Process Amherst History

Nicholas Horton (nhorton@amherst.edu)

September 27, 2020

Clean up Amherst History to make it easier for students to wrangle

```
process_chapter <- function(lines) {
  # we are looking to remove the page numbers, given by three blank lines, a line of text, then a blank
  # here's an example
  #
  #
  #
  # THE COLLEGE A SCHOOL OP AND FOE CHRIST. 443
  #
  # All those lines should disappear
  # I am assuming that the three blank lines are sufficient here to distinguish the pattern
  # Further checking may be warranted
  counter <- 0
  to_prune <- c()
  for (line in 1:(length(lines) - 2)) {
    if (lines[line] != "" & counter >= 2) {
      counter <- 0
    } else if (counter < 2) {
      counter <- counter + 1
    } else if (counter >= 2) { # remove some blank lines + page ref
      to_prune <- c(to_prune, (line - 2):(line + 1))
    }
  }
  return(lines[-to_prune])
}
```

```
history <- readLines("amherst_history.txt")
total_lines <- length(history)
total_lines
```

```
## [1] 33950
```

```
history[total_lines + 1] <- "CHAPTER END"

# first process chapter breaks
chapter_lines <- grep("^CHAPTER ", history)
length(chapter_lines)
```

```
## [1] 29
```

```
line_num <- 1
for (chapter in seq(chapter_lines)) {
```

```
  last_line <- chapter_lines[chapter] - 3    # last lines are blank
  cleaned <- process_chapter(history[line_num:last_line])
  sink(file = paste("chapter", sprintf("%02d", chapter - 1), ".txt", sep = ""))
  for (line in 1:length(cleaned)) { # last two lines are blank
    cat(cleaned[line], "\n")  # copy line
  }
  line_num <- chapter_lines[chapter]
  sink()
}
```

```
for (line_num in 1:length(history)) {
  if (history[line_num]) != "" { # non blank line
    counter <- 0
    cat(history[line_num], "\n")  # copy line
  } else {
    if (counter <= 2)
  }
}
```