- Tema nr. 2

-PSDT-

Student: Irimia Mihaela

Instrucțiuni

Aplicați demersul din cursul 2 (începând cu Sentiment Analysis- fără prezentarea lexicoanelor de sentimente) cu următoarele precizări:

- În loc de "joy" folosiți "trust"
- Secțiunile vor avea 50 rânduri și 100 de rânduri (în curs aveau 80 de rânduri) (prin urmare veți avea 2 analize: una pe 50 rânduri și una pe 100 de rânduri).
- NU mai comparați cele două lexicoane nrc și bing (vezi codurile de la "De ce există aceste diferențe? ")
- Reprezentați grafic cuvintele pozitive și negative
- NU utilizați custom stop words (cuvinte comune particularizate)
- Wordclouds: grupate după sentimente și sortate după sentimente (100 cuvinte)
- Pentru predările întârziate nota pleacă de la 7 în jos.

Rezolvare

- Vom trasforma textul celor 6 romane in Tidy Text folosind unnest_tokens().
- Creez doua coloane linenumber si chapter pentru a urmări din ce rând și capitol al cărții provine fiecare cuvânt, folosind group_by() și mutate().

```
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)

# avem 6 carti
tidy_books %>%
  distinct(book)
```

```
> tidy_books
# A tibble: 725,055 × 4
   book                linenumber chapter word
   <fct>                    <int>   <int> <chr>
 1 Sense & Sensibility          1       0 sense
 2 Sense & Sensibility          1       0 and
 3 Sense & Sensibility          1       0 sensibility
 4 Sense & Sensibility          3       0 by
 5 Sense & Sensibility          3       0 jane
 6 Sense & Sensibility          3       0 austen
 7 Sense & Sensibility          5       0 1811
 8 Sense & Sensibility         10       1 chapter
 9 Sense & Sensibility         10       1 1
10 Sense & Sensibility         13       1 the
# i 725,045 more rows
```

- folosim lexiconul NRC și filtram după cuvintele care exprimă încredere -trust- în romanul *'Emma'*.

```r
# cele mmai frecvente cuvinte utilizate in cele 6 carti
nrctrust <- get_sentiments("nrc") %>%
  filter(sentiment == "trust")
nrctrust

# cautam cuvantul trust in cartea Emma si sortam
tidy_books %>%
  filter(book == "Emma") %>%
  inner_join(nrctrust) %>%
  count(word, sort = TRUE)
```

```
# A tibble: 452 × 2
   word      n
   <chr>   <int>
 1 good      359
 2 frank     200
 3 father    168
 4 friend    166
 5 hope      143
 6 happy     125
 7 doubt      98
 8 word       94
```

Observam ca cele mai frecvente cuvinte utilizate care exprimă încredere (trust) în ***romanul Emma*** sunt: ***good, frank, father, friend, hope, happy*** etc.

Analiza pe 50 rânduri

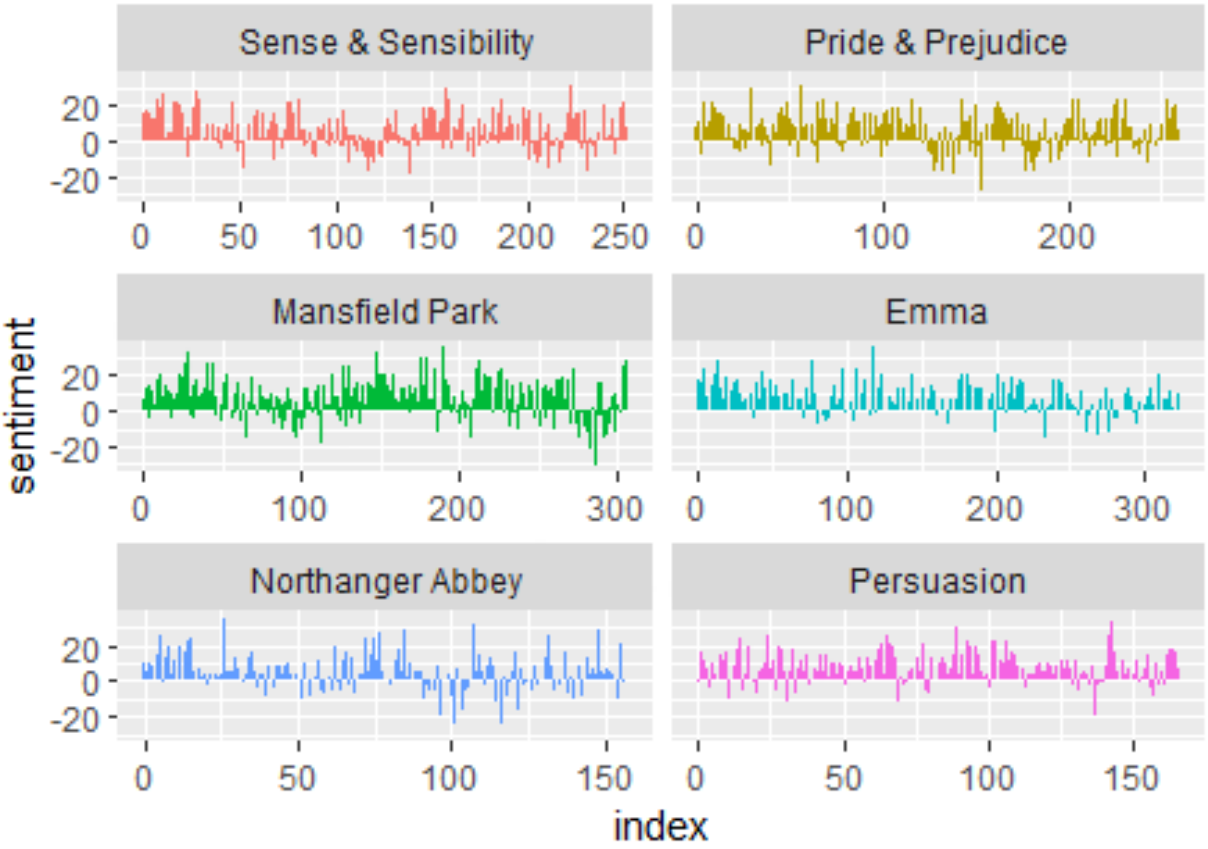
- atribuim un scor fiecarui cuvânt utilizând lexiconul bing și funcția inner_join().
- Dupa care contorizam câte cuvinte pozitive și negative sunt în fiecare roman.
- Definim un index ce numără secțiuni de text de 50 de rânduri pentru a stii unde ne aflăm în cadrul romanului, utilizând operatorul %/% 50.

```r
janeaustensentiment1 <- tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(book, index = linenumber %/% 50, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
janeaustensentiment1
```

```
# A tibble: 1,471 × 5
   book                index negative positive sentiment
   <fct>               <dbl>    <dbl>    <dbl>     <dbl>
 1 Sense & Sensibility     0        6       18        12
 2 Sense & Sensibility     1       14       29        15
 3 Sense & Sensibility     2       14       31        17
 4 Sense & Sensibility     3        8       23        15
 5 Sense & Sensibility     4        6       19        13
 6 Sense & Sensibility     5       10       16         6
 7 Sense & Sensibility     6        7       19        12
 8 Sense & Sensibility     7       13       26        13
 9 Sense & Sensibility     8       11       35        24
10 Sense & Sensibility     9       10       28        18
# i 1,461 more rows
```

- **Index** - este referinta numerica pentru fiecare sectiune de text de 50 de randuri din fiecare carte
- **Negative** – numarul de cuvinte cu conotatie negative identificate in fiecare sectiune de text
- **Positive** – numarul de cuvinte cu conotatie pozitiva identificate in fiecare sectiune de text
- **Sentiment** – scorul total de sentiment pe fiecare sectiune calculat ca diferenta dintre numarul de cuvinte positive si numarul de cuvinte negative (de exemplu: pentru prima linie observam ca scorul sentimentelor este 12, ceea ce indica faptul ca predomina un sentiment pozitiva – 18 cuvinte pozitive - 6 cuvinte negative)
- Reprezentăm grafic scorurile sentimentelor din fiecare roman –
    - pe axa OX avem indexul care ține evidența numărului de secțiuni a câte 50 de rânduri.
    - Pe axa OY sunt reprezentate scorurile sentimentelor care pot fi pozitive sau negative, dupa caz

```
ggplot(janeaustensentiment1, aes(index, sentiment, fill = book)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~book, ncol = 2, scales = "free_x")
```



Observam ca în fiecare roman avem atat cuvinte cu o conotatie pozitiva, cat si cuvinte cu o conontatie negative, insa, pentru majoirtatea cartilor analizate, sentimentul dominant este positiv. Romanele unde apar cele mai multe sentimente negative sunt ***Pride & Prejuduce***, ***Mansfield Park*** si ***Northanger Abbey***. In timp ce, la polul opus se situeaza romanul ***Persuasion***.

```
afinn1 <- pride_prejudice %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(index = linenumber %/% 50) %>%
  summarise(sentiment = sum(value)) %>%
  mutate(method = "AFINN")


bing_and_nrc1 <- bind_rows(
  pride_prejudice %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  pride_prejudice %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive",
                                         "negative"))) %>%
    mutate(method = "NRC")) %>%
  count(method, index = linenumber %/% 50, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
bing_and_nrc1

bind_rows(afinn1,
          bing_and_nrc1) %>%
  ggplot(aes(index, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")
```

- Comparăm cele trei dicționare de sentimente și examinăm modul în care sentimentul se schimbă în functie de firul narativ al romanului ***Pride & Prejudice***.

Am obținut sentimentul dominant (pozitiv sau negativ) pe fiecare secțiune din textul analizat, pentru cele 3 lexicoane. Prin unirea celor 3 lexicoane, putem vizualiza sau reprezenta grafic sentimentele dominante. Fiecare dintre aceste lexicoane măsoară sentimentul secțiunilor textului într-un mod distinct, iar graficul arată cum variază sentimentele dominante în funcție de index. In toate cele 3 grafice, sentimentul pozitiv pare sa fie dominant, existand si sectiuni in care sentimentul negativ este dominat, insa foarte putine. **AFINN** pare să aibă cea mai mare variație de scoruri, urmat de **Bing et al.**, respectiv **NRC** care pare să fie cel mai echilibrat, cu mai puține extreme. Sentimentele dominante positive sunt mai mari ca zero si sunt dominante si cele negative sunt mai mici ca zero, insa nu domina:

- ***Lexiconul AFFIN*** valorile sentimentului variază între aproximativ -20 și +60 si prezintă cel mai mic număr de cuvinte negative.
- ***Lexiconul Bing*** valorile sentimentului variază între aproximativ -20 și +30 si indica o prevalenta a cuvintelor positive in text.
- ***Lexiconul NRC*** valorile sentimentului variază între aproximativ -10 și +30 si are cel mai mare număr de cuvinte pozitive.

Analiza pe 100 rânduri

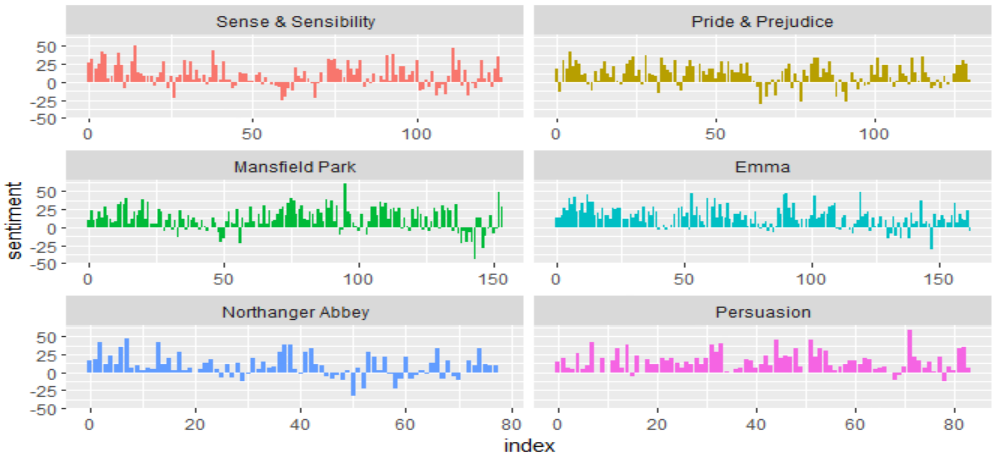
- atribuim un scor fiecarui cuvânt utilizând lexiconul bing și funcția inner_join().

- Dupa care numaram câte cuvinte pozitive și negative sunt în fiecare carte.
- Definim un index ce numără secțiuni de text de 100 de rânduri pentru a stii unde ne aflăm în cadrul romanului, utilizând operatorul %/% 100.

```
janeaustensentiment2 <- tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(book, index = linenumber %/% 100, sentiment)
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
janeaustensentiment2
```

```
# A tibble: 738 × 5
   book                index negative positive sentime
   <fct>               <dbl>    <dbl>    <dbl>    <dl
 1 Sense & Sensibility     0       20       47
 2 Sense & Sensibility     1       22       54
 3 Sense & Sensibility     2       16       35
 4 Sense & Sensibility     3       20       45
 5 Sense & Sensibility     4       21       63
 6 Sense & Sensibility     5       25       63
 7 Sense & Sensibility     6       39       44
 8 Sense & Sensibility     7       23       31
 9 Sense & Sensibility     8       15       39
10 Sense & Sensibility     9       22       63
# i 728 more rows
```

- **Index** - este referinta numerica pentru fiecare sectiune de text de 100 de randuri din fiecare carte
- **Negative** – numarul de cuvinte cu conotatie negative identificate in fiecare sectiune de text
- **Positive** – numarul de cuvinte cu conotatie pozitiva identificate in fiecare sectiune de text
- **Sentiment** – scorul total de sentiment pe fiecare sectiune calculat ca diferenta dintre numarul de cuvinte positive si numarul de cuvinte negative (de exemplu: pentru prima linie observam ca scorul sentimentelor este 12, ceea ce indica faptul ca predomina un sentiment pozitiva)
- Reprezentăm grafic scorurile sentimentelor din fiecare roman –
    - pe axa OX avem indexul care ține evidența numărului de secțiuni a câte 100 de rânduri.
    - Pe axa OY sunt reprezentate scorurile sentimentelor care pot fi pozitive sau negative, dupa caz



Observam ca în fiecare roman avem atat cuvinte cu o conotatie pozitiva, cat si cuvinte cu o conontatie negative, insa, pentru majoirtatea cartilor analizate, sentimentul dominant este positiv. Romanele unde apar cele mai multe sentimente negative sunt ***Mansfield Park***, ***Pride & Prejuduce*** si ***Northanger Abbey***. In timp ce, la polul opus se situeaza romanul ***Persuasion***.

```
ggplot(janeaustensentiment2, aes(index, sentiment, fill = book)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~book, ncol = 2, scales = "free_x")
```

Observam în fiecare roman analizat se întâlnesc atât sentimente positive, cât și negative.

- Comparăm cele trei dicționare de sentimente și examinăm modul în care sentimentul se schimbă în functie de firul narativ al romanului “Pride & Prejudice”.

```
afinn2 <- pride_prejudice %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(index = linenumber %/% 100) %>%
  summarise(sentiment = sum(value)) %>%
  mutate(method = "AFINN")

bing_and_nrc2 <- bind_rows(
  pride_prejudice %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  pride_prejudice %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive",
                                         "negative"))) %>%
    mutate(method = "NRC")) %>%
  count(method, index = linenumber %/% 100, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
bing_and_nrc2

bind_rows(afinn2,
          bing_and_nrc2) %>%
  ggplot(aes(index, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")
```

Am obținut sentimentul dominant (pozitiv sau negativ) pe fiecare secțiune din textul analizat, pentru cele 3 lexicoane. Unirea celor 3 lexicoane permite vizualiza sau reprezenta grafica a sentimentelor dominante. Fiecare dintre aceste lexicoane măsoară sentimentul secțiunilor textului într-un mod distinct, iar graficul arată cum variază sentimentele dominante în funcție de index. In toate cele 3 grafice, sentimentul pozitiv pare sa fie dominant, existand si sectiuni in care sentimentul negativ este dominant, insa foarte putine. **AFINN** pare să aibă cea mai mare variație de scoruri, urmat de **Bing et al.**, respectiv **NRC** care pare să fie cel mai echilibrat, cu mai puține extreme. Sentimentele dominante positive sunt mai mari ca zero si sunt dominante si cele negative sunt mai mici ca zero, insa nu domina:

- ***Lexiconul AFFIN*** valorile sentimentului variază între aproximativ -20 și +80 si prezintă cel mai mic număr de cuvinte negative.
- ***Lexiconul Bing*** valorile sentimentului variază între aproximativ -20 și +30 si indica o prevalenta a cuvintelor positive in text.
- ***Lexiconul NRC*** valorile sentimentului variază între aproximativ -10 și +30 si are cel mai mare număr de cuvinte positive si un numar extrem de mic de cuvinte negative.

- A**nalizam frecvența cuvintelor** care contribuie la fiecare sentiment prin utilizarea lexiconului Bing.

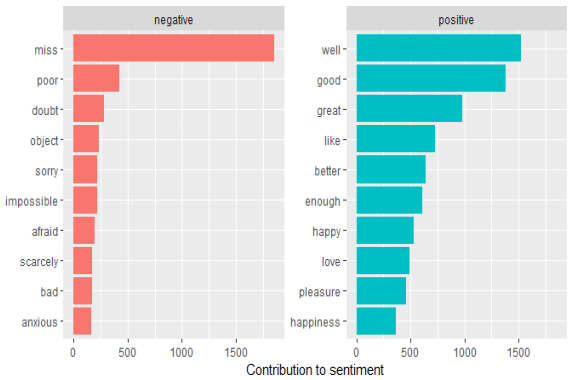
```
bing_word_counts <- tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
bing_word_counts
```

```
> bing_word_counts
# A tibble: 2,585 × 3
   word       sentiment     n
   <chr>      <chr>     <int>
 1 miss       negative   1855
 2 well       positive   1523
 3 good       positive   1380
 4 great      positive    981
 5 like       positive    725
 6 better     positive    639
 7 enough     positive    613
 8 happy      positive    534
 9 love       positive    495
10 pleasure   positive    462
# i 2,575 more rows
```

Cuvântul cu cea mai mare frecvență de apariție este ***miss*** (care transimite un sentiment negativ), urmat de ***well*** si ***good*** care transmit un sentiment pozitiv.

Reprezentăm grafic frecvențele calculate.

```
bing_word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```



Observam că cel mai frecvent cuvant negative utilizat este ***miss***, regăsit de peste 1855 de ori iar cel mai utilizat cuvânt pozitiv este ***well***, întâlnit de mai mult de 1523 de ori.

Norul de cuvinte (wordclouds): grupate după sentimente și sortate după sentimente (100 cuvinte)

- Folosind funcția wordcloud vom construi un nor de cuvinte format din cele mai frecvent utilizate 100 de cuvinte

```
library(wordcloud)
tidy_books %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```

Observam ca cel mai frecvent cuvânt utilizat este ***miss***, fiind urmat de cuvântul ***time***.

- reprezentam norul de cuvinte în funcție de sentimentul transmis (pozitiv sau negativ)

```
tidy_books %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("gray20", "gray80"),
                   max.words = 100)
```

Observam că cel mai întâlnit cuvânt negativ este ***miss*** iar cel mai frecvent utilizat cuvânt pozitiv este ***well***, respectiv ***good***.

- Creez un set de date in care specific pentru fiecare carte/roman numarul de capitole

```
austen_chapters <- austen_books() %>%
  group_by(book) %>%
  unnest_tokens(chapter, text, token = "regex",
                pattern = "Chapter|CHAPTER [\\dIVXLC]") %>%
  ungroup()


austen_chapters %>%
  group_by(book) %>%
  summarise(chapters = n()) %>%
  arrange(desc(chapters))
```

```
# A tibble: 6 x 2
  book                chapters
  <fct>                  <int>
1 Pride & Prejudice         62
2 Emma                      56
3 Sense & Sensibility       51
4 Mansfield Park            49
5 Northanger Abbey          32
6 Persuasion                25
```

Observam că cele mai multe capitole le are romanul ***Pride & Prejudice*** (62 de capitole), urmata de romanul ***Emma*** (56 de capitole) si ***Sense & Sensibility*** (51 de capitole).

- Vom crea o listă ce continue doar cuvintele negative din lexiconul Bing.

```
bingnegative <- get_sentiments("bing") %>%
  filter(sentiment == "negative")
bingnegative
```

```
# A tibble: 4,781 x 2
   word        sentiment
   <chr>       <chr>
 1 2-faces     negative
 2 abnormal    negative
 3 abolish     negative
 4 abominable  negative
 5 abominably  negative
 6 abominate   negative
 7 abomination negative
```

- adăugăm o coloană cu numărul de cuvinte din fiecare capitol pentru a obtine dimensiunea fiecarui capitol.

```
wordcounts <- tidy_books %>%
  group_by(book, chapter) %>%
  summarize(words = n())
wordcounts
```

```
# Groups:   book [6]
   book                chapter words
   <fct>                 <int> <int>
 1 Sense & Sensibility       0     7
 2 Sense & Sensibility       1  1571
 3 Sense & Sensibility       2  1970
 4 Sense & Sensibility       3  1538
 5 Sense & Sensibility       4  1952
 6 Sense & Sensibility       5  1030
 7 Sense & Sensibility       6  1353
 8 Sense & Sensibility       7  1288
 9 Sense & Sensibility       8  1256
10 Sense & Sensibility       9  1863
```

- Obținem frecventa relativa a cuvintelor negative din fiecare capitol față de numărul total de cuvinte din fiecare capitol.

```
tidy_books %>%
  semi_join(bingnegative) %>%
  group_by(book, chapter) %>%
  summarize(negativewords = n()) %>%
  left_join(wordcounts, by = c("book", "chapter")) %>%
  mutate(ratio = negativewords/words) %>%
  filter(chapter != 0) %>%
  top_n(1) %>%
  ungroup() %>%
  arrange(desc(ratio))
```

```
# A tibble: 6 x 5
  book                chapter negativewords words  ratio
  <fct>                 <int>         <int> <int>  <dbl>
1 Pride & Prejudice        34           111  2104 0.0528
2 Northanger Abbey         21           149  2982 0.0500
3 Sense & Sensibility      43           161  3405 0.0473
4 Mansfield Park           46           173  3685 0.0469
5 Emma                     15           151  3340 0.0452
6 Persuasion                4            62  1807 0.0343
```

Interpretare: Capitolul 34 al romanului ***Pride & Prejudice*** are cea mai mare proportie de cuvinte negative, cu un procent de 5,28%. In timp ce, la polul opus se afla capitolul 4 al romanului ***Persuasion*** are cea mai mică proporție de cuvinte negative, 3,43%.