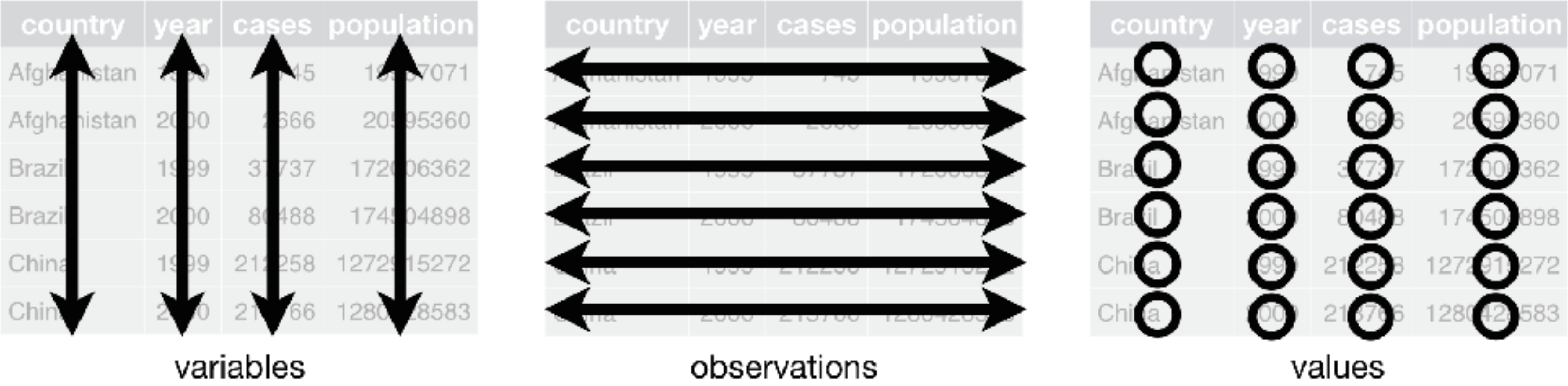
Natural Language Processing

STOR 390
4/18/17

Kurt Vonnegut on the Shapes of Stories

https://www.youtube.com/watch?v=oP3c1h8v2ZQ

We know how to work with **tidy data**



country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

observations

values

We know how to work with **tidy data**

Regression

linear model, polynomial terms

Classification

K-nearest-neighbors, SVM

Clustering

K-means

Unstructured data: not all data is tidy

Networks

Text

Images

Network data

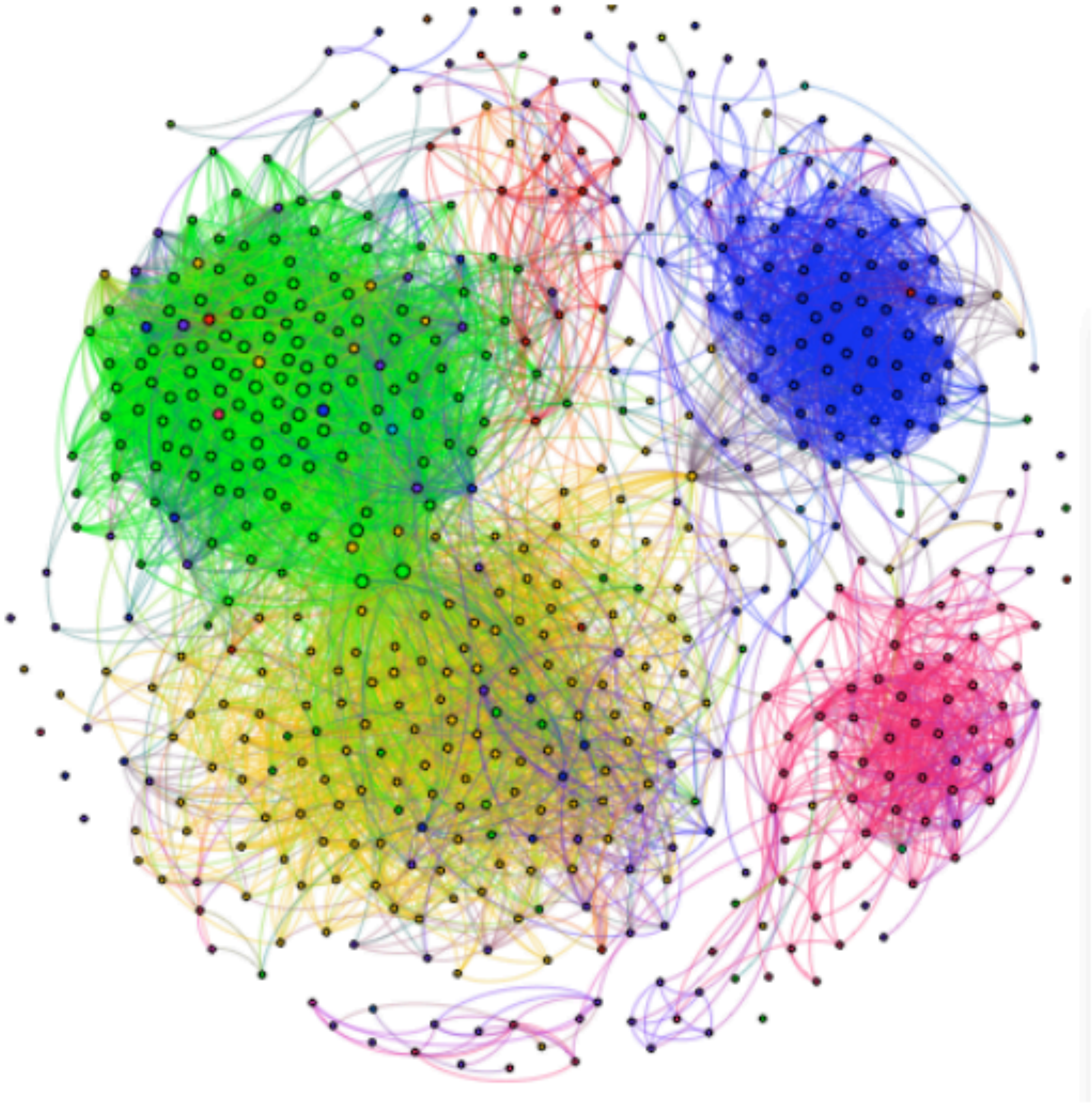
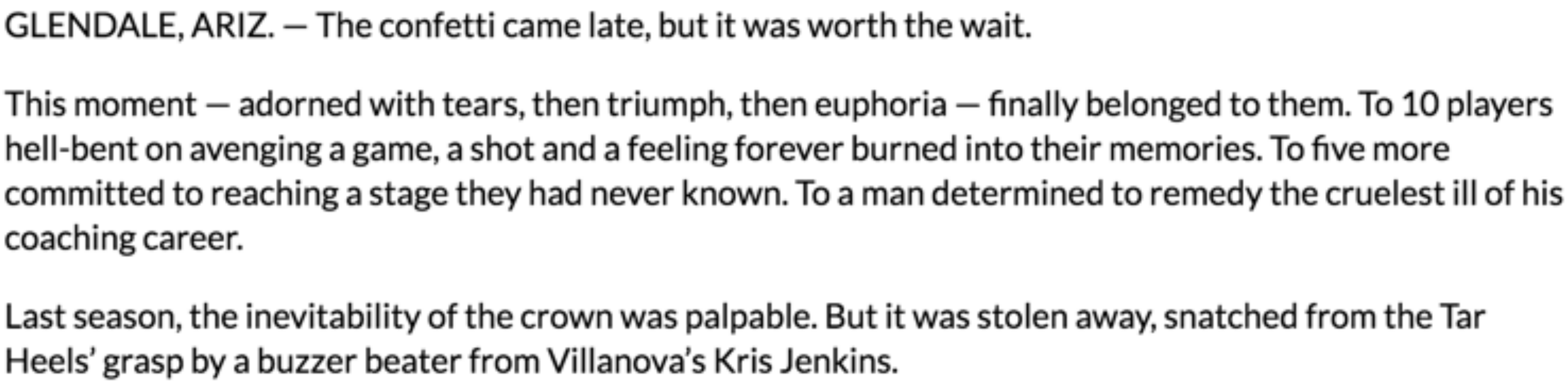


Image data

http://dogtime.com/puppies/255-puppies

http://www.dailytarheel.com/article/2017/04/a-title-to-remember-north-carolina-wins-its-sixth-ncaa-championship

Text data



GLENDALE, ARIZ. — The confetti came late, but it was worth the wait.

This moment — adorned with tears, then triumph, then euphoria — finally belonged to them. To 10 players hell-bent on avenging a game, a shot and a feeling forever burned into their memories. To five more committed to reaching a stage they had never known. To a man determined to remedy the cruelest ill of his coaching career.

Last season, the inevitability of the crown was palpable. But it was stolen away, snatched from the Tar Heels' grasp by a buzzer beater from Villanova's Kris Jenkins.

https://emeraldcitybookreview.com/2014/06/beautiful-books-picturing-jane-austen_20.html

Unstructured ≠ no structure

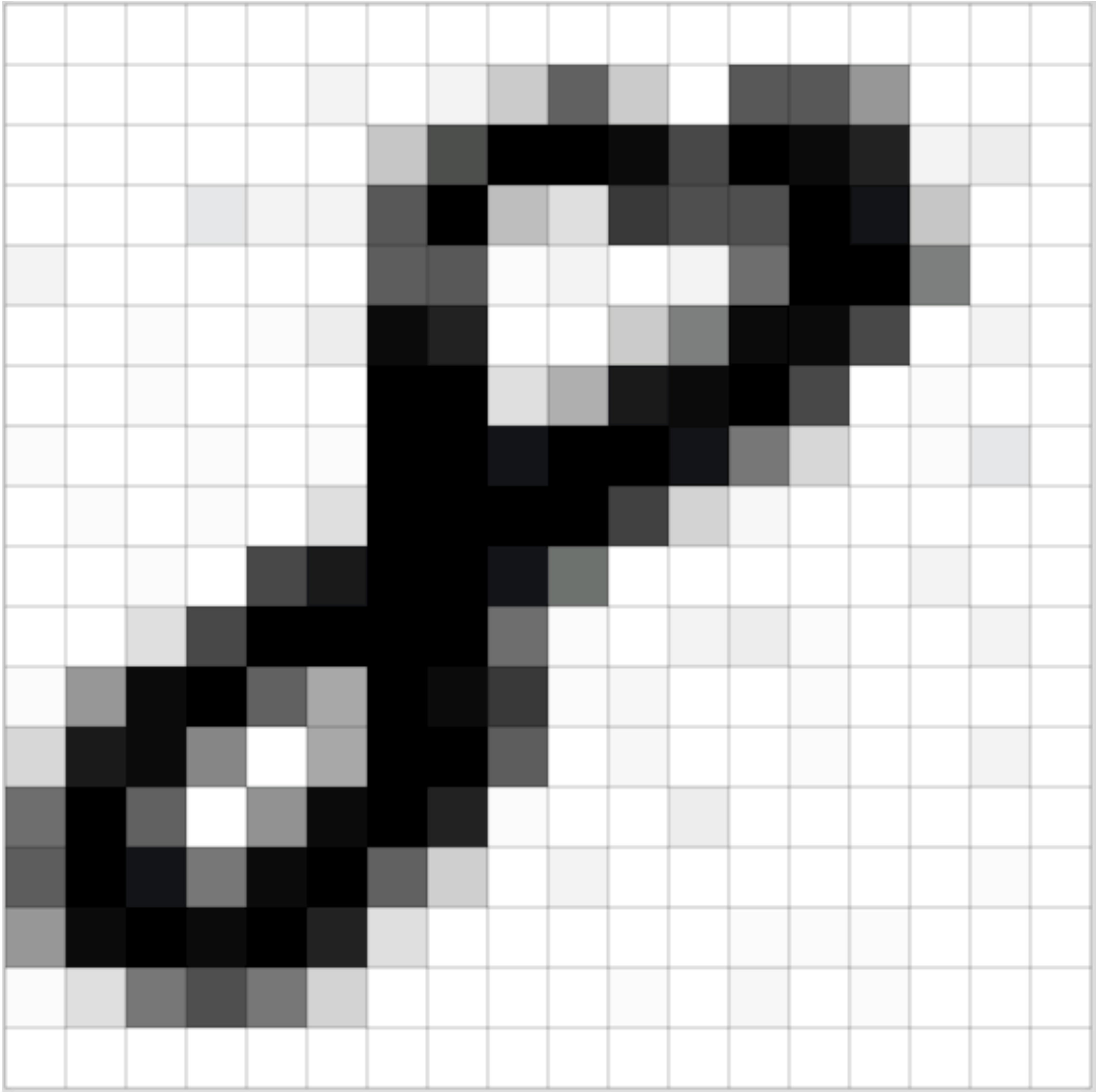
Two strategies

Invent new tools

PageRank

Turn it into tidy data

Images are numbers

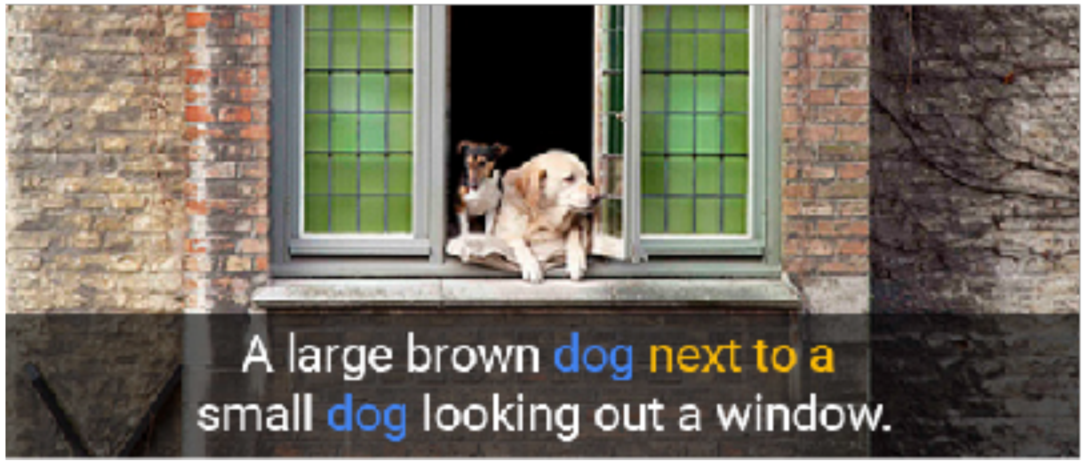


https://medium.com/@ageitgey/machine-learning-is-fun-part-3-deep-learning-and-convolutional-neural-networks-f40359318721

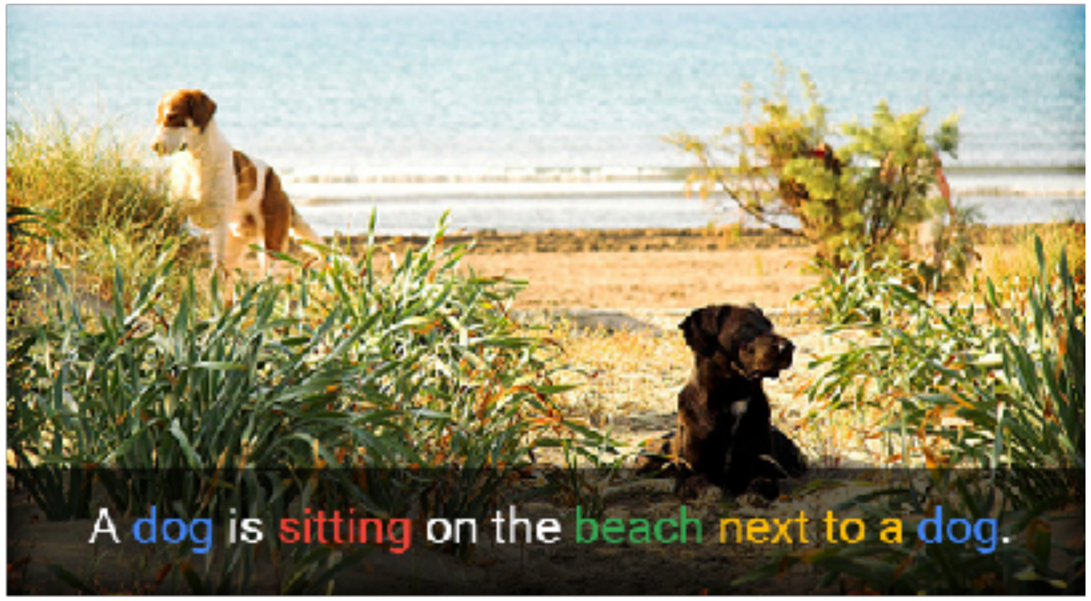
Human captions from the training set

A cute little dog sitting in a heart drawn on a sandy beach.

A dog walking next to a little dog on top of a beach.


A large brown dog next to a small dog looking out a window.

Automatically captioned


A dog is sitting on the beach next to a dog.

https://research.googleblog.com/2016/09/show-and-tell-image-captioning-open.html

Text data

One document = string of words

Corpus = collection of documents

“**A token is a meaningful unit of text**, most often a word, that we are interested in using for further analysis, and tokenization is the process of splitting text into tokens.”

—Text Mining with R

Tokenization turns text into tidy format

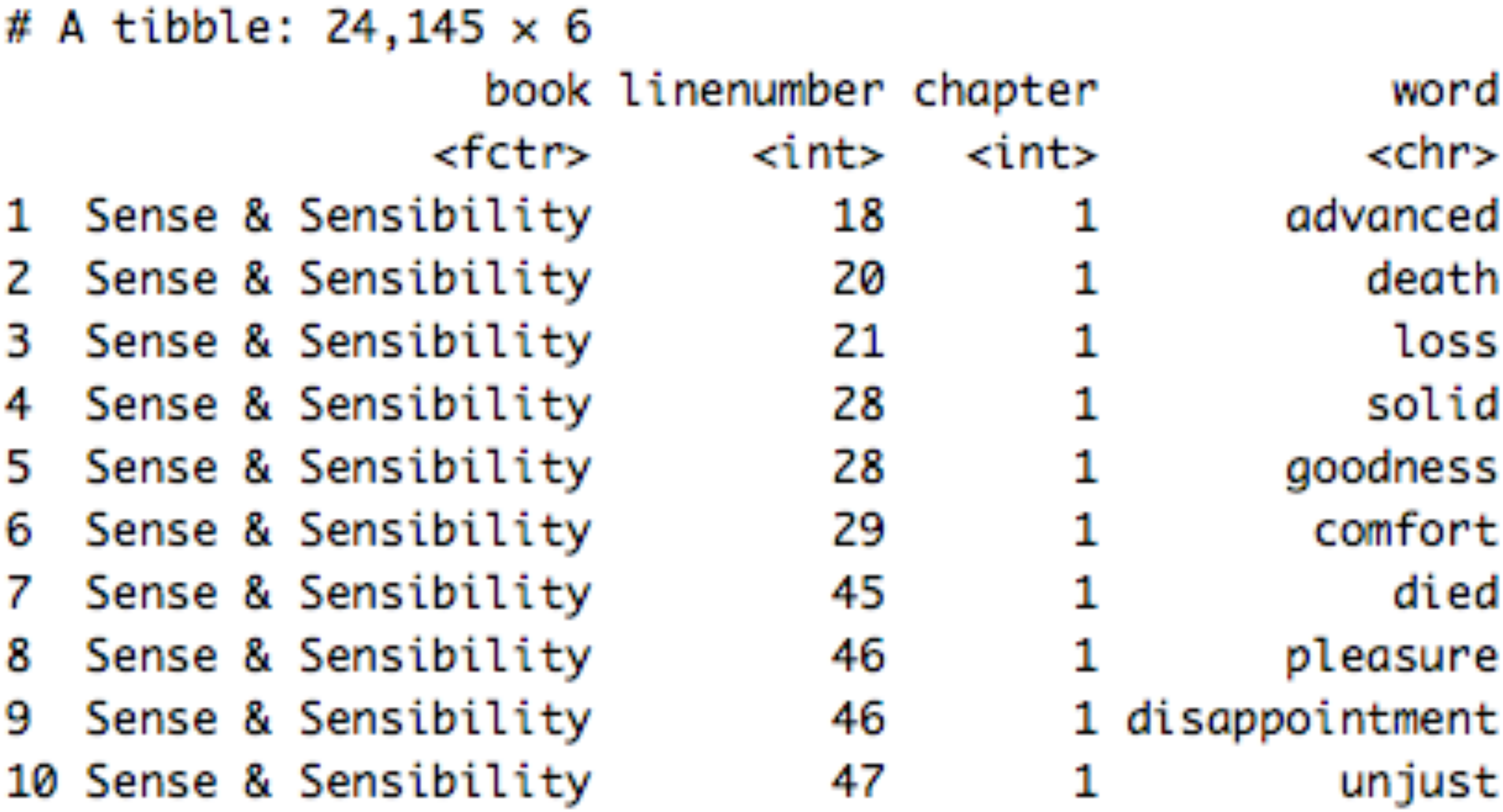
Word

Sentence

Paragraph

Chapter

Jane Austen’s books tokenized by word



```
# A tibble: 24,145 × 6
                 book linenumber chapter           word
               <fctr>      <int>   <int>          <chr>
1  Sense & Sensibility         18       1       advanced
2  Sense & Sensibility         20       1          death
3  Sense & Sensibility         21       1           loss
4  Sense & Sensibility         28       1          solid
5  Sense & Sensibility         28       1       goodness
6  Sense & Sensibility         29       1        comfort
7  Sense & Sensibility         45       1           died
8  Sense & Sensibility         46       1       pleasure
9  Sense & Sensibility         46       1 disappointment
10 Sense & Sensibility         47       1         unjust
```

Make text **lower case**

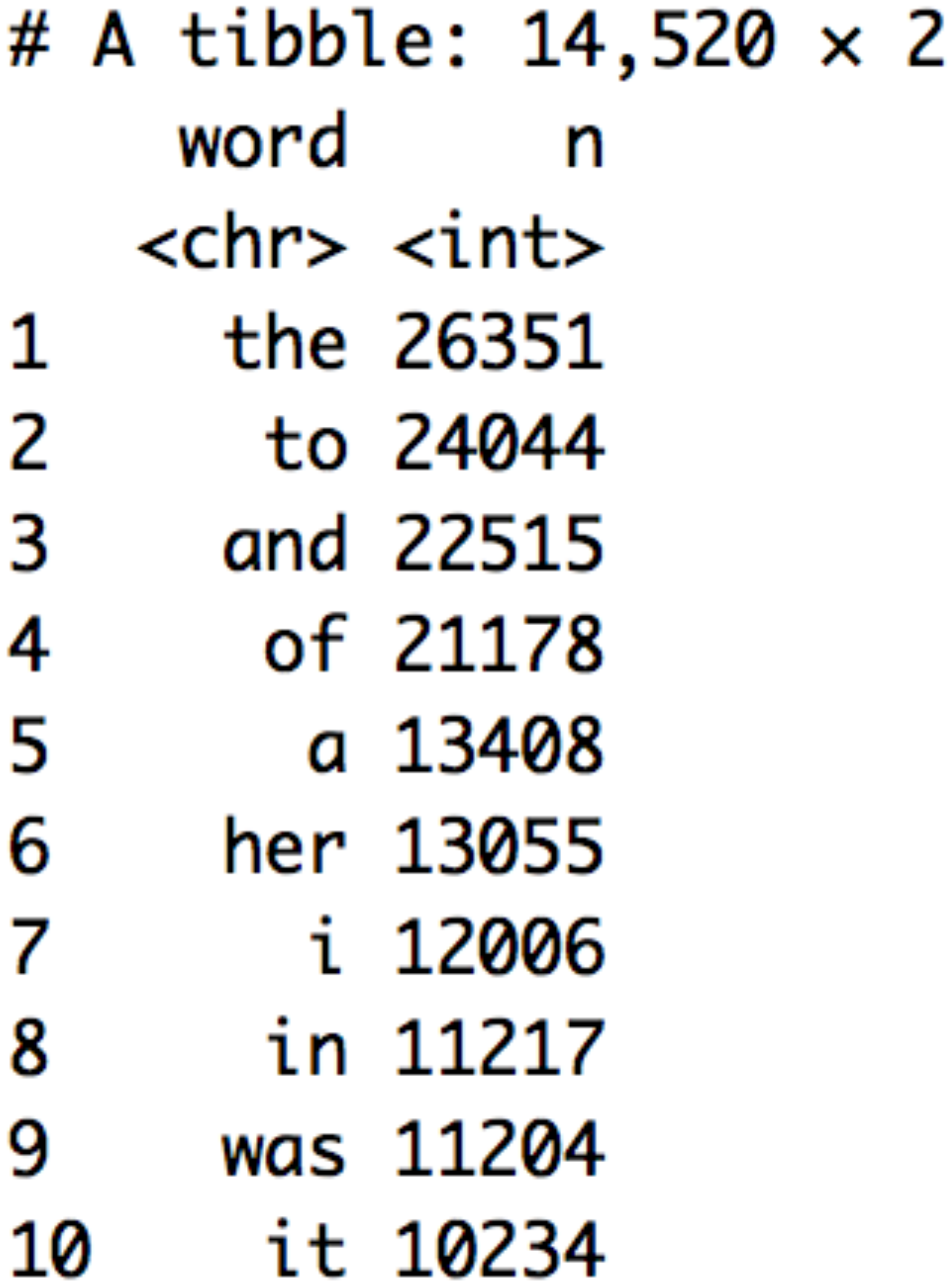
Make words more comparable

Door —> door

Tokenization loses information

Ignores word order

Most frequently appearing words



```
# A tibble: 14,520 × 2
    word     n
   <chr> <int>
1    the 26351
2     to 24044
3    and 22515
4     of 21178
5      a 13408
6    her 13055
7      i 12006
8     in 11217
9    was 11204
10    it 10234
```

Remove **stop words**

Commonly occurring words

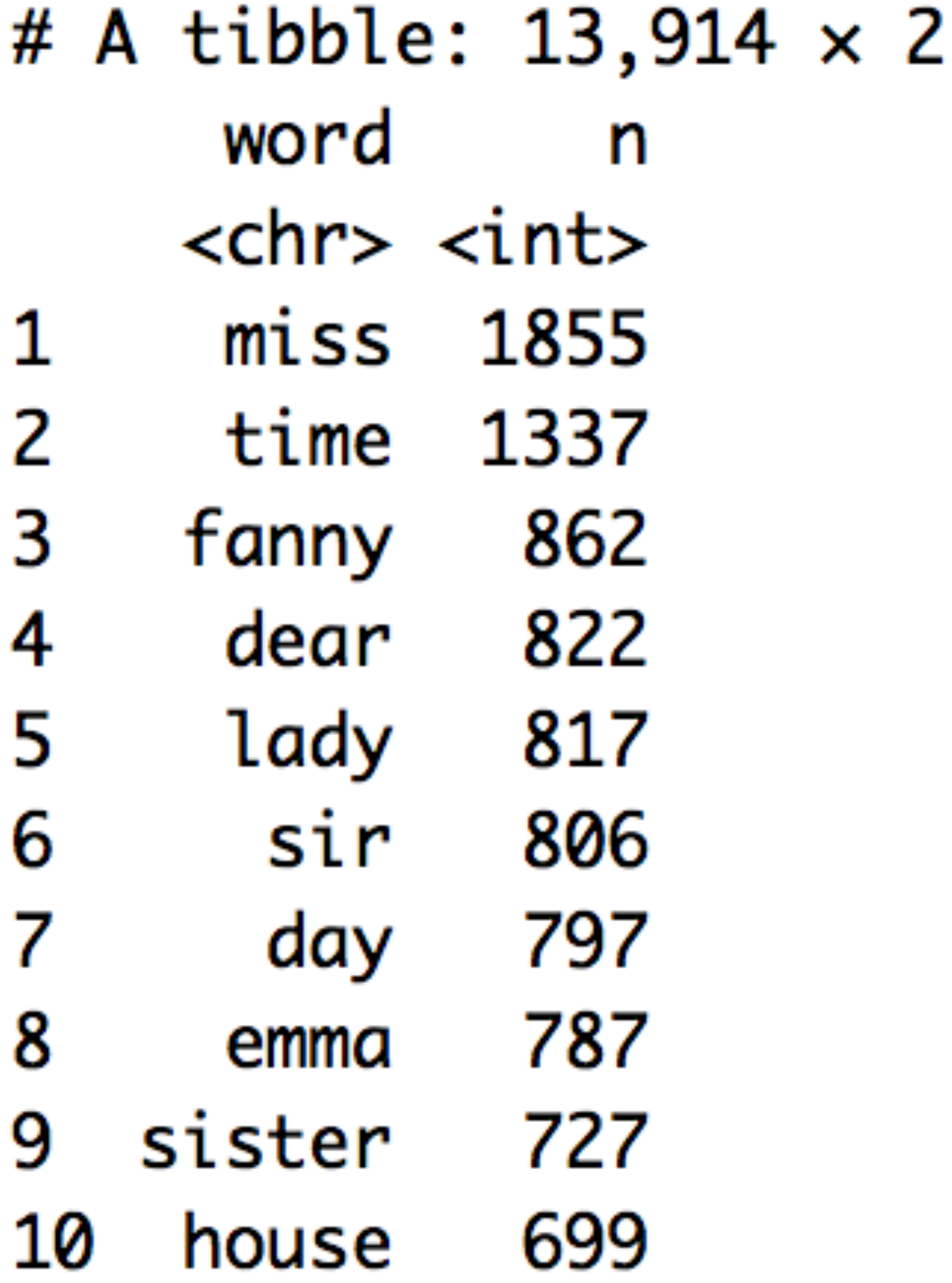
the

to

and

Hand code a list of words

Most frequently occurring words (no stop words)



```
# A tibble: 13,914 × 2
     word     n
    <chr> <int>
1    miss  1855
2    time  1337
3   fanny   862
4    dear   822
5    lady   817
6     sir   806
7     day   797
8    emma   787
9  sister   727
10  house   699
```

Sentiment analysis attempts to quantify emotional content

Assign each word an emotional value

positive/negative

trust, fear, sadness, anger, surprise, disgust, joy, anticipation”

-5, -4, … 4, 5

There are precompiled lexicons

Hand coded

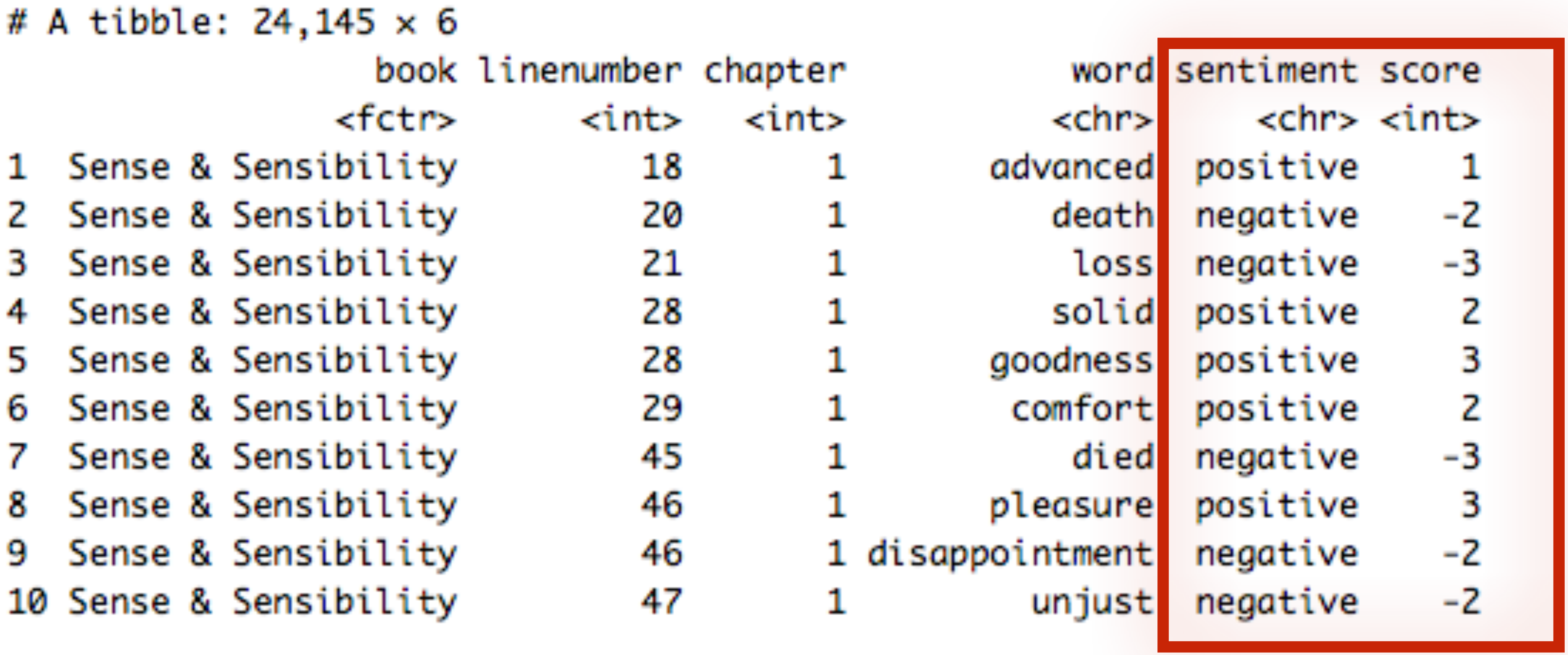
Crowdsourced

Amazon turk

Online reviews

Yelp

Assign each word a sentiment



```
# A tibble: 24,145 × 6
                  book linenumber chapter           word sentiment score
                <fctr>      <int>   <int>          <chr>     <chr> <int>
1  Sense & Sensibility         18       1       advanced  positive     1
2  Sense & Sensibility         20       1          death  negative    -2
3  Sense & Sensibility         21       1           loss  negative    -3
4  Sense & Sensibility         28       1          solid  positive     2
5  Sense & Sensibility         28       1       goodness  positive     3
6  Sense & Sensibility         29       1        comfort  positive     2
7  Sense & Sensibility         45       1           died  negative    -3
8  Sense & Sensibility         46       1       pleasure  positive     3
9  Sense & Sensibility         46       1 disappointment  negative    -2
10 Sense & Sensibility         47       1         unjust  negative    -2
```

Sentiment analysis is noisy

Sentiment analysis is noisy

Lexicons may not generalize

Unigrams

no good

Context

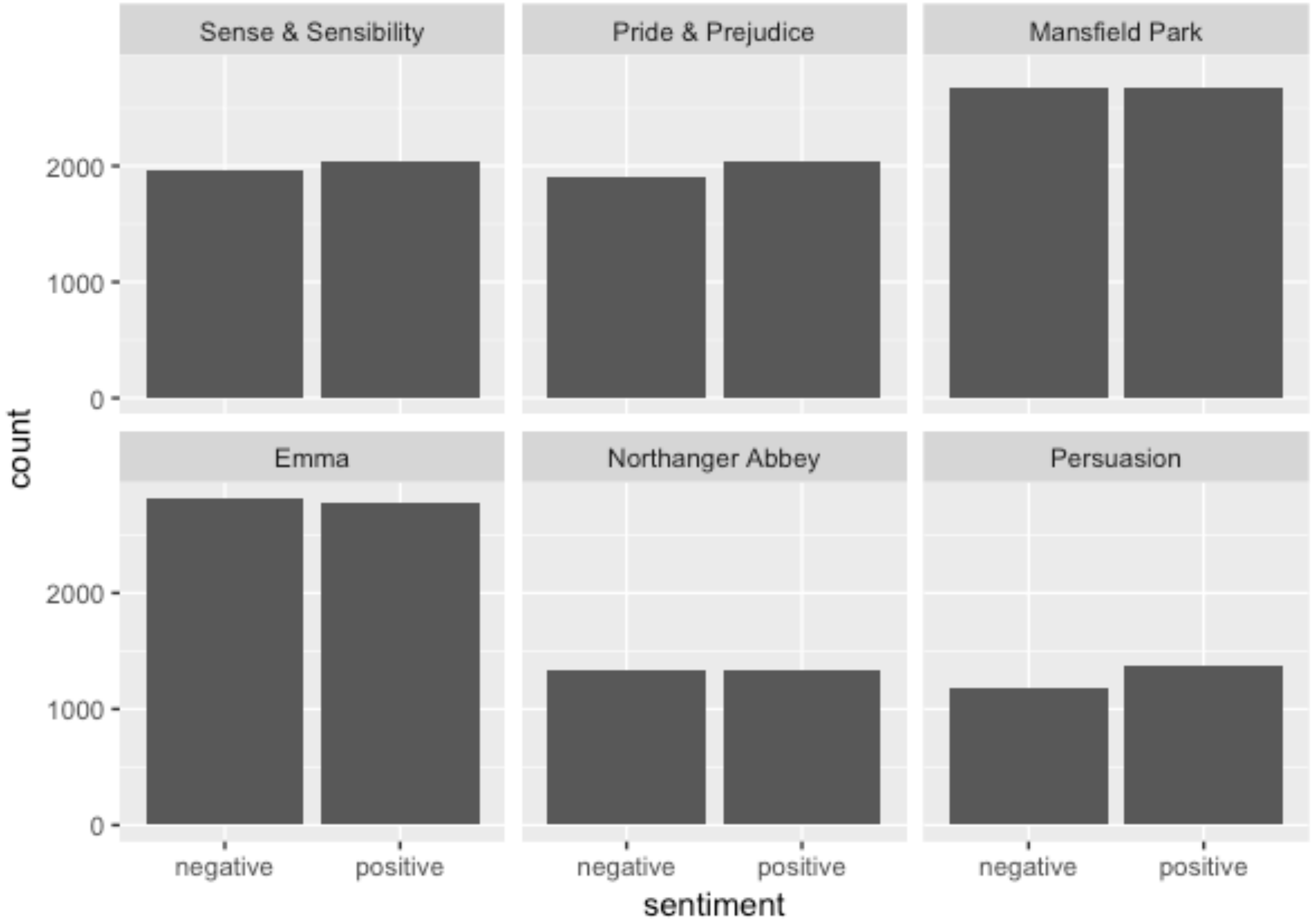
Sentiment analysis is noisy

Statistics is so much fun

vs.

Statistics is so much fun

Jane Austen novels are fairly balanced



Different ways to quantify “time"

chapter

paragraph

line

sentence

Different ways to quantify “time"

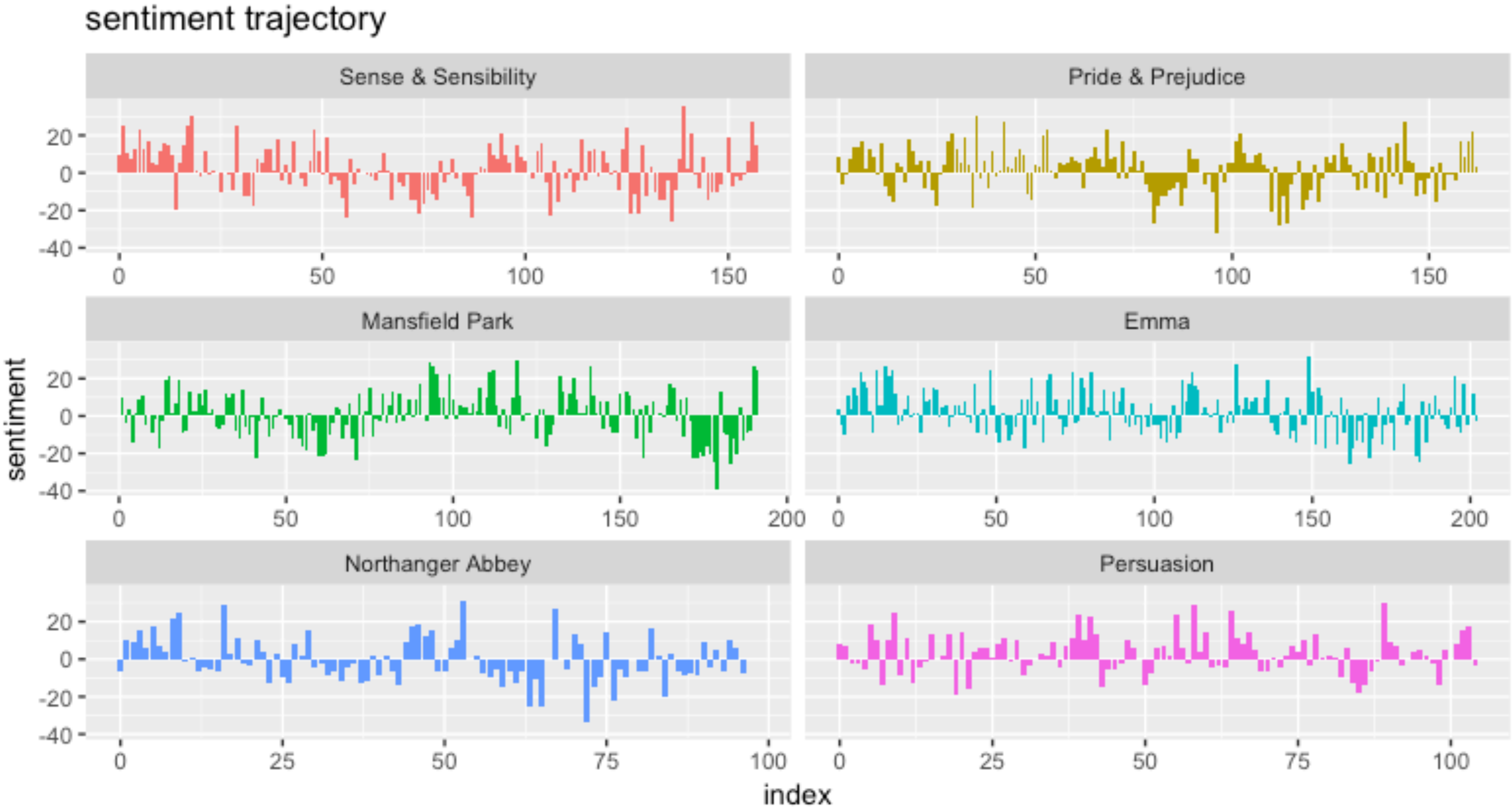
chapter

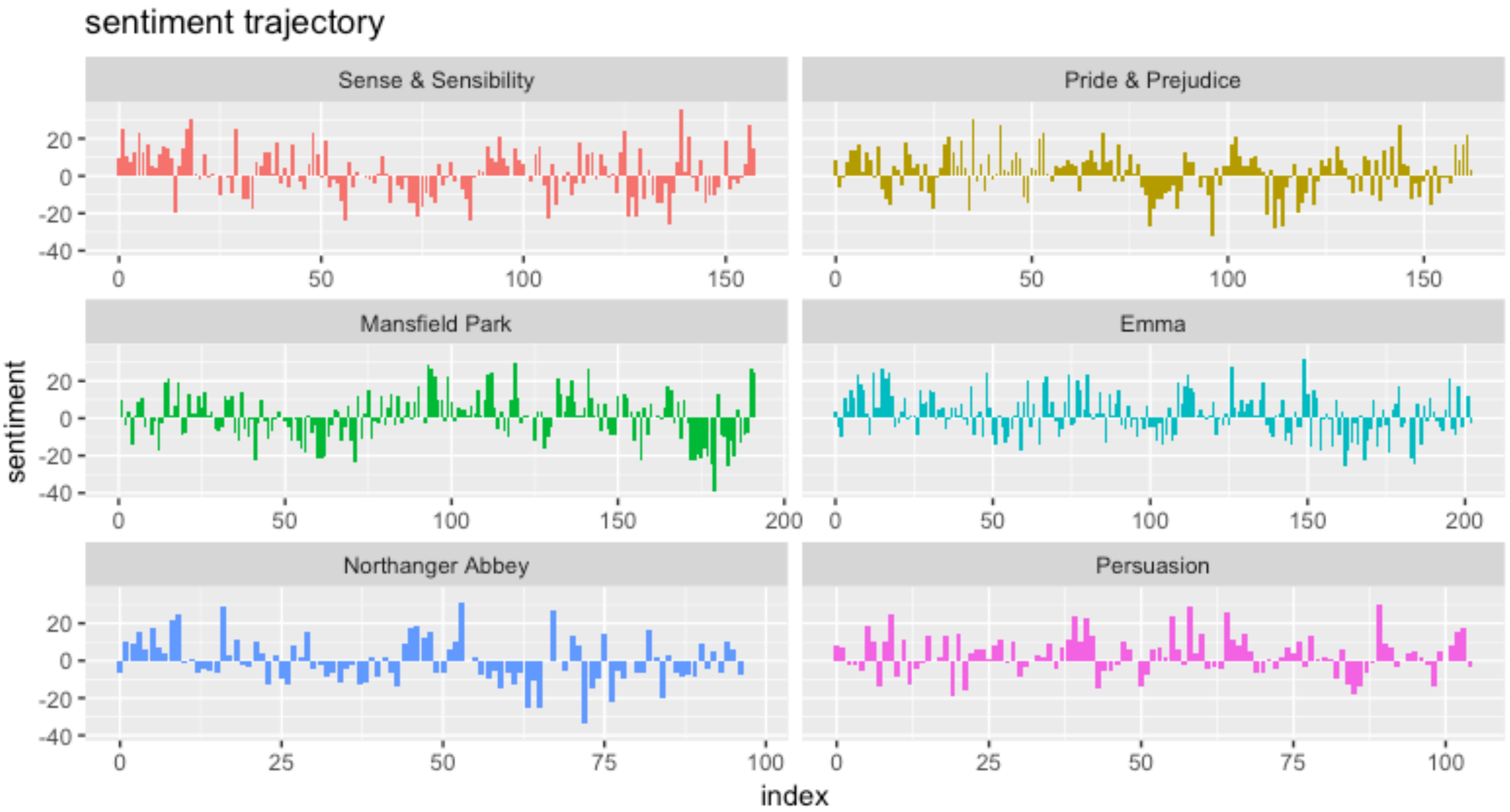
paragraph

line

sentence

we choose
one unit of time = 80 lines


sentiment trajectory
Sense & Sensibility
Pride & Prejudice
Mansfield Park
Emma
Northanger Abbey
Persuasion
sentiment
index
20
0
-20
-40
0
50
100
150
200
25
75

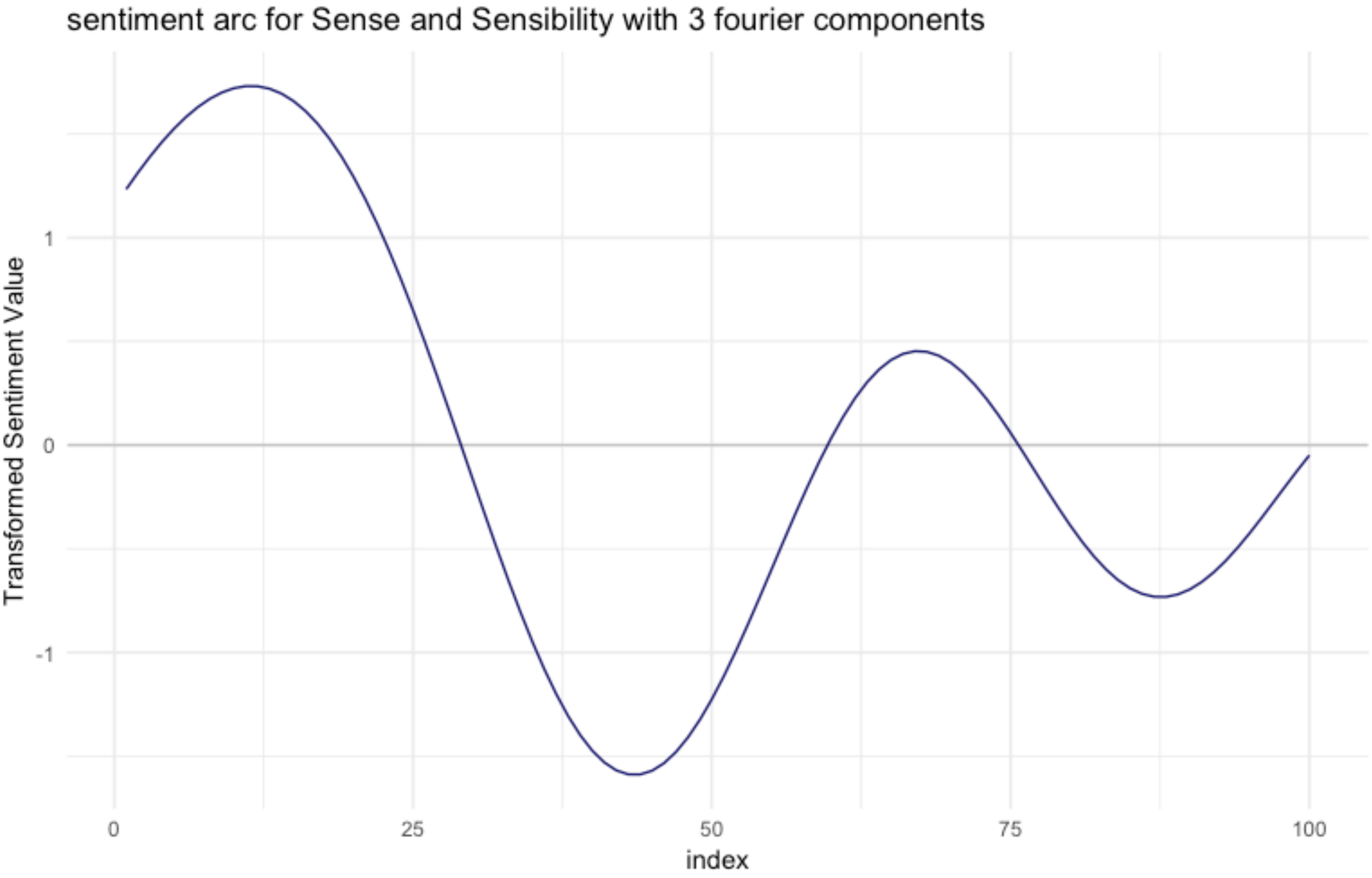


index = line number %/% 80

sentiment = (# positive words) - (# negative words)

Smooth time series with a low band pass filter

http://www.matthewjockers.net/2015/02/02/syuzhet/


sentiment arc for Sense and Sensibility with 3 fourier components
Transformed Sentiment Value
1
0
-1
0
25
50
75
100
index

References

Text Mining with R
http://tidytextmining.com/

Revealing Sentiment and Plot Arcs with the Syuzhet Package
http://www.matthewjockers.net/2015/02/02/syuzhet/