

#выкачивание текстового файла в Rstudio

```
content<-readLines(file.choose(), n = -1L)
```

#удаление пустых строк

```
blank <- str_detect(content, "\"username\\": \"\"")
```

```
dat <- content[blank == FALSE]
```

#удаление записей, где не просматривалось видео

```
pdf <- str_detect(dat, "pdf")
```

```
dat <- dat[pdf == FALSE]
```

```
null <- str_detect(dat, "null")
```

```
dat <- dat>null == FALSE]
```

#поиск записей, содержащих события

```
event <- str_detect(dat, "\"event_type\\": \"[a-z]+_[a-z]+\"")
```

```
dat <- dat[event == TRUE]
```

#удаление ненужного текста (для удобства)

```
dat <- str_replace(dat, "\"agent\\": \"Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/59.0.3071.115 Safari/537.36\"","")
```

#создание вектора ID

```
ID <- str_extract(dat, "\"user_id\\": [[:digit:]]{1,}")
```

```
ID <- str_replace(ID, "\"user_id\\": ", "")
```

#вектор ID лекций

```
lecture <- str_extract(dat, "/courseware/[^ ]+/")
```

```
lecture <- str_replace_all(lecture, c("/courseware/" = "", "/" = "", " " = ""))
```

```
lecture <- str_sub(lecture, 34, nchar(lecture))
```

#создание вектора лекций

```
events <- str_extract(dat, "\"event_type\\": [^ ]+")
```

```
events <- str_replace_all(events, c("event_type" = "", "[^a-z]" = ""))
```

```
sum(is.na(events))
```

#вектор времени

```
time_spent <- str_extract(dat, "time\\": \"2017[^ ]+00:00")
```

```
time_spent <- str_replace(time_spent, "time\\": \"\" , \"\")
```

```
time_spent <- as.character(parse_datetime(time_spent))
```

```
library(readr)
```

#создание дата фрейма из четырех векторов

```
HSE <- data.frame(ID, lecture, events, time_spent, stringsAsFactors = FALSE)
```

#самые просматриваемые видео

```
HSE %>% select(events, lecture) %>%
```

```
  filter(events == "play_video") %>% group_by(events, lecture) %>% count(sort = TRUE)
```

#самые «загружаемые» видео

```
HSE %>% filter(events == "load_video") %>% select(ID, lecture) %>% group_by(ID) %>% count(sort = TRUE) -> loaders
```

#дата фрейм с лекциями и всеми событиями

```
HSE %>% select(lecture, events) %>% group_by(lecture, events) %>% count() -> with_video
```

```
with_video %>% filter(events == "seek_video") %>% arrange(desc(n))
```

```
spr <- spread(with_video, key = events, value = n, fill = 0)
```