NLP - Assignment 4

In this assignment you will analyze the emotional arc of several books.

Preparations

1) Select five books that you find interesting and that do not drastically differ in length. The smallest book should be more than 50% of the size of the largest book.
2) Following the steps of the the previous assignments, read in the texts, extract the main text from them, and create a tibble with five rows that looks like this.

```
main_text_fun <- function(file){

    # load text
  text <- read_file(file)

  # define regex
  regex <- '\\*{3}[:print:]*\\*{3}'

  # cut text into sections
  text_split = str_split(text, '\\*{3}[:print:]*\\*{3}')

  # get sections
  sections <- text_split[[1]]

  # select main text
  main_text <- sections[2]

  }


# file
files <- list.files('books', full.names = T)

# process texts
texts <- sapply(files, main_text_fun)

# as tibble
text_tbl <- as_tibble(cbind(book = c('Alice in Wonderland','Dorian Gray', 'Huckleberry Finn', 'Peter Pa
), text = texts))

# print
text_tbl
```

```
## # A tibble: 5 x 2
##   book             text
##   <chr>            <chr>
## 1 Alice in Wonder~ "\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\nALICE'S AD~
## 2 Dorian Gray      "\r\n\r\n\r\n\r\n\r\nProduced by Judith Boss.  HTML ver~
## 3 Huckleberry Finn "\r\n\r\nProduced by David Widger\r\n\r\n\r\n\r\n\r\n\r~
## 4 Peter Pan        "\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\n\r\nPETER PAN\~
## 5 Treasure Island  "\r\n\r\n\r\n\r\n\r\nProduced by Judy Boss, John Hamm a~
```

2) Next use `unnest_tokens` to tokenize the text.

```
# tokenize
token_tbl <- text_tbl %>%
  unnest_tokens(word, "text")
```

3) Now use `tidyverse`'s `group_by()` and `mutate()` functions to add a variable `pos` that codes the position of a word inside the respective books.

```
# add pos variable
token_tbl <- token_tbl %>%
  group_by(XX) %>%
  mutate(pos = 1:n(),
         rel_pos = pos / max(pos)) %>%
  ungroup()
```

```
# add pos variable
token_tbl <- token_tbl %>%
  group_by(book) %>%
  mutate(pos = 1:n(),
         rel_pos = pos / max(pos)) %>%
  ungroup()
```

Sentiment analysis

1) Extract the *afinn* sentiment dictionary using the `get_sentiments` function and store it in an object called `afinn`.

2) Use `inner_join` to combine your `token_tbl` with `afinn`.

```
# add sentiments
token_tbl <- token_tbl %>%
  inner_join(get_sentiments("afinn"))
```

Smoothing

1. Use the `group_by` - `mutate` idiom along with the smooth function below to calculate more interpretable, smoothed sentiment scores for each of the books.

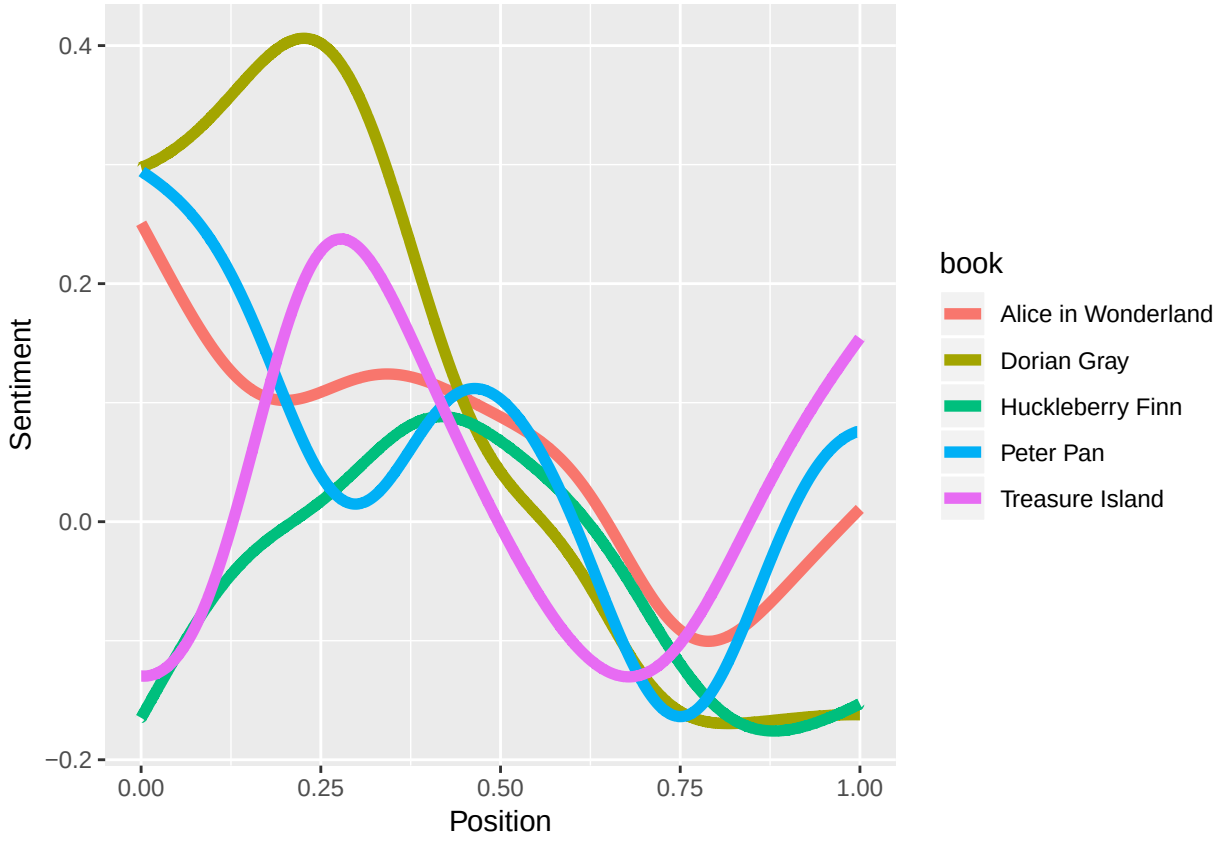
```
# smoothing function
smooth = function(pos, score){
  sm = sapply(pos, function(x) {
    weights = dnorm(pos, x, max(pos) / 10)
    sum(score * (weights / sum(weights)))
    })
  }
```

```
# smooth scores
token_tbl <- token_tbl %>%
  group_by(book) %>%
  mutate(smooth_score = smooth(pos, score))
```

2. Use the code below to create a plot like this:

```
ggplot(token_tbl,
       aes(rel_pos, smooth_score,color=book)) +
```

```
  geom_line(lwd=2) +
  labs(x = "Position", y = 'Sentiment')
```



Project proposal

Come up with 1 or 2 project proposals, each about half a page long. Address which question you would like to address and which data you want or would like to use for it.