Sentiment analysis Dostoyevsky's The Idiot and comparison with other works

In this report I use syuzhet R package to conduct analysis of one of my favourite books - Fyodor Dostoyevsky's The Idiot. We explore syuzhet package capabilites to see how sentiments change throughout the book. We compare different method of measuring sentiment and visualising it. Next we compare The Idiot with other works of Fyodor Dostoyevsky. Finally we do similar comparison with works of Leo Tolstoy, another great russian writer from the same epoch.

```
#devtools::install_github("mjockers/syuzhet")
Sys.setenv(JAVA_HOME="/Library/Java/JavaVirtualMachines/jdk1.8.0_20.jdk/Contents/Home")
library(syuzhet)
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(gutenbergr)
library(forcats)
library(ggthemes)
library(viridis)
```

```
## Loading required package: viridisLite
```

We will use Gutenberg Project Library to get the text. First let's examine which work of Dostoyevsky we have available. We use gutenbergr R package.

```
dostoyevsky <- gutenberg_authors %>% filter(author == "Dostoyevsky, Fyodor")
```

We focus on work which I remeber well - The Idiot.

```
(idiot_gutenberger_entry <- gutenberg_works(author == "Dostoyevsky, Fyodor", language == "en", title ==
```

```
## # A tibble: 1 x 8
##   gutenberg_id     title              author gutenberg_author_id language
##          <int>     <chr>               <chr>               <int>    <chr>
## 1         2638 The Idiot Dostoyevsky, Fyodor                 314       en
## # ... with 3 more variables: gutenberg_bookshelf <chr>, rights <chr>,
## #   has_text <lgl>
```

First we check the format, and see if text was downloaded correctly.

```
idiot <- gutenberg_download(idiot_gutenberger_entry["gutenberg_id"])
```

```
## Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest
```

```
## Using mirror http://aleph.gutenberg.org
```

```
head(idiot)
```

```
## # A tibble: 6 x 2
##   gutenberg_id                     text
##          <int>                    <chr>
## 1         2638                THE IDIOT
## 2         2638
## 3         2638    By Fyodor Dostoyevsky
## 4         2638
## 5         2638
## 6         2638 Translated by Eva Martin
```

The Idiot constitutes of 4 parts. As the work is long (Mordern Library edition has 667 pages), we will split into part for sake of visualisation. First we extract the sentences from the text using get_sentences method from syuzhet. Next we find where different parts of the book begin.

```
idiot_v <- get_sentences(idiot$text)
part_1_start = grep("PART I", idiot_v)[1]
part_2_start = grep("PART II", idiot_v)[1]
part_3_start = grep("PART III", idiot_v)[1]
part_4_start = grep("PART IV", idiot_v)[1]
```

We evaluate emotional valence on each sentence using 4 methods available in Syuzhet - defualt(syuzhet), bing, afinn and nrc. Additionally we also extract different sentiment from nrc lexicon.

```
linenumber = seq_along(idiot_v)
syuzhet_vector <- get_sentiment(idiot_v, method="syuzhet")
bing_vector <- get_sentiment(idiot_v, method="bing")
afinn_vector <- get_sentiment(idiot_v, method="afinn")
nrc_vector <- get_sentiment(idiot_v, method="nrc")
nrc_sentiment <- get_nrc_sentiment(idiot_v)

idiot_sentiment <- cbind(
                    tibble(linenumber = linenumber,
                           text = idiot_v,
                           syuzhet_emotional_valence = syuzhet_vector,
                           bing_emotional_valence = bing_vector,
                           afinn_emotional_valence = afinn_vector,
                           nrc_emotional_valence = nrc_vector),
                    nrc_sentiment)
head(idiot_sentiment)
```

```
##   linenumber
## 1          1
## 2          2
## 3          3
## 4          4
## 5          5
## 6          6
##
## 1 THE IDIOT\n\nBy Fyodor Dostoyevsky\n\n\nTranslated by Eva Martin\n\n\n\n\nPART I\n\nI.\n\nTowards
## 2                                                The morning was so damp and misty that it was only\nw
## 3       Some of the passengers by this particular train were returning from\nabroad; but the third-cl
## 4                                                                          All of them seemed wea
## 5
## 6
```

```
##   syuzhet_emotional_valence bing_emotional_valence afinn_emotional_valence
## 1                     -1.00                     -1                      -3
## 2                     -0.60                     -1                       3
## 3                     -0.25                      0                       1
## 4                     -0.75                     -1                      -2
## 5                      0.60                      0                       0
## 6                     -0.35                     -1                       0
##   nrc_emotional_valence anger anticipation disgust fear joy sadness
## 1                     0     0            1       1    0   0       0
## 2                    -2     1            0       0    1   0       2
## 3                    -1     1            0       0    0   0       1
## 4                    -1     0            0       0    0   0       1
## 5                     1     0            0       0    0   1       0
## 6                     0     0            3       0    1   2       0
##   surprise trust negative positive
## 1        0     0        1        1
## 2        0     0        2        0
## 3        0     0        1        0
## 4        0     0        1        0
## 5        0     1        0        1
## 6        2     1        2        2
```

We annotate different parts of the book and provide additional numering for each part to make visualisation of results easier.

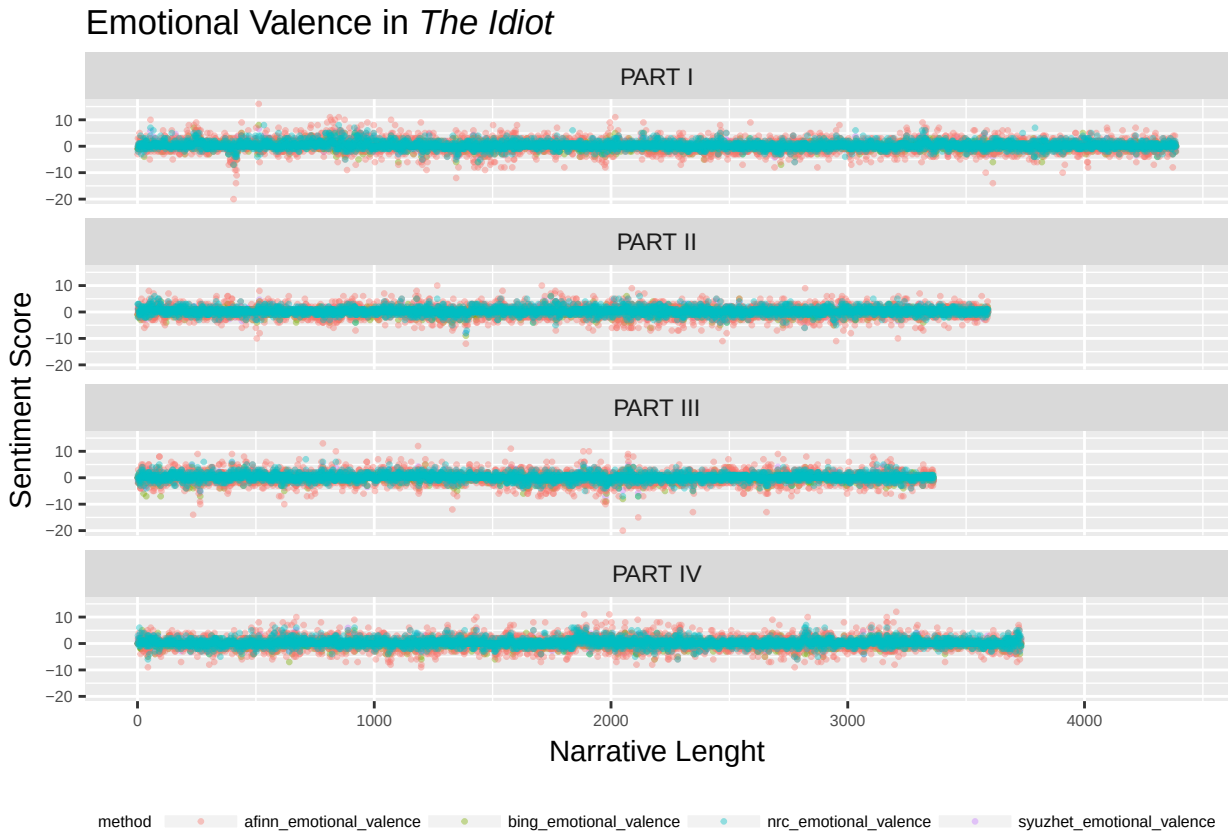
```
idiot_sentiment["part"] = "PART I"
idiot_sentiment[part_2_start:part_3_start, "part"] = "PART II"
idiot_sentiment[part_3_start:part_4_start, "part"] = "PART III"
idiot_sentiment[part_4_start:dim(idiot_sentiment)[1], "part"] = "PART IV"
idiot_sentiment$part = as.factor(idiot_sentiment$part)
idiot_sentiment[idiot_sentiment$part == "PART I", "part_linenumber"] = seq_along(1:(part_2_start - 1))
idiot_sentiment[idiot_sentiment$part == "PART II", "part_linenumber"] = seq_along(part_2_start:(part_3_s
idiot_sentiment[idiot_sentiment$part == "PART III", "part_linenumber"] = seq_along(part_3_start:(part_4_
idiot_sentiment[idiot_sentiment$part == "PART IV", "part_linenumber"] = seq_along(part_4_start:dim(idiot
#head(idiot_sentiment$part)
#tail(idiot_sentiment$part)
```

```
theme_syuzhet <- #theme_tufte() +
  theme(axis.ticks =element_line()) +
  theme(axis.text =element_text(size=6)) +
  theme(panel.border=element_blank()) +
  theme(legend.title=element_text(size=6)) +
  theme(legend.title.align=1) +
  theme(legend.text=element_text(size=6)) +
  theme(legend.position="bottom") +
  theme(legend.key.size=unit(0.2, "cm")) +
  theme(legend.key.width=unit(1, "cm"))

theme_syuzhet_no_x_axis <- theme_syuzhet +
  theme(axis.ticks=element_blank(), axis.text.x=element_blank())
```

Finally we can plot extracted sentiment. We use different colours for different sentiment extraction method.

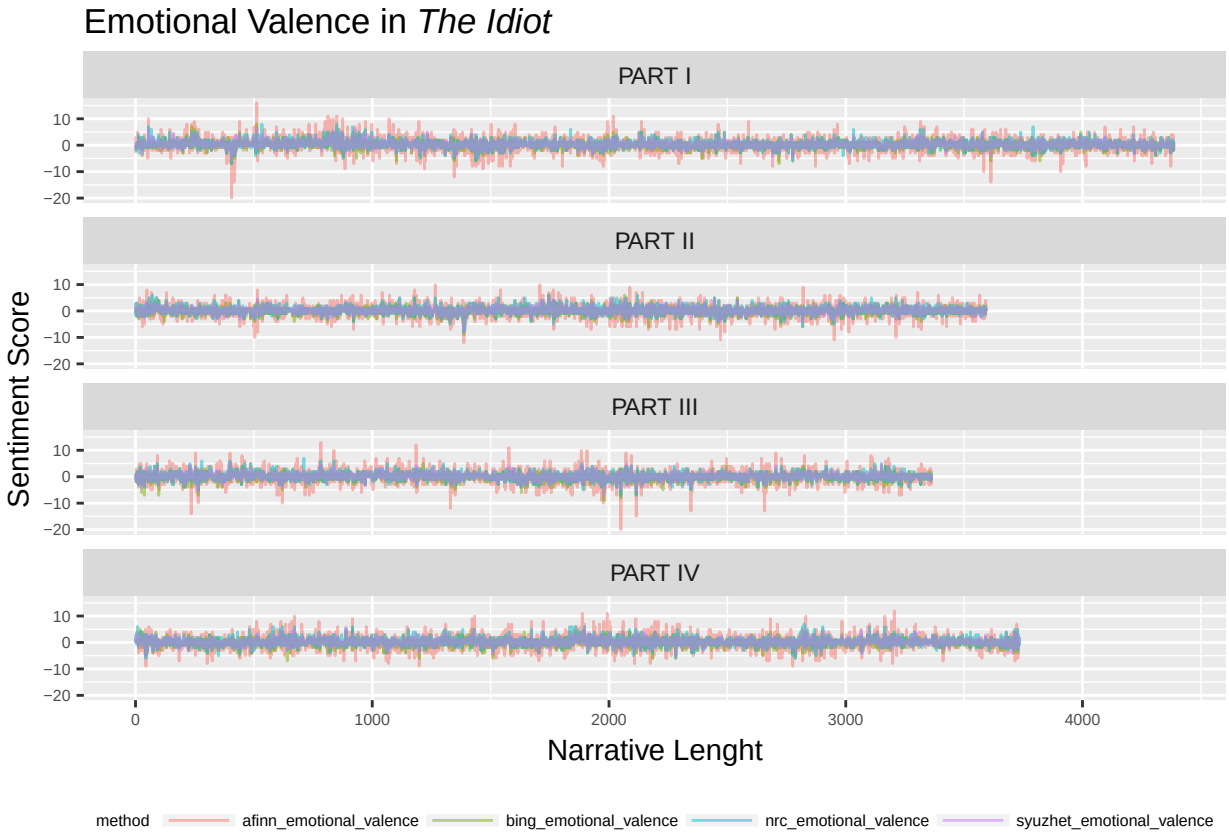
```
idiot_sentiment_by_method <- idiot_sentiment %>%
            select(linenumber, part_linenumber, part,
                   syuzhet_emotional_valence,
```

```
                  bing_emotional_valence,
                  afinn_emotional_valence,
                  nrc_emotional_valence) %>%
        gather(method, emotional_valence, syuzhet_emotional_valence : nrc_emotional_valence)
ggplot(idiot_sentiment_by_method, aes(x = part_linenumber, y = emotional_valence, color = method)) +
  geom_point(alpha = 0.4, size = 0.5) +
  facet_wrap(~part, nrow = 4) +
  theme_syuzhet +
  labs(y="Sentiment Score", x="Narrative Lenght", title = expression(paste("Emotional Valence in ", ita
```



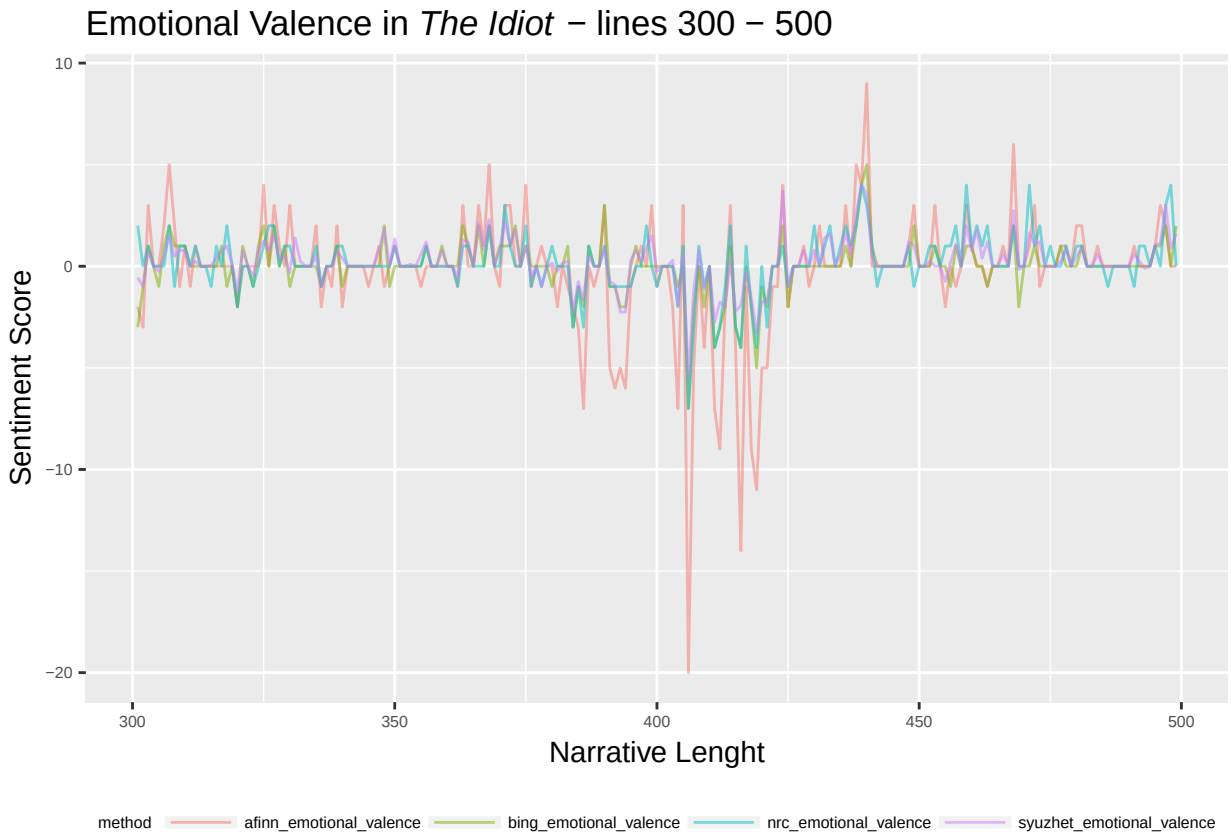
These plots need zooming to be reabable, but we can see that scales used by different methods are slightly different and some of them give only discrete values. Definately plotting these as point plots might not the best idea, let's try line plot instead.

```
ggplot(idiot_sentiment_by_method, aes(x = part_linenumber, y = emotional_valence, color = method)) +
  geom_line(alpha = 0.5) +
  facet_wrap(~part,  nrow = 4) +
  theme_syuzhet +
  labs(y="Sentiment Score", x="Narrative Lenght", title = expression(paste("Emotional Valence in ", ita
```



With lineplots we start two notice better general variability of emotional valance, and some extreme points. We will examine these extreme points now to get better intutions about numbers we are obtaining. We zoom in a bit at points with strongest negative emotions.

```
idiot_sentiment_by_method %>% filter(linenumber > 300, linenumber < 500) %>%
ggplot( aes(x = linenumber, y = emotional_valence, color = method)) +
  geom_line(alpha = 0.5) +
  theme_syuzhet +
  labs(y="Sentiment Score", x="Narrative Lenght", title = expression(paste("Emotional Valence in ", ita
```




```
idiot_sentiment %>% filter(linenumber >= 405, linenumber < 415)
```

```
##    linenumber
## 1         405
## 2         406
## 3         407
## 4         408
## 5         409
## 6         410
## 7         411
## 8         412
## 9         413
## 10        414
##
## 1
## 2
## 3  But _here_ I should imagine the\nmost terrible part of the whole punishment is, not the bodily pa
## 4
## 5
## 6
## 7
## 8
## 9
## 10
##    syuzhet_emotional_valence bing_emotional_valence
## 1                       0.50                      0
## 2                      -5.50                     -7
## 3                      -1.20                     -2
```

```
## 4                      0.80                    0
## 5                     -1.15                   -2
## 6                      0.00                    0
## 7                     -2.75                   -4
## 8                     -1.75                   -3
## 9                     -2.00                   -2
## 10                     0.35                    1
##    afinn_emotional_valence nrc_emotional_valence anger anticipation
## 1                        3                     1     0            0
## 2                      -20                    -7     2            1
## 3                       -5                    -3     2            0
## 4                        0                     1     0            0
## 5                       -4                    -1     1            0
## 6                        0                     0     0            1
## 7                       -7                    -4     2            0
## 8                       -9                    -3     4            2
## 9                       -2                    -1     1            2
## 10                       3                     2     0            0
##    disgust fear joy sadness surprise trust negative positive   part
## 1        0    0   1       0        1     0        0        1 PART I
## 2        2    5   0       6        0     1        7        0 PART I
## 3        2    3   0       2        0     0        4        1 PART I
## 4        0    0   0       0        0     0        0        1 PART I
## 5        1    1   0       1        0     1        2        1 PART I
## 6        0    0   0       0        0     0        0        0 PART I
## 7        2    3   0       2        2     0        4        0 PART I
## 8        4    4   0       3        1     1        4        1 PART I
## 9        1    2   0       2        1     0        2        1 PART I
## 10       0    0   0       0        0     0        0        2 PART I
##    part_linenumber
## 1              405
## 2              406
## 3              407
## 4              408
## 5              409
## 6              410
## 7              411
## 8              412
## 9              413
## 10             414
```

Seems like line 406 is the most negative one.

```
idiot_sentiment %>% filter(linenumber == 406) %>% select(text)
```

```
##
## 1 Now with the rack and tortures and so on--you suffer terrible pain of\ncourse; but then your tortu
```

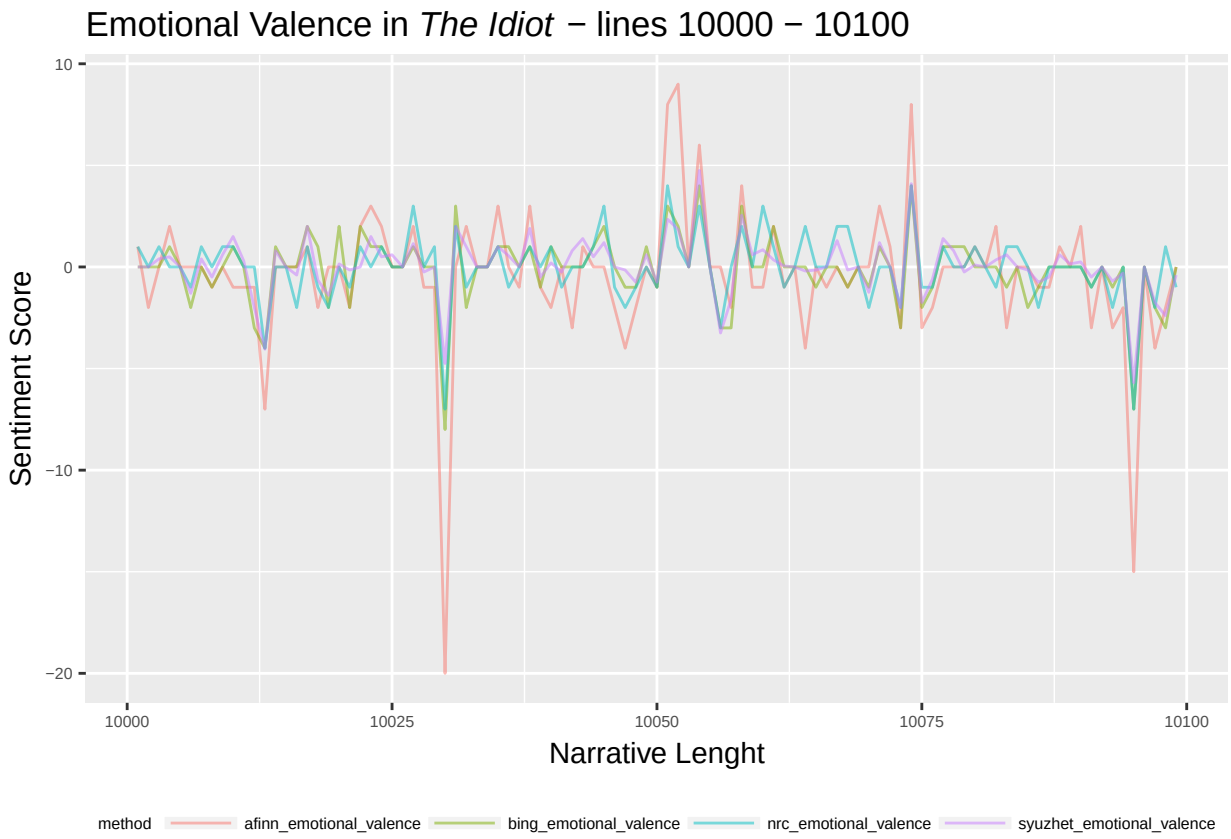
This surely sounds negative - lets see what other emotions it conveys.

```
emotions <- c("anger", "anticipation", "disgust", "fear",
                              "joy", "sadness", "surprise", "trust")
idiot_sentiment %>% filter(linenumber == 406) %>% select(one_of(emotions))
```

```
##   anger anticipation disgust fear joy sadness surprise trust
## 1     2            1       2    5   0       6        0     1
```

Emotions captured in the sentence obviously depend on larger context, but judging by sentence itself, authomatically recognized emotions are quite accurate. Let's look at another low point - around line number 10000.

```
idiot_sentiment_by_method %>% filter(linenumber > 10000, linenumber < 10100) %>%
ggplot( aes(x = linenumber, y = emotional_valence, color = method)) +
  geom_line(alpha = 0.5) +
  theme_syuzhet +
  labs(y="Sentiment Score", x="Narrative Lenght", title = expression(paste("Emotional Valence in ", ita
```




```
idiot_sentiment %>% filter(linenumber >= 10025, linenumber < 10035)
```

```
##    linenumber
## 1       10025
## 2       10026
## 3       10027
## 4       10028
## 5       10029
## 6       10030
## 7       10031
## 8       10032
## 9       10033
## 10      10034
##
## 1
## 2
## 3
## 4
```

```
## 5
## 6  What if I were now to\ncommit some terrible crime--murder ten fellow-creatures, for instance,\nor
## 7
## 8
## 9
## 10
##    syuzhet_emotional_valence bing_emotional_valence
## 1                       0.60                      0
## 2                       0.00                      0
## 3                       1.15                      1
## 4                      -0.25                      0
## 5                       0.00                      0
## 6                      -4.75                     -8
## 7                       2.00                      3
## 8                       1.00                     -2
## 9                       0.00                      0
## 10                      0.00                      0
##    afinn_emotional_valence nrc_emotional_valence anger anticipation
## 1                        0                     0     0            0
## 2                        0                     0     0            1
## 3                        2                     3     0            1
## 4                       -1                     0     0            0
## 5                       -1                     1     0            0
## 6                      -20                    -7     6            3
## 7                        0                     2     0            0
## 8                        2                    -1     0            0
## 9                        0                     0     0            0
## 10                       0                     0     0            0
##    disgust fear joy sadness surprise trust negative positive     part
## 1        0    0   0       0        0     0        1        1 PART III
## 2        0    0   0       0        0     1        0        0 PART III
## 3        0    1   0       0        0     1        0        3 PART III
## 4        0    0   0       0        0     0        0        0 PART III
## 5        0    0   1       0        0     0        0        1 PART III
## 6        5    6   0       6        1     1        8        1 PART III
## 7        0    2   1       2        0     4        1        3 PART III
## 8        0    0   0       0        0     0        1        0 PART III
## 9        0    0   0       0        0     0        0        0 PART III
## 10       0    0   0       0        0     0        0        0 PART III
##    part_linenumber
## 1             2045
## 2             2046
## 3             2047
## 4             2048
## 5             2049
## 6             2050
## 7             2051
## 8             2052
## 9             2053
## 10            2054
```

This time line 10030 is the most negative.

```
idiot_sentiment %>% filter(linenumber == 10030) %>% select(text)
```

```
##
## 1 What if I were now to\ncommit some terrible crime--murder ten fellow-creatures, for instance,\nor a
```

```
idiot_sentiment %>% filter(linenumber == 10030) %>% select(one_of(emotions))
```

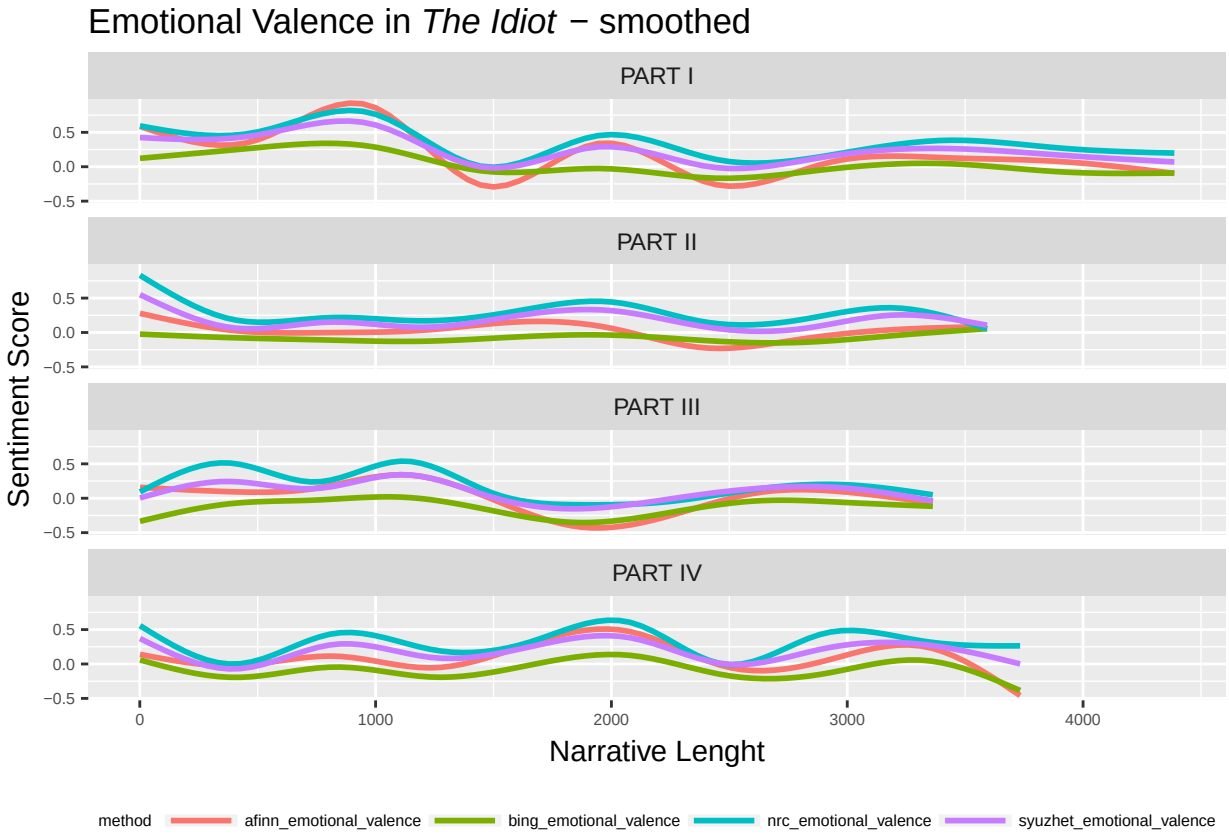
```
##   anger anticipation disgust fear joy sadness surprise trust
## 1     6            3       5    6   0       6        1     1
```

Again - it seems that our method might be off, but again reasonably so.

Having better understanding and some proof of accuracy of obtained scores let's come back to our comparison. Let's try plotting smoothed values for different scoring methods.

```
ggplot(idiot_sentiment_by_method, aes(x = part_linenumber, y = emotional_valence, color = method)) +
  geom_smooth(alpha = 0.5, se= FALSE) +
  facet_wrap(~part, nrow = 4) +
  theme_syuzhet +
  labs(y="Sentiment Score", x="Narrative Lenght", title = expression(paste("Emotional Valence in ", ital
```

```
## `geom_smooth()` using method = 'gam'
```

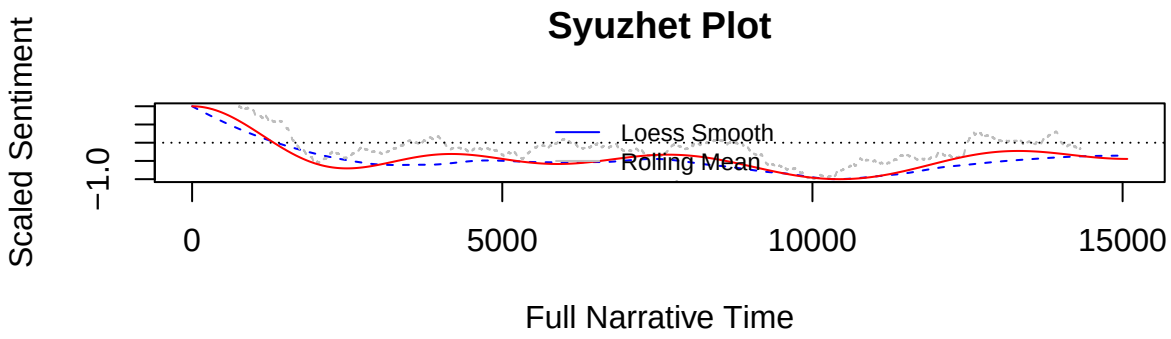


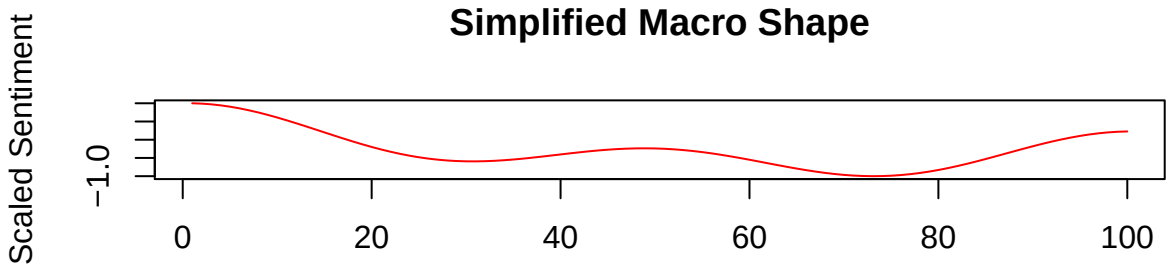
Plot presents smoothed sentiments comuted with different methods. Smoothing is done by fitting GAM(Generalized Additive Model) to the data, curve is approximated by spline. We see that all curves share the same basic shape. For the further analysis we will use Syuzhet default sentiment valence estimation method. Rest of the sentiments will be measured with NRC method.

First let's again look at the plot. This time instead of ggplot we use plotting capabilities of Syuzhet. Shape is different here, as we use single curve for the whole novel, and we use different smothing method. General shape of the curve seems to be acurrately representing the novel. At the beginning of the novel there are many

postive emotions which go steadily more negative until the middle of first part. From this point sentimnt is oscilating but stays on the negative side. Finally we reach the lowest point around line 10000 - 11000. This is the point round the end of part 3 and beggining of part 4. From this point emotion steadily rise util the ending which is again on the sad note. This reflects event in the novel well, althought some of the plots - e.g. simplified macro shape don't capture dip at the end of the novel.

```
simple_plot(idiot_sentiment$syuzhet_emotional_valence)
```





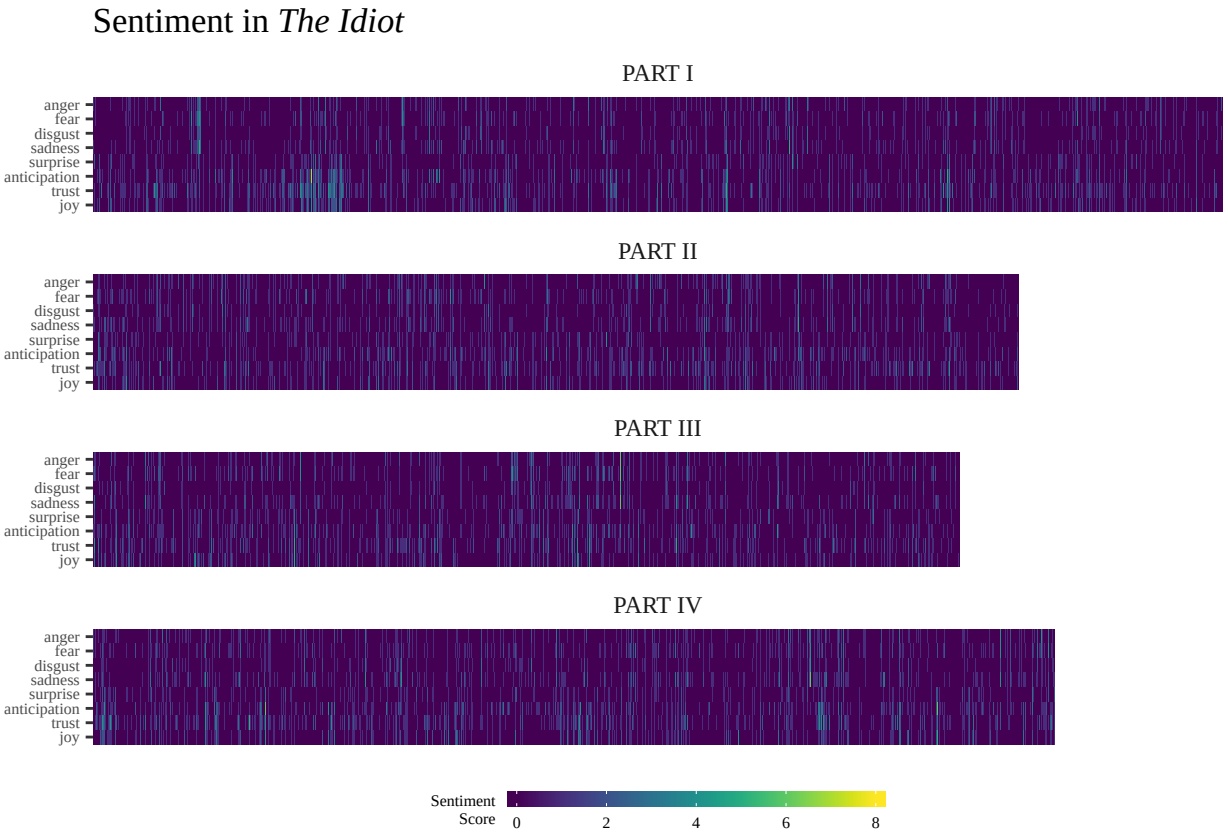
Now let's inspect how different emotions appear in the novel.

```
emotions <- idiot_sentiment %>% select(linenumber, part, part_linenumber, anger, anticipation,
                                       disgust, fear, joy, sadness, surprise,
                                       trust) %>%
        gather(sentiment, value, anger:trust)
emotions$sentiment <- as_factor(emotions$sentiment)
emotions$sentiment <- fct_relevel(emotions$sentiment, c("joy","trust","anticipation","surprise","sadness
head(emotions)
```

```
##   linenumber   part part_linenumber sentiment value
## 1          1 PART I               1     anger     0
## 2          2 PART I               2     anger     1
## 3          3 PART I               3     anger     1
## 4          4 PART I               4     anger     0
## 5          5 PART I               5     anger     0
## 6          6 PART I               6     anger     0
```

```
emotions %>%
  ggplot(aes(x = part_linenumber, y = sentiment, fill = value)) +
  geom_tile(width = 3) +
  facet_wrap(~part, nrow = 4) +
  scale_fill_viridis(name="Sentiment\nScore") +
  labs(x=NULL, y=NULL, title=expression(paste("Sentiment in ", italic("The Idiot")))) +
```

```
  theme_tufte() +
  theme_syuzhet +
  scale_x_discrete(expand=c(0,0))
```



On the plot we negative emotion in the higher parts of the stripe and positive ones in the lower part. In general we see that sentiment score stays low most of the time, but there are particular episodes(in particular in fist part) with stronger negative and positive emotions. Now let's try making use of more complex functionalities of Syuzhet.

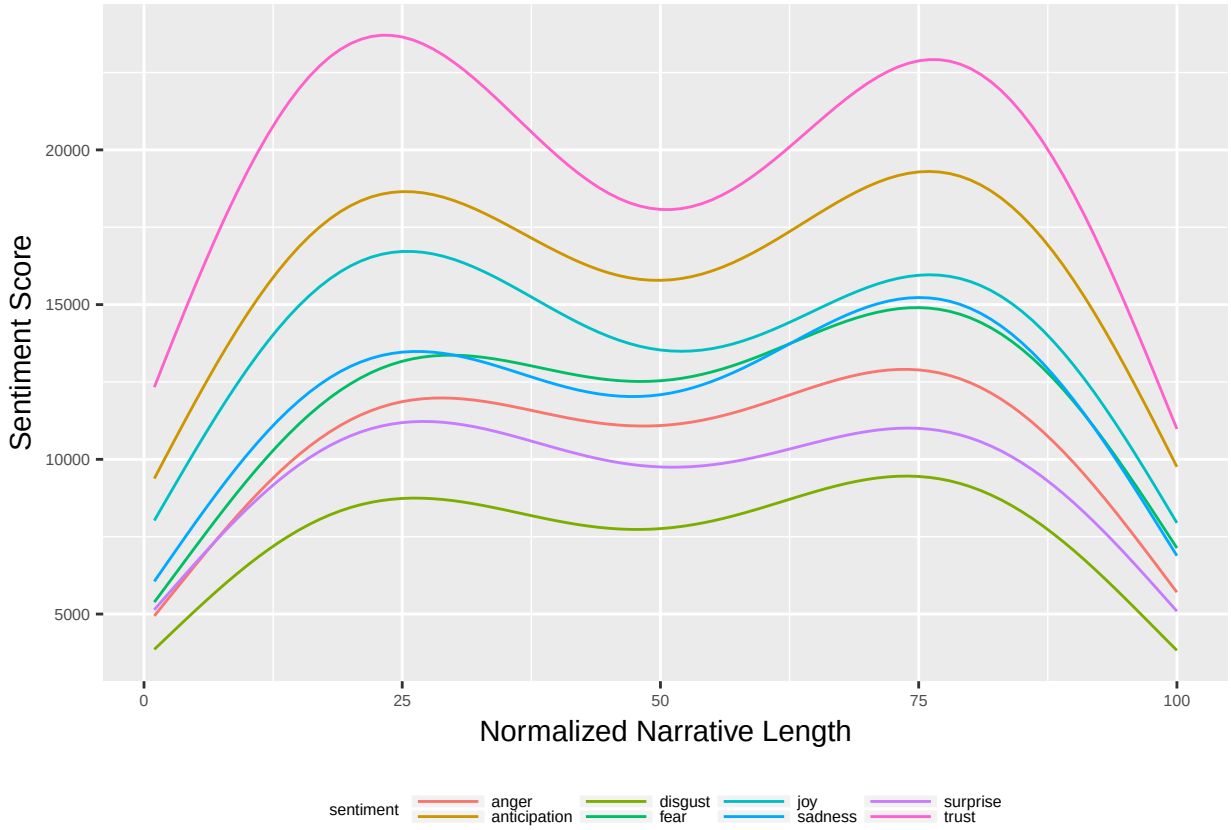
```
gutenberg_sentiments <- function(work) {
  sentence_v <- get_sentences(work$text)
  linenumber <- seq_along(sentence_v)
  emotional_valence <- get_sentiment(sentence_v, method="syuzhet")
  nrc_sentiment <- get_nrc_sentiment(sentence_v)
  cbind(
    tibble(linenumber = linenumber,
           text = sentence_v,
           emotional_valence = emotional_valence),
    nrc_sentiment)
}

sentiment_tranformed <- function(sentiment,
                                 columns,
                                 func = get_dct_transform) {
  transformed_list <- columns %>% map(~ func(sentiment[,.x]))
  names(transformed_list) <- columns
  transformed_list[["index"]] <- seq_along(transformed_list[[1]])
  as.tibble(transformed_list)
```

```
}

emotion_columns = c("joy","trust","anticipation","surprise","sadness","disgust","fear","anger")
valence_and_emotions = c("emotional_valence", emotion_columns)

emotional_summary <- function(sentiment) {
  colSums(prop.table(sentiment[,emotion_columns]))
}
```

```
idiot_sentiment["emotional_valence"] <- idiot_sentiment["syuzhet_emotional_valence"]
idiot_dct_tranformed <- sentiment_tranformed(idiot_sentiment, valence_and_emotions)
idiot_fourier_transformed <- sentiment_tranformed(idiot_sentiment, valence_and_emotions, get_transforme
```

Syuzhet have two methods of creating curves aproximating "emotional shape" of the novel - by Fourier transform with low pass filter and by Discrete Cosine Transport(dct). We inspect both methods.

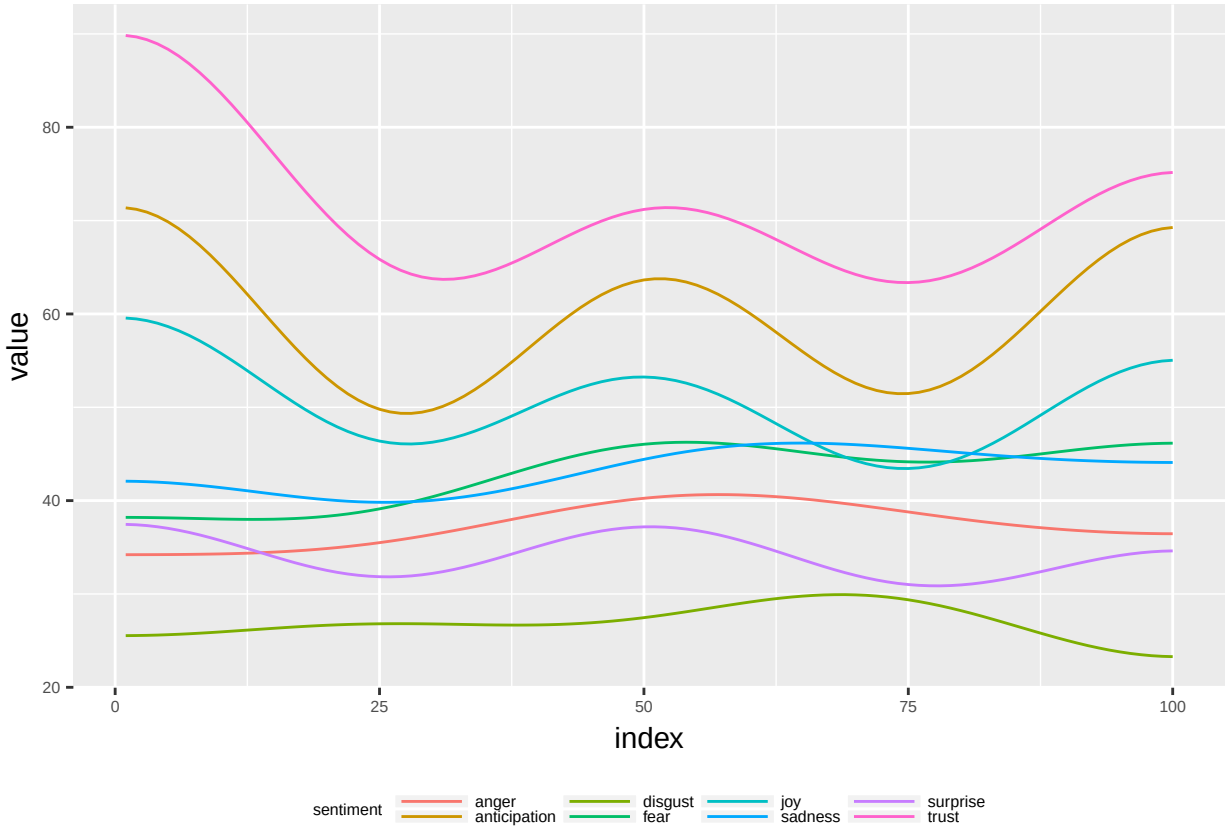
```
idiot_dct_tranformed["method"] <- "dct"
idiot_fourier_transformed["method"] <- "fourier"
idiot_fourier_transformed %>% gather(sentiment, value, joy:anger) %>%
ggplot(aes(x = index, y = value, color = sentiment)) +
  labs(y="Sentiment Score", x="Normalized Narrative Length") +
  geom_line() +
  theme_syuzhet
```



First let's see effect of using Fourier transform to get smoother version of sentiment curves for The Idiot. We see artifacts at the beginning and at the end of the novel. They are due to periodicity of fourier functions. This problem was noticed by authors and others https://annieswafford.wordpress.com/2015/03/30/why-syuzhet-doesnt-work-and-how-we-know/ nd other http://www.matthewjockers.net/ . This is why

currently author recommends using other method. Still plot gives us idea which emotiona are most strongly expressed through the narratve. We see that all emotions have similar base shape, but slightly different levels.

```
idiot_dct_tranformed %>% gather(sentiment, value, joy:anger) %>%
ggplot(aes(x = index, y = value, color = sentiment)) + geom_line() + theme_syuzhet
```



DCT tranfrom gives slightly different picture. Artifical dips at the beggining and at the end of the novel and not present anymore. We see that trust, anticipation and joy and surprise oscilating through the narrative time, starting high, and having two low points around 1/4-th and 3/4th of the narrive time, while negative emotions like anger, sadness and disgust start on the lower point and slowly raise, until the highest point which happens after the half of the novel(later for disgust). From this point they lower slighly until the end of the novel. Having well annotated novel could allow us to connect certain sentiments to specific moment in the narrative. At the current moment conducting such analysis is rather hard, nor long and complex novel like The Idiot and requires very good knowlegde of the text.

Although trasnformations are usefull in conducting analysis of general theme of the text, i believe one of it's main advantedeg is it's normalizing effet which allows ust to compare different texts. To explore this possibility I conduct simple analysis of emotional valence and expressed sentiment in the most famous works of Fyodor Dostoyevsky and Leo Tolstoy.

```
gutenberg_works(author == "Dostoyevsky, Fyodor", language == "en")
```

```
## # A tibble: 12 x 8
##    gutenberg_id
##           <int>
## 1           600
## 2          2197
## 3          2302
## 4          2554
```

```
## 5          2638
## 6          8117
## 7          8578
## 8         28054
## 9         36034
## 10        37536
## 11        38241
## 12        40745
## # ... with 7 more variables: title <chr>, author <chr>,
## #   gutenberg_author_id <int>, language <chr>, gutenberg_bookshelf <chr>,
## #   rights <chr>, has_text <lgl>
```

```
notes_from_underground <- gutenberg_download(600)
gambler <-gutenberg_download(2197)
crime_and_punishment <- gutenberg_download(2554)
brothers_karamazov <- gutenberg_download(28054)
white_nights <- gutenberg_download(36034)
```

```
gutenberg_works(author == "Tolstoy, Leo, graf", language == "en")
```

```
## # A tibble: 41 x 8
##    gutenberg_id                                title             author
##           <int>                                <chr>              <chr>
## 1           243  The Forged Coupon, and Other Stories Tolstoy, Leo, graf
## 2           689 The Kreutzer Sonata and Other Stories Tolstoy, Leo, graf
## 3           985                         Father Sergius Tolstoy, Leo, graf
## 4           986                         Master and Man Tolstoy, Leo, graf
## 5          1399                          Anna Karenina Tolstoy, Leo, graf
## 6          1938                           Resurrection Tolstoy, Leo, graf
## 7          2142                              Childhood Tolstoy, Leo, graf
## 8          2450                                Boyhood Tolstoy, Leo, graf
## 9          2600                          War and Peace Tolstoy, Leo, graf
## 10         2637                                  Youth Tolstoy, Leo, graf
## # ... with 31 more rows, and 5 more variables: gutenberg_author_id <int>,
## #   language <chr>, gutenberg_bookshelf <chr>, rights <chr>,
## #   has_text <lgl>
```

```
#father_sergius <- gutenberg_download(985)
war_and_peace <- gutenberg_download(2600)
master_and_man <- gutenberg_download(986)
anna_karenina <- gutenberg_download(1399)
resurrection <- gutenberg_download(1938)
```

I use Project Guteberg again for this aim. I take 6 works from Fyodor Dostoyevsky - Notes from the Underground, Gambler, White Nights, Crime and Punishment, Idiot and Brothers Karamazov. I wanted to incude both 3 'greatest' works of Dostoyevsky as well as some acclaimed shorter works. Similarly for Leo Toylsto I analyze his 2 most famoust works Anna Karenina and War and Peace, and two of the later work - Master and Man and Resurrection.

```
author_sentiments <- function(works) {
   works %>%
    map(gutenberg_sentiments)
}

author_emotional_summary <- function(works) {
  works %>%
```

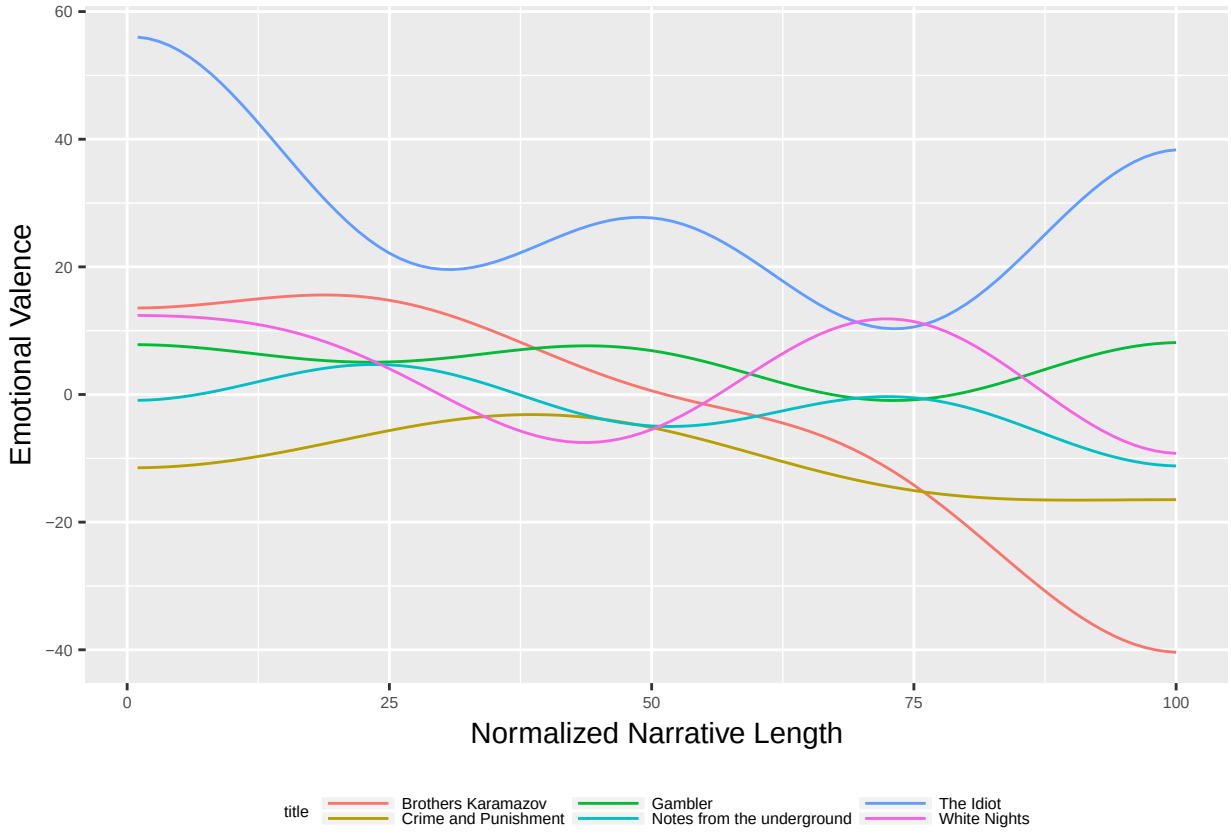
```
    map(~emotional_summary(gutenberg_sentiments(.x))) %>%
    map2(names(works), ~ c(.x, title = .y)) %>%
    reduce(rbind)
}

author_valence <- function(works) {
  works %>%
    map(~ sentiment_tranformed(gutenberg_sentiments(.x), c("emotional_valence"))) %>%
    map2_df(names(works), ~ mutate(.x, title = .y))
}
```

```
tolstoy_works <- list(war_and_peace, anna_karenina, master_and_man, resurrection)
names(tolstoy_works) <- c("War and Peace", "Anna Karenina","Master and Man", "Resurrection")
tolstoy_emotional_valence <- author_valence(tolstoy_works)
tolstoy_emotional_summary <- author_emotional_summary(tolstoy_works)
```

```
dostoyevsky_works <- list(white_nights, notes_from_underground, gambler, crime_and_punishment, idiot, bi
names(dostoyevsky_works) <- c("White Nights", "Notes from the underground", "Gambler", "Crime and Punish
dostoyevsky_emotional_valence <- author_valence(dostoyevsky_works)
dostoyevsky_emotional_summary <- author_emotional_summary(dostoyevsky_works)
```

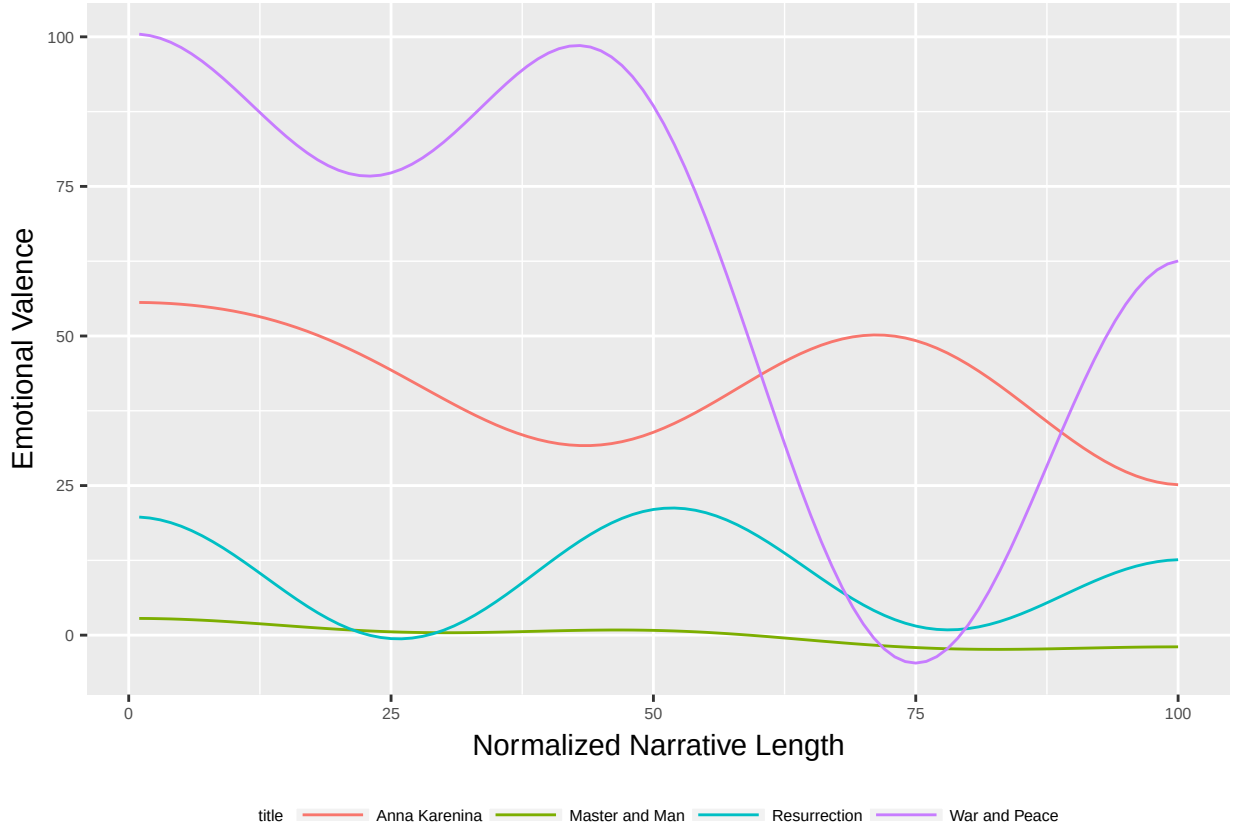
First we look at emotional valence in the works of Dostoyevsky.

```
dostoyevsky_emotional_valence %>%
ggplot(aes(x = index, y = emotional_valence, color = title)) +
  geom_line() +
  theme_syuzhet +
  labs(y="Emotional Valence", x="Normalized Narrative Length")
```



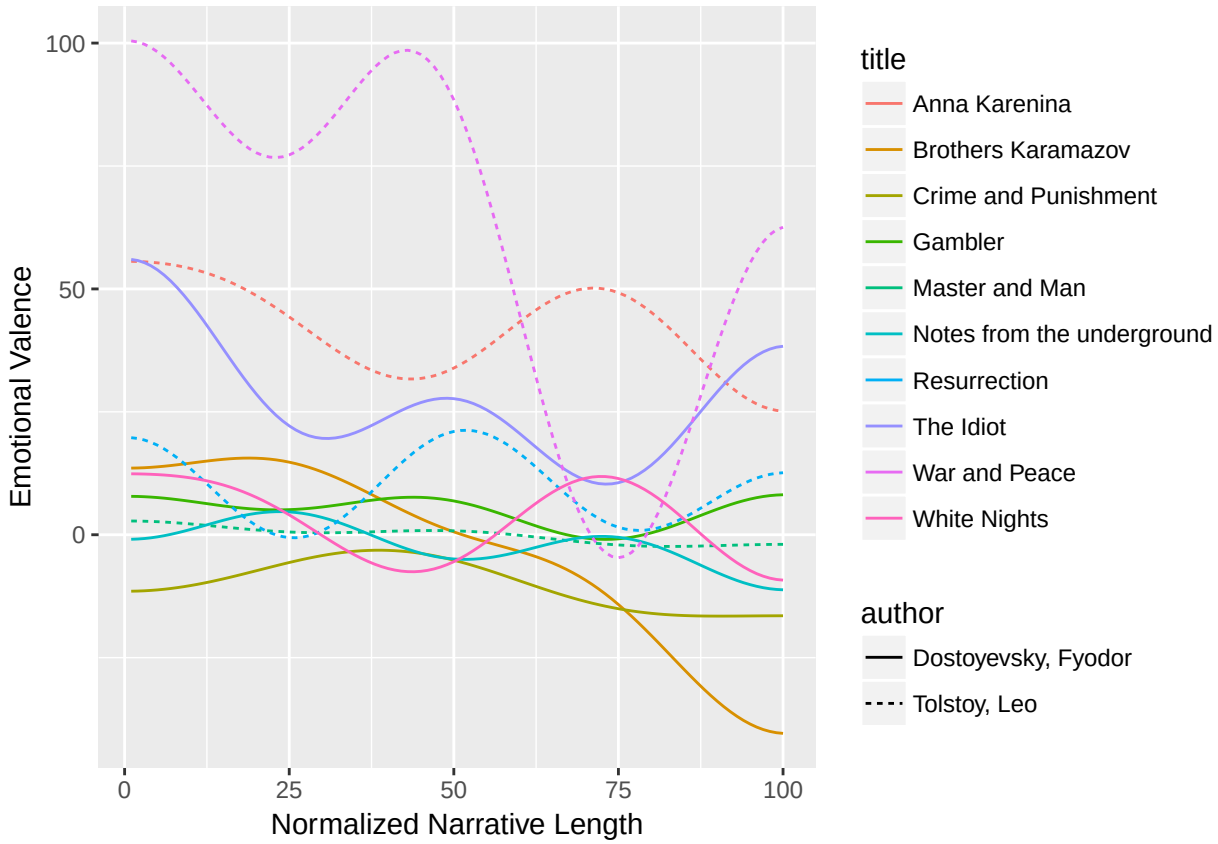
It seems that Idiot is quite an unsual works for Dostoyevsky, it is much more positive than his other works. We see that most of the works either oscilate between positive and negative emotiona , and largly negative. We see also that shapes differ quite singificantly between different text, nd there doesn't seem to be Dostoyesky formula for a novel. To get better perspective let's now take a look at works of the other great russian writer from the 19th century - Leo Tolstoy.

```
tolstoy_emotional_valence %>%
ggplot(aes(x = index, y = emotional_valence, color = title)) +
  geom_line() +
  theme_syuzhet +
  labs(y="Emotional Valence", x="Normalized Narrative Length")
```



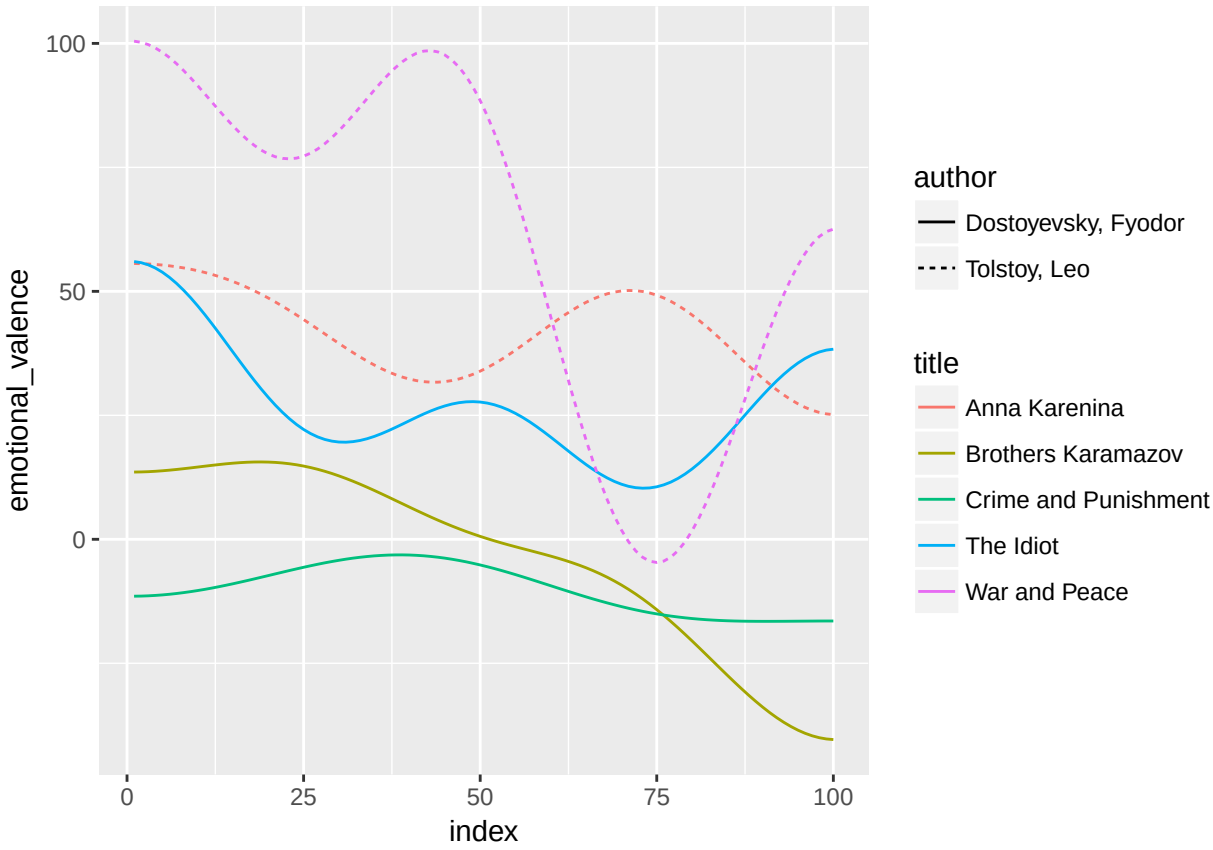
One common theme of Tolstoy's works is that the thye containe more positive emotions and his most famoust works "Anna Karenina" and War and Peace have their emotional valence higher than later work.

```
tolstoy_emotional_valence["author"] = "Tolstoy, Leo"
dostoyevsky_emotional_valence["author"] = "Dostoyevsky, Fyodor"
rbind(tolstoy_emotional_valence,dostoyevsky_emotional_valence) %>%
ggplot(aes(x = index, y = emotional_valence, color = title, linetype = author)) +
  geom_line() +
  #theme_syuzhet +
  labs(y="Emotional Valence", x="Normalized Narrative Length")
```



When we compare both writers we see that Tolstoy works are much stronger in postive emotions than works of Dostoyevsky. The Idiot is quite outstanding, compared to other Dostoyevsky works, and actually is closer to works of Toystoy thank other works of Dostoyevsky in terms of emotional valence. One interesting observation is that works which are considered the greatest for both authors have also biggest changes in emotional valence.

```
tolstoy_emotional_valence["author"] <- "Tolstoy, Leo"
dostoyevsky_emotional_valence["author"] <- "Dostoyevsky, Fyodor"
rbind(tolstoy_emotional_valence,dostoyevsky_emotional_valence) %>%
  filter(title %in% c("Anna Karenina", "War and Peace", "The Idiot", "Brothers Karamazov", "Crime and Pu
ggplot(aes(x = index, y = emotional_valence, color = title, linetype = author)) + geom_line()
```



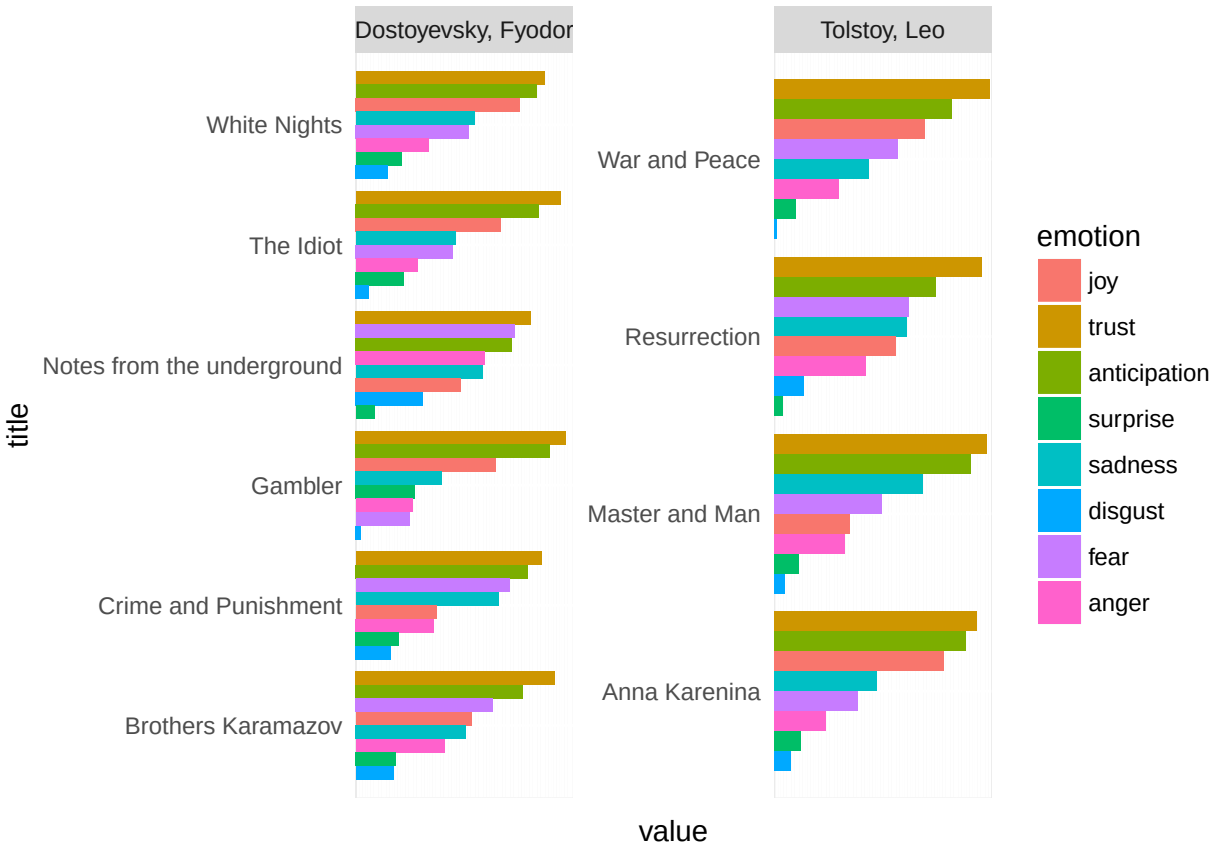
Let's look at perspecitive of distinct emotions in the works:

```
row.names(dostoyevsky_emotional_summary) <- 1:6
dostoyevsky_emotional_summary_df <- as.data.frame(dostoyevsky_emotional_summary)
dostoyevsky_emotional_summary_df["author"] <- "Dostoyevsky, Fyodor"

row.names(tolstoy_emotional_summary) <- 1:4
tolstoy_emotional_summary_df <- as.data.frame(tolstoy_emotional_summary)
tolstoy_emotional_summary_df["author"] <-  "Tolstoy, Leo"

emotional_summary_df <- rbind(dostoyevsky_emotional_summary_df, tolstoy_emotional_summary_df) %>% gathe
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
emotional_summary_df["emotion"] <- fct_relevel(as_factor(emotional_summary_df[,"emotion"]), c("joy","tr
```

```
ggplot(emotional_summary_df, aes(x = title, y = value, fill = emotion)) +
  geom_col(position = position_dodge()) +
  coord_flip() +
  scale_color_brewer(2) +

  theme(axis.ticks=element_blank(), axis.text.x=element_blank()) +
  facet_wrap(~author, scales = "free_y")
```



First we notice for both authors is that universally strongest emotion are trust and anticipation. With information we have it is hard to judge if this is some charateristic of work, or maybe it is caused by some flaw in analysis method. To evalute it properly we should conduct proper analysis of the bigger corpora of works from the period. Some impressions one could take is that negative emotions are more common in works of Dostoyevsky than in writing of Tolstoy. In Crime and Punishment and Brothers Karamazov, and in Notes fom the Underground fear is one of the dominating emotions. The Idiot seems to be most positive of Dostoyevsky works from the one we examined. When examining works of Tolstoy we see that some later work like Master and Man and Resurection contains more negative emotions, while in both Anna Karenina and War and Peace joy is one of the most common sentiments. Still we should treat this analysis as tool for hypothesis generation.

Syuzhet is fantastic tool. I makes sentiment analysis easy and pleasant. I found it particular useful for exploratory analysis and hypothesis generation of longer textual data. Still there is a number of hyperparametrs we have to tune to use, we have to choose the right lexicon and sentiment analysis method and it is not clear which methods of smoothing work best. Most of existing benchmarks for sentiment analysis are done for differnt kinds of texts - e.g. movie reviews and therefore are not always applicable for corpora of fiction, which often uses very specific, older language. To evaluate it properly it would be great to have bigger, sentiment annotated corpora of literature.