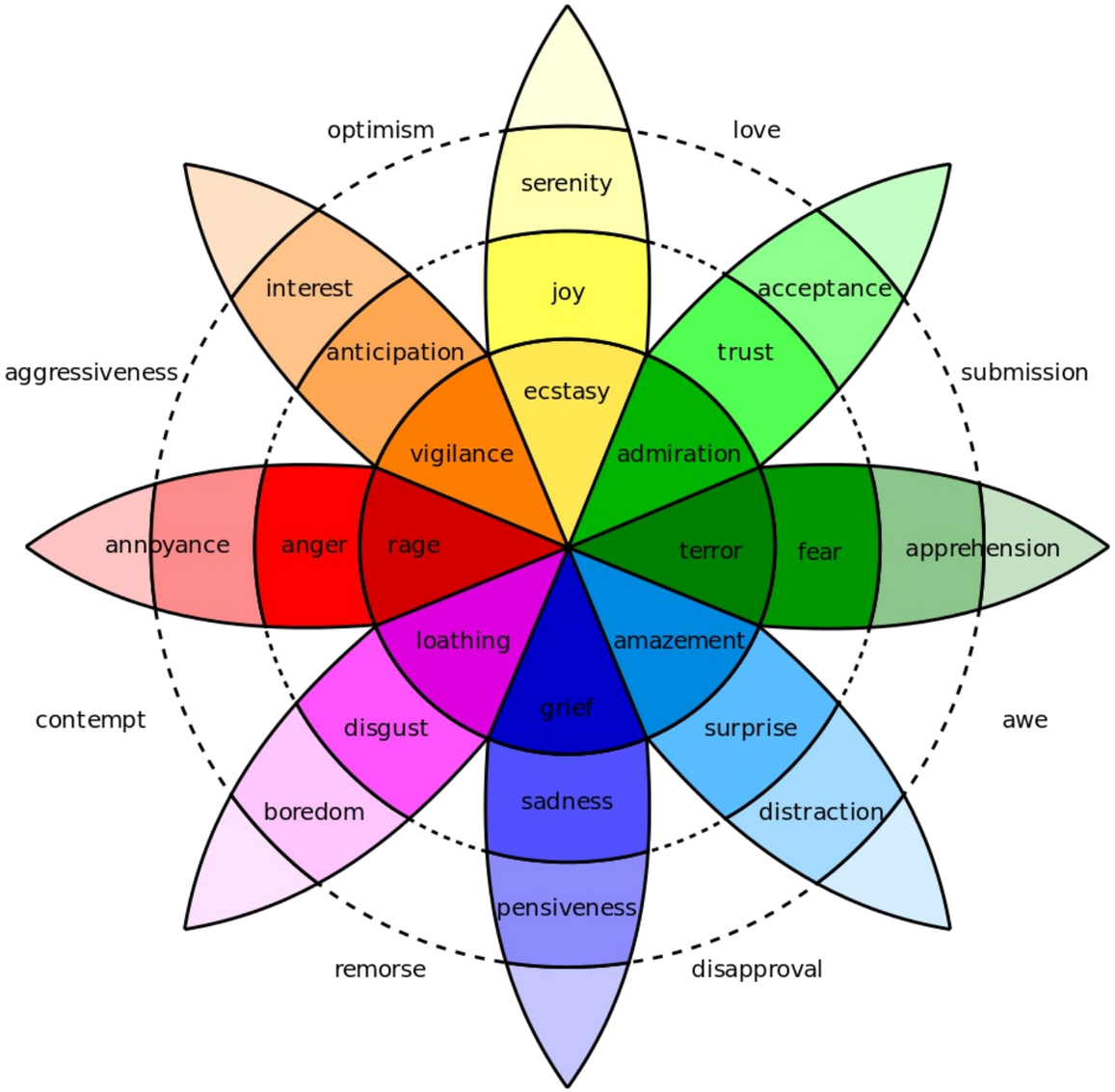
In reality sentiment is more complex than +/-

Plutchik’s Wheel of Emotion



A more complex emotional framework for comparison




optimism
love
serenity
interest
joy
acceptance
anticipation
trust
aggressiveness
ecstasy
submission
vigilance
admiration
annoyance
anger
rage
terror
fear
apprehension
loathing
amazement
contempt
grief
awe
disgust
surprise
boredom
sadness
distraction
pensiveness
remorse
disapproval

SENTIMENT ANALYSIS IN R

Let's practice!

SENTIMENT ANALYSIS IN R

Bing lexicon with an inner join

Ted Kwartler
Data Dude

Table Joins

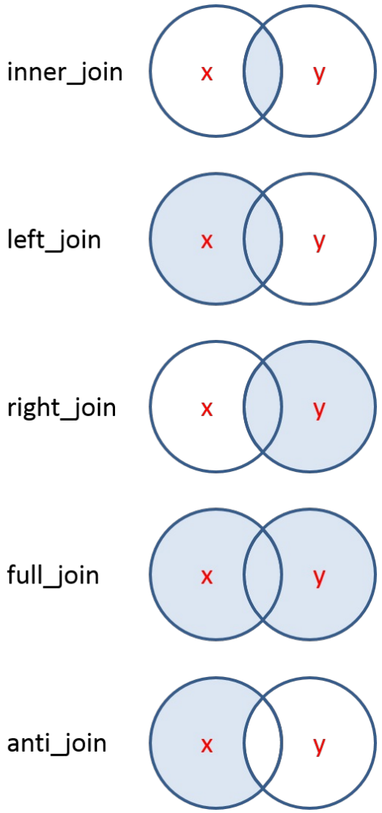

inner_join
x
y
left_join
x
y
right_join
x
y
full_join
x
y
anti_join
x
y

Table Joins

dplyr **Joins**

```
inner_join(x, y, ...)
left_join(x, y, ...)
right_join(x, y, ...)
full_join(x, y, ...)
semi_join(x, y, ...)
anti_join(x, y, ...)
```

Declaring the by parameter:

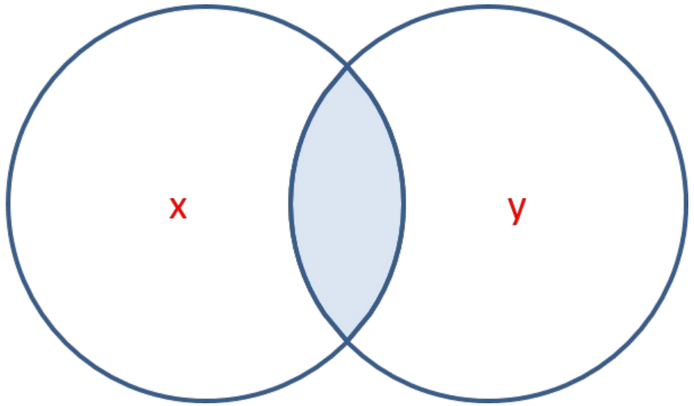
```
inner_join(x, y, by = "shared_column")
```

or

```
inner_join(x, y, by = c("a" = "b"))
```

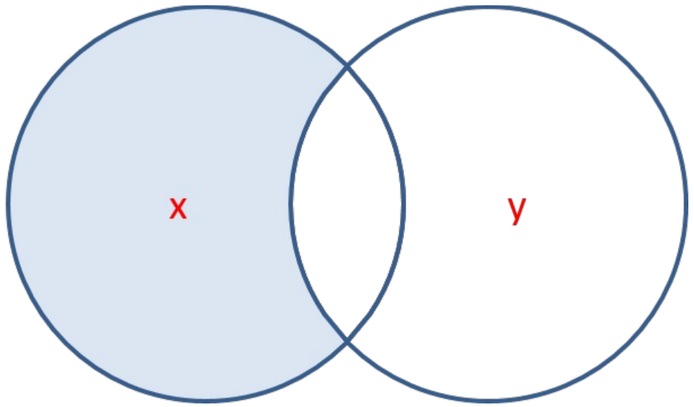
Comparing Inner and Anti Joins

inner_join



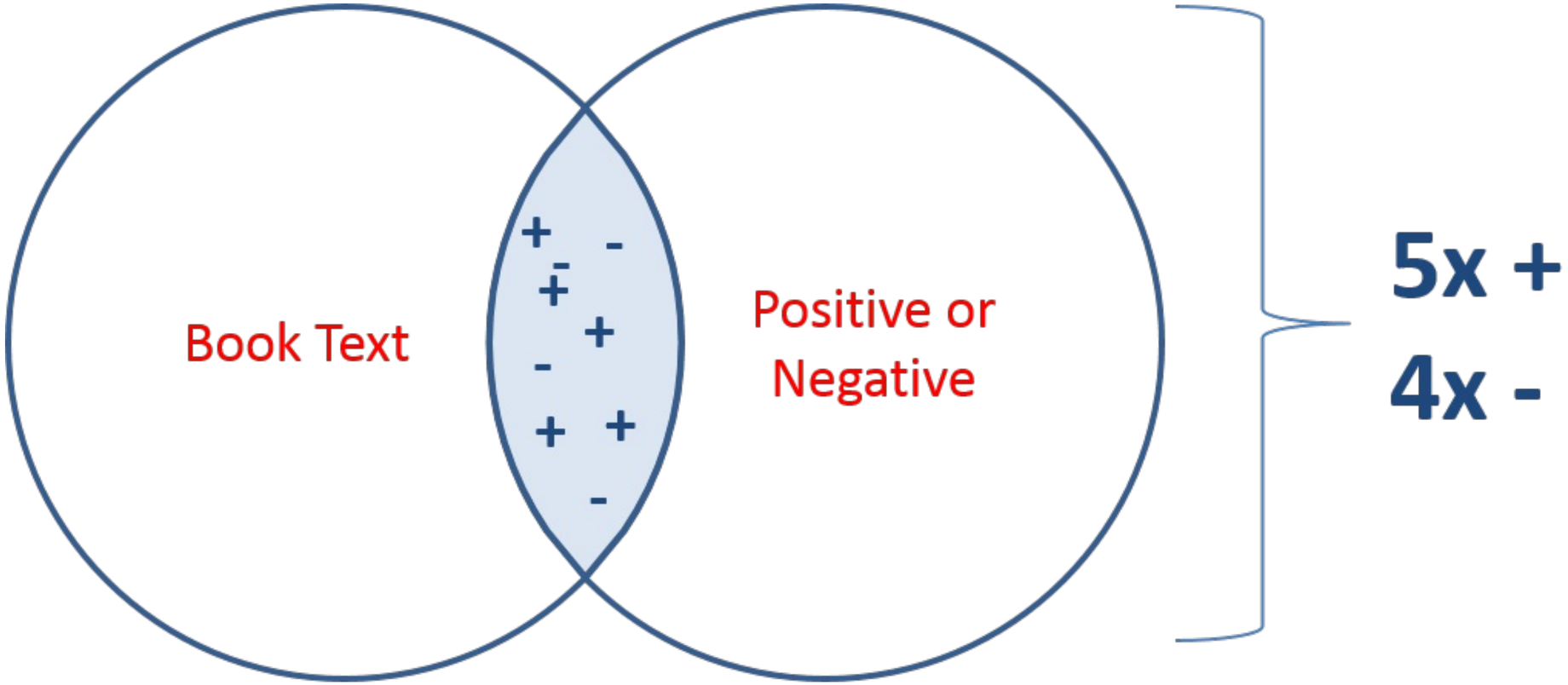

```
inner_join(
  text_table,
  subjectivity_lexicon,
  by = "word_column"
)
```

anti_join




```
anti_join(
  text_table,
  stopwords_table,
  by = "word_column"
)
```

Starting with positive/negative



SENTIMENT ANALYSIS IN R

Let's practice!

SENTIMENT ANALYSIS IN R

AFINN & NRC inner joins

Ted Kwartler
Data Dude

AFINN

Load & Subset

```
> library(tidytext)
> data(sentiments)
> afinn <- subset(sentiments, sentiments$lexicon == "AFINN")
```

Result

```
> tail(afinn)
# A tibble: 6 × 4
      word sentiment lexicon score
     <chr>     <chr>   <chr> <int>
1 youthful      <NA>   AFINN     2
2    yucky      <NA>   AFINN    -2
3    yummy      <NA>   AFINN     3
4   zealot      <NA>   AFINN    -2
5  zealots      <NA>   AFINN    -2
6  zealous      <NA>   AFINN     2
```

NRC

Load & Subset

```
> library(tidytext)
> data(sentiments)
> nrc <- subset(sentiments, sentiments$lexicon == "nrc")
```

Result

```
> tail(nrc)
# A tibble: 6 × 4
     word    sentiment lexicon score
    <chr>        <chr>   <chr> <int>
1 zealous        trust     nrc    NA
2    zest anticipation     nrc    NA
3    zest          joy     nrc    NA
4    zest     positive     nrc    NA
5    zest        trust     nrc    NA
6     zip     negative     nrc    NA
```

Huckleberry Finn


```
> tidy_huck
# A tibble: 55,198 x 3
   document          term count
      <chr>         <chr> <dbl>
 1        1          finn     1
 2        1  huckleberry     1
 3        3           ago     1
 4        3         fifty     1
 5        3         forty     1
 6        3  mississippi     1
 7        3         scene     1
 8        3           the     1
 9        3          time     1
10        3        valley     1
# ... with 55,188 more rows
```

Huck Finn Joined to AFINN

```
> huck_finn_join <- tidy_huck %>%
+   inner_join(afinn, by = c("term" = "word"))


> huck_finn_join
# A tibble: 4,849 x 6
   document       term count sentiment lexicon score
      <chr>      <chr> <dbl>     <chr>   <chr> <int>
 1       11 adventures     1      <NA>   AFINN     2
 2       11     matter     1      <NA>   AFINN     1
 3       14       lied     1      <NA>   AFINN    -2
 4       17       true     1      <NA>   AFINN     2
 5       20        hid     1      <NA>   AFINN    -1
 6       20       rich     1      <NA>   AFINN     2
# ... with 4,843 more rows
```

Using summarize()

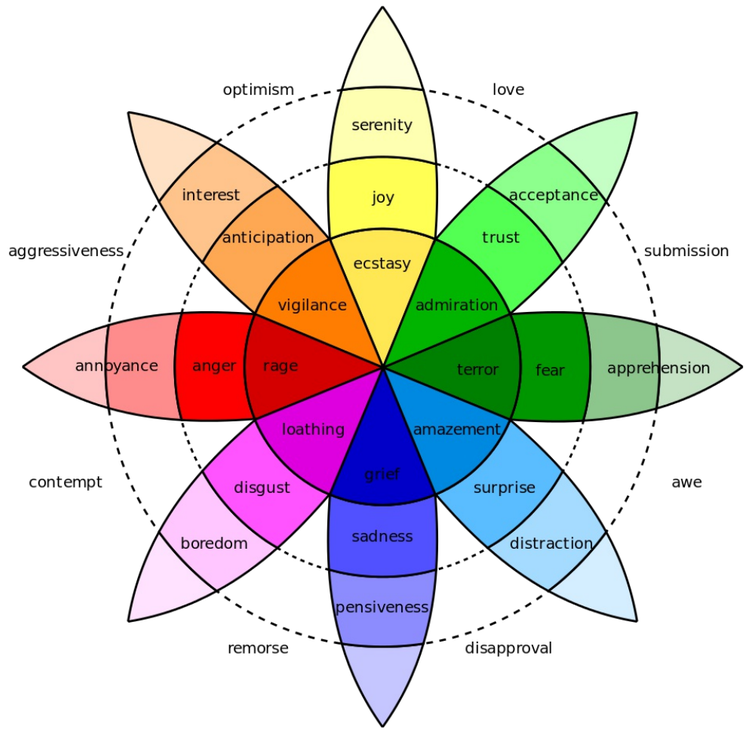
```
> sample_df
# A tibble: 2 x 6
  document  term count sentiment lexicon score
     <dbl> <chr> <dbl>     <chr>   <chr> <dbl>
1       22 judge     1      <NA>   AFINN    -3
2       22  took     1      <NA>   AFINN     1


> sample_df %>%
+   group_by(document) %>%
+   summarize(total_score = sum(score))
# A tibble: 1 x 2
  document total_score
     <dbl>       <dbl>
1       22          -2
```

Using filter()

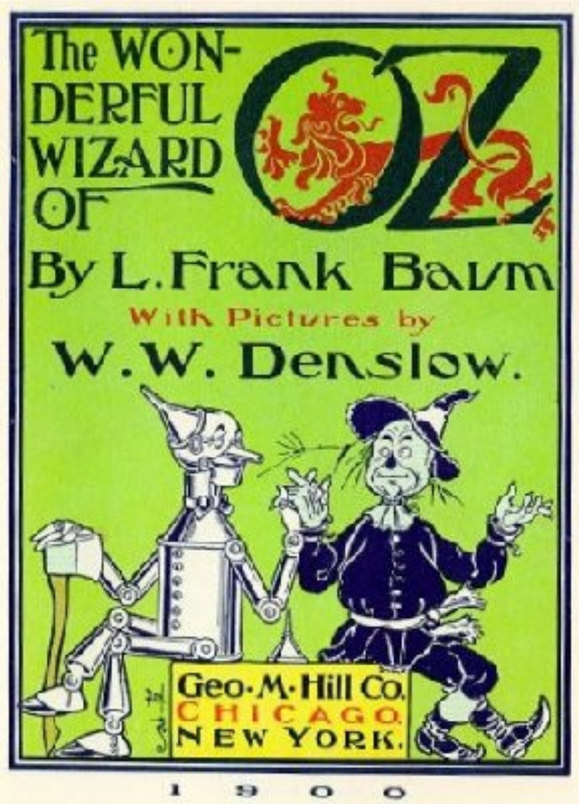
```
> filter(huck_finn_join, document == 20)
# A tibble: 2 x 6
  document  term count sentiment lexicon score
     <chr> <chr> <dbl>     <chr>   <chr> <int>
1       20   hid     1      <NA>   AFINN    -1
2       20  rich     1      <NA>   AFINN     2
```

Plutchik & NRC




```
> filter(sentiments,
+        lexicon == "nrc")
# A tibble: 13,901 x 4
          word sentiment lexicon score
         <chr>     <chr>   <chr> <int>
 1       abacus     trust     nrc    NA
 2      abandon      fear     nrc    NA
 3      abandon  negative     nrc    NA
 4      abandon   sadness     nrc    NA
 5    abandoned     anger     nrc    NA
 6    abandoned      fear     nrc    NA
 7    abandoned  negative     nrc    NA
 8    abandoned   sadness     nrc    NA
 9 abandonment     anger     nrc    NA
10 abandonment      fear     nrc    NA
# ... with 13,891 more rows
```

The Wonderful Wizard of NRC




```
> oz
# A tibble: 19,007 x 3
   document          term count
      <chr>         <chr> <dbl>
 1        1           the     1
 2        1        wizard     1
 3        1     wonderful     1
 4        6          baum     1
 5        6         frank     1
 6       10      contents     1
 7       12 introduction     1
 8       13       cyclone     1
 9       13           the     1
10       14       council     1
# ... with 18,997 more rows
```

%in% operator

```
> x <- c("text", "mining", "python")

> y <- c("text", "tm", "qdap", "R", "mining")

> x %in% y
[1] TRUE TRUE FALSE

> y %in% x
[1] TRUE FALSE FALSE FALSE TRUE
```

SENTIMENT ANALYSIS IN R

Let's practice!