Homework #5

2019150432 임효진

December 4, 2020

2

(a)

(1)

```
#"\" doesn't esape the backslash.
#"\\" means "\" but needs 2 \\ to escape.
#"\\\" needs another \ to escape.
```

(2)

```
#"\"'\\\\"
```

(3)

```
#It will match string such as .a.a.a
#It could be represented as a string by "\\..\\..\\..".
```

(b)

(1)

“[1]” “[^aeiou]” “[^e]ed""$i(ng|se)$”

(2)

One can use str_detect with regex “(^ie)|[^c]ie”.

(3)

One can use str_detect with regex “q[^u]” and see whether there are True values.

[1] aeiou

(4)

Considering the tendency of British English one can use regex “ou|ae|oe|ise|yse”

(5)

“^(010|02)-₩d₩d₩d₩d-₩d₩d₩d₩d”

(c)

(1)

“a?” is equivalent to “a{0,1}”. “a+” is equivalent to “a{1,}”. ”a*” is equivalent to “a{0,}”.

(2)

```
#^.*$ matches any string except a new line.
#"\\{.+\\}" matches at least one character inside curly brackets.
#\d{4}-\d{2}-\d{2} matches 4 digitits folowed by - followed by 2 digits followed
#by - followed by 2 digits.
#"\\\\{4}" matches 4 \.
```

(d)

(1)

```
str_subset(words, "^x|x$")
```

```
## [1] "box" "sex" "six" "tax"
```

```
str_subset(words, "^[aeiou].*[^aeiou]") %>%
    head()
```

```
## [1] "able"     "about"    "absolute" "accept"   "account"  "achieve"
```

```
words[str_detect(words, "a")&
      str_detect(words, "e")&
      str_detect(words, "i")&
      str_detect(words, "o")&
      str_detect(words, "u")]
```

```
## character(0)
```

(2)

```
max_num=max(str_count(words, "[aeiou]"))
words[(str_count(words, "[aeiou]"))==max_num]
```

```
## [1] "appropriate" "associate"   "available"   "colleague"   "encourage"
## [6] "experience"  "individual"  "television"
```

```
prop=str_count(words, "[aeiou]")/str_length(words)
words[prop==max(prop)]
```

```
## [1] "a"
```

3

(1)

```
library(gutenbergr)
gutenberg_metadata %>%
    filter(str_detect(title,
                      "^Pride and Prejudice$")) %>%
    select(gutenberg_id)
```

```
## # A tibble: 5 x 1
##   gutenberg_id
##          <int>
## 1         1342
## 2        20686
## 3        20687
## 4        26301
## 5        42671
```

(2)

```
gutenberg_works(languages="en") %>%
    filter(str_detect(title,
                      "^Pride and Prejudice$")) %>%
    select(gutenberg_id)
```

```
## # A tibble: 1 x 1
##   gutenberg_id
##          <int>
## 1         1342
```

(3)

```
book=gutenberg_download(1342)
```

(4)

```
library(tidytext)
words=book %>% unnest_tokens(word, text)
head(words)
```

```
## # A tibble: 6 x 2
##   gutenberg_id word
##          <int> <chr>
## 1         1342 there
## 2         1342 is
## 3         1342 an
## 4         1342 illustrated
## 5         1342 edition
## 6         1342 of
```

(5)

```
words=words %>%
    mutate(word_num=1:length(word))
head(words)
```

```
## # A tibble: 6 x 3
##   gutenberg_id word        word_num
##          <int> <chr>          <int>
## 1         1342 there              1
## 2         1342 is                 2
## 3         1342 an                 3
## 4         1342 illustrated        4
## 5         1342 edition            5
## 6         1342 of                 6
```

(6)

```
words=words %>% anti_join(stop_words) %>%
    filter(!str_detect(word, "^\\d+$"))
head(words)
```

```
## # A tibble: 6 x 3
##   gutenberg_id word        word_num
##          <int> <chr>          <int>
## 1         1342 illustrated        4
## 2         1342 edition            5
## 3         1342 title              8
## 4         1342 viewed            11
## 5         1342 ebook             13
## 6         1342 cover             15
```

(7)

```
sentiment=get_sentiments("afinn")
words=words %>% inner_join(sentiment, by="word")
head(words)
```

```
## # A tibble: 6 x 4
##   gutenberg_id word      word_num value
##          <int> <chr>        <int> <dbl>
## 1         1342 dear           218     2
## 2         1342 cried          279    -2
## 3         1342 dear           302     2
## 4         1342 delighted      344     3
## 5         1342 agreed         349     1
## 6         1342 dear           392     2
```

(8)

```
words %>% ggplot(aes(x=word_num, y=value))+
    geom_point(size=.3, alpha=.5)+
    geom_smooth()
```