Tidy og Sentiment

Tekstanalyse

Nrc

word	sentiment
abacus	trust
abandon	fear
abandon	negative
abandon	sadness
abandoned	anger
abandoned	fear
abandoned	negative
abandoned	sadness

Afinn

word	value
abandon	-2
abandoned	-2
abandons	-2
abducted	-2
abduction	-2
abductions	-2

Bing

	word	sentiment
1	2-faces	negative
2	abnormal	negative
3	abolish	negative
4	abominable	negative
5	abominably	negative
6	abominate	negative
7	abomination	negative
8	abort	negative

```
> summary(afinnsent)
     word                value
 Length:2477        Min.   :-5.0000
 Class :character   1st Qu.:-2.0000
 Mode  :character   Median :-2.0000
                    Mean   :-0.5894
                    3rd Qu.: 2.0000
                    Max.   : 5.0000
```

```
> summary(bingst)
     word              sentiment
 Length:6786        Length:6786
 Class :character   Class :charac
 Mode  :character   Mode  :charac
```

```
> summary(nrc)
     word              sentiment
 Length:13875       Length:13875
 Class :character   Class :character
 Mode  :character   Mode  :character
> unique(nrc$sentiment)
 [1] "trust"        "fear"         "negative"     "sadness"      "anger"
 [6] "surprise"     "positive"     "disgust"      "joy"          "anticipation"
```

```
tidy_books <- austen_books() %>%
  group_by(book) %>%
  mutate(linenumber = row_number(),
         chapter = cumsum(str_detect(text, regex("^chapter [\\divxlc]",
                                                 ignore_case = TRUE)))) %>%
  ungroup() %>%
  unnest_tokens(word, text)
```

Filter

	text	book
1	SENSE AND SENSIBILITY	Sense & Sensibility
2		Sense & Sensibility
3	by Jane Austen	Sense & Sensibility
4		Sense & Sensibility
5	(1811)	Sense & Sensibility
6		Sense & Sensibility
7		Sense & Sensibility
8		Sense & Sensibility
9		Sense & Sensibility
10	CHAPTER 1	Sense & Sensibility
11		Sense & Sensibility
12		Sense & Sensibility
13	The family of Dashwood had long been settled in Sus...	Sense & Sensibility
14	was large, and their residence was at Norland Park, in...	Sense & Sensibility

book	linenumber	chapter	word
Sense & Sensibility	1	0	sense
Sense & Sensibility	1	0	and
Sense & Sensibility	1	0	sensibility
Sense & Sensibility	3	0	by
Sense & Sensibility	3	0	jane
Sense & Sensibility	3	0	austen
Sense & Sensibility	5	0	1811
Sense & Sensibility	10	1	chapter
Sense & Sensibility	10	1	1

10	Sense & Sensibility	13	1	the
11	Sense & Sensibility	13	1	family
12	Sense & Sensibility	13	1	of
13	Sense & Sensibility	13	1	dashwood
14	Sense & Sensibility	13	1	had
15	Sense & Sensibility	13	1	long
16	Sense & Sensibility	13	1	been

Joy in "Emma"

```
nrcjoy <- get_sentiments("nrc") %>%
  filter(sentiment == "joy")
```

	word	sentiment
1	abacus	trust
2	abandon	fear
3	abandon	nega
4	abandon	sadn
5	abandoned	anger
6	abandoned	fear
7	abandoned	negative

```
22 coutjoy=tidy_books %>%
23   filter(book == "Emma") %>%
24   inner_join(nrcjoy) %>%
25   count(word, sort = TRUE)
```

book	linenumber	chapter	word	sentiment
Emma	16	1	happy	joy
Emma	16	1	blessings	joy
Emma	21	1	marriage	joy
Emma	22	1	mother	joy
Emma	24	1	excellent	joy
Emma	25	1	mother	joy

word	n
good	359
friend	166
hope	143
happy	125
love	117

Changes of sentiment I

Sense & Sensibility	16	1	respectable
Sense & Sensibility	16	1	a
Sense & Sensibility	16	1	manner
Sense & Sensibility	16	1	as
Sense & Sensibility	16	1	to
Sense & Sensibility	16	1	engage
Sense & Sensibility	16	1	the
Sense & Sensibility	16	1	general
Sense & Sensibility	16	1	good

	word	sentiment
1	abound	positive
2	abounds	positive
3	abundance	positive
4	abundant	positive
5	accessable	positive

book	linenumber	chapter	word	sentiment
Sense & Sensibility	16	1	respectable	positive
Sense & Sensibility	16	1	good	positive
Sense & Sensibility	18	1	advanced	positive
Sense & Sensibility	20	1	death	negative
Sense & Sensibility	20	1	great	positive

```
#80lines seqs
js2 <- tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(book, seqindex = linenumber %/% 80, sentiment)
```

book	seqindex	sentiment	n
Sense & Sensibility	0	negative	16
Sense & Sensibility	0	positive	32
Sense & Sensibility	1	negative	19
Sense & Sensibility	1	positive	53
Sense & Sensibility	2	negative	12

Changes of sentiment II

book	seqindex	sentiment	n
Sense & Sensibility	0	negative	16
Sense & Sensibility	0	positive	32
Sense & Sensibility	1	negative	19
Sense & Sensibility	1	positive	53
Sense & Sensibility	2	negative	12

```
js3 <- tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(book, seqindex = linenumber %/% 80, sentiment) %>%
  spread(sentiment, n, fill = 0)
```

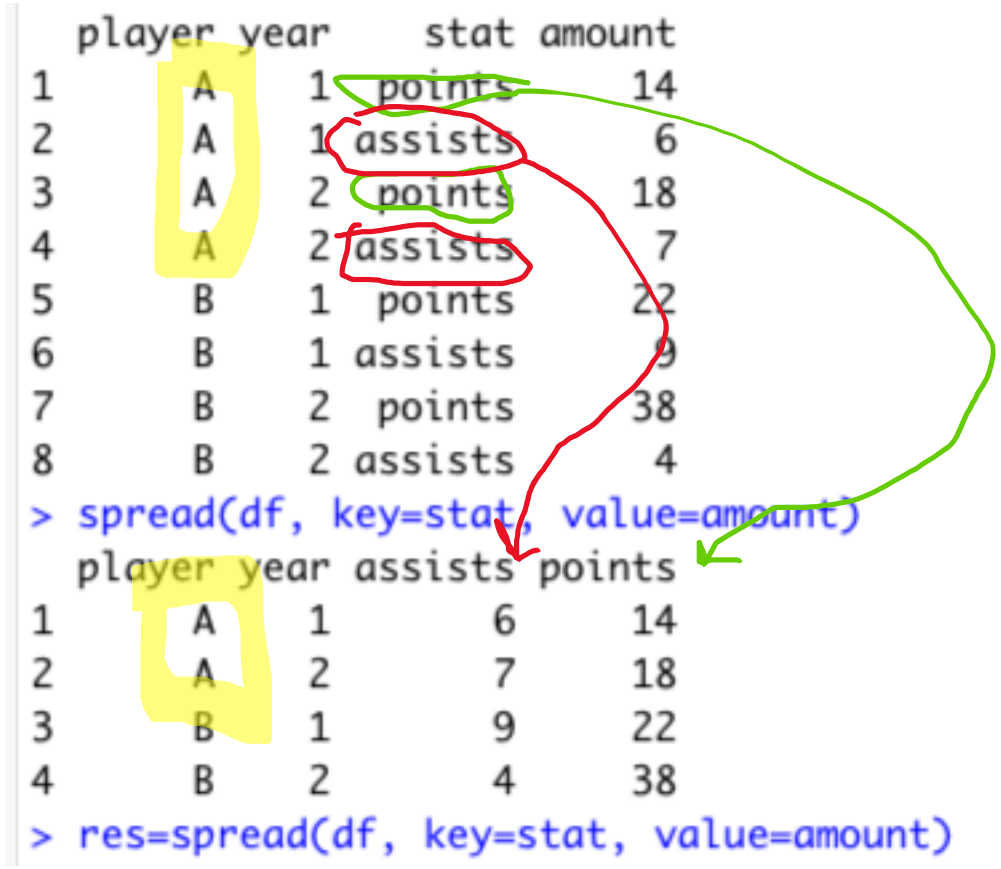
spread()

book	seqindex	negative	positive
Sense & Sensibility	0	16	32
Sense & Sensibility	1	19	53
Sense & Sensibility	2	12	31
Sense & Sensibility	3	15	31
Sense & Sensibility	4	16	34

book	seqindex	negative	positive	sentiment
Sense & Sensibility	0	16	32	16
Sense & Sensibility	1	19	53	34
Sense & Sensibility	2	12	31	19
Sense & Sensibility	3	15	31	16
Sense & Sensibility	4	16	34	18
Sense & Sensibility	5	16	51	35

sentiment=neg-pos

```
js4 <- tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(book, seqindex = linenumber %/% 80, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
```


player year stat amount
1 A 1 points 14
2 A 1 assists 6
3 A 2 points 18
4 A 2 assists 7
5 B 1 points 22
6 B 1 assists 9
7 B 2 points 38
8 B 2 assists 4
> spread(df, key=stat, value=amount)
player year assists points
1 A 1 6 14
2 A 2 7 18
3 B 1 9 22
4 B 2 4 38
> res=spread(df, key=stat, value=amount)

Check sentlibs I

22	Pride & Prejudice	11	1	a
23	Pride & Prejudice	11	1	good
24	Pride & Prejudice	11	1	fortune
25	Pride & Prejudice	11	1	must
26	Pride & Prejudice	11	1	be
27	Pride & Prejudice	11	1	in
28	Pride & Prejudice	11	1	want
29	Pride & Prejudice	11	1	of
30	Pride & Prejudice	11	1	a
31	Pride & Prejudice	11	1	wife
32	Pride & Prejudice	13	1	however

want

join with afinn

	linenumber	chapter	word	value
ice	11	1	good	3
ice	11	1	want	1
ice	18	1	dear	2
Pride & Prejudice	26	1	no	-1
Pride & Prejudice	28	1	want	1
Pride & Prejudice	28	1	cried	-2

seqindex	sentiment	method
0	29	AFINN
1	0	AFINN
2	20	AFINN
3	30	AFINN
4	62	AFINN

```
group_by(seqindex = linenumber %/% 80) %>%
summarise(sentiment = sum(value)) %>%
mutate(method = "AFINN")
```

Check sentlibs II

22	Pride & Prejudice	11	1	a
23	Pride & Prejudice	11	1	good
24	Pride & Prejudice	11	1	fortune
25	Pride & Prejudice	11	1	must
26	Pride & Prejudice	11	1	be
27	Pride & Prejudice	11	1	in
28	Pride & Prejudice	11	1	want
29	Pride & Prejudice	11	1	of
30	Pride & Prejudice	11	1	a
31	Pride & Prejudice	11	1	wife
32	Pride & Prejudice	13	1	however

want

join with nrc and bing

linenumber	chapter	word	sentiment	method
1	0	pride	positive	nrc
1	0	prejudice	negative	nrc
10	1	truth	positive	nrc
10	1	possession	negative	nrc
11	1	good	positive	nrc

book	linenumber	chapter	word	sentiment	method
Pride & Prejudice	1	0	pride	positive	Bing et al.
Pride & Prejudice	1	0	prejudice	negative	Bing et al.
Pride & Prejudice	1	0	pride	positive	nrc
Pride & Prejudice	1	0	prejudice	negative	nrc
Pride & Prejudice	10	1	truth	positive	nrc
Pride & Prejudice	10	1	possession	negative	nrc

book	linenumber	chapter	word	sentiment	method
Pride & Prejudice	1	0	pride	positive	Bing et al.
			prejudice	negative	Bing et al.
			good	positive	Bing et al.
			fortune	positive	Bing et al.

Check sentlibs II

19623 obs

book	linenumber	chapter	word	sentiment	method
Pride & Prejudice	1	0	pride	positive	Bing et al.
Pride & Prejudice	1	0	prejudice	negative	Bing et al.
Pride & Prejudice	1	0	pride	positive	nrc
Pride & Prejudice	1	0	prejudice	negative	nrc
Pride & Prejudice	10	1	truth	positive	nrc
Pride & Prejudice	10	1	possession	negative	nrc

COUNT is short for group_by, summary and count=n()

count(method,index = linenumber %/%80,sentiment)

652 obs

method	index	sentiment	n
Bing et al.	0	negative	7
Bing et al.	0	positive	21
Bing et al.	1	negative	20
Bing et al.	1	positive	19
Bing et al.	2	negative	16
Bing et al.	2	positive	20

count(method)

2 obs

method	n
Bing et al.	8704
nrc	10989

4 obs

method	sentiment	n
Bing et al.	negative	3652
Bing et al.	positive	5052
nrc	negative	3656
nrc	positive	7333

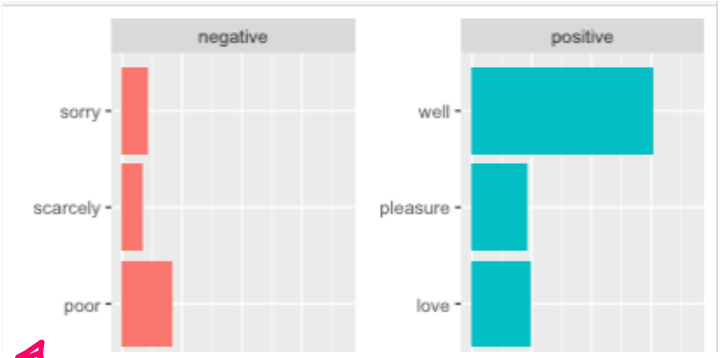
Check sentlibs III

method	index	negative	positive	sentiment
Bing et al.	0	7	21	14
NRC	0	10	30	20
Bing et al.	1	20	19	-1
NRC	1	17	33	16

Most common pos and negs

```
bing_word_counts1 <- bing_word_counts%>%
  group_by(sentiment) %>%
  top_n(10)
```

	word	sentiment	n
1	miss	negative	1855
2	well	positive	1523
3	good	positive	1380
4	great	positive	981
5	like	positive	725
5	better	positive	639




```
bb <- bing_word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  mutate(word = reorder(word, n))
```

reorder the factor-levels
by the count. For ggplot

Most common pos and negs

Stopwords

word	lexicon
a	SMART
a's	SMART
able	SMART
about	SMART
above	SMART
according	SMART
accordingly	SMART

Mywords

word	lexicon
miss	custom

rbind(stopw,mywords)

word	lexicon
miss	custom
a	onix
about	onix
above	onix

Tidy og Silvan

Most common pos and negs