Assignment 3

Steve Kerr (211924)

2023-10-02

String manipulation

The following homework sets you a challenging task to give structure to raw text. You will need to look through the text to discover text patterns that will allow you to split, join, and extract in order to create the structure we are looking for. If you struggle to implement this, describe in words how you would plan to go about the task, thinking about how to split this into subtasks and how to describe these.

To start with, you are asked to retrieve Songs of Innocence and of Experience by William Blake from Project Gutenberg. It is located at https://www.gutenberg.org/cache/epub/1934/pg1934.txt. This is a collection of poems in two books: *Songs of Innocence* and *Songs of Experience*.

The goal of the task is to parse this into a dataframe where each row is a line of a poem (there should be no empty lines). The following columns should describe where each line was found:

- line_number
- stanza_number
- poem_title
- book_title

Substeps

Think about how to split this up into smaller tasks before bringing this all together. Remember that when working on loops it is often easier to work with a single item in the list first, before putting this into a loop. There will be other ways to approach this, but a step-by-step approach will do the following:

- Get the content of the book (removing publisher information, contents, and the copyright notice)
- Split the book into the two sub-books (Songs of Innocence and Songs of Experience)
- Split each book into poems
- Split each poem into stanzas (verses)
- Split each stanza into lines

```
# load packages
pacman::p_load(
  dplyr,
  kableExtra,
  readr,
  stringr,
  tibble,
  tidyr
)
```

```r
# get text
url <- "https://www.gutenberg.org/cache/epub/1934/pg1934.txt"
text <- read_lines(url, skip = 104, n_max = 958) # lines 104-958 contain content we want

# wrangle data
df <- text |>
  as_tibble_col(column_name = "text") |> # convert chr vec to df col
  mutate(
    book_start = str_detect(text, "^SONGS OF"), # detect start of each book
    poem_start = str_detect(text, "^(?:[A-Z,'-]+\\s*)+$"), # detect start of each poem
    stanza_start = !book_start & !poem_start & text != "" & # detect start of each stanza
      lag(str_detect(text, "^\\s*$"), default = FALSE), # (detects lines that follow line breaks)
    book_title = ifelse(book_start, text, NA), # extract book title
    poem_title = ifelse(poem_start, text, NA)  # extract poem title
  ) |>
  fill(book_title, poem_title) |> # complete book title and poem title info
  filter(text != "") |> # remove empty lines
  group_by(book_title, poem_title) |>
  mutate(stanza_number = ifelse(stanza_start, cumsum(stanza_start), NA)) |> # assign stanza number
  fill(stanza_number) |> # complete stanza number info
  group_by(book_title, poem_title, stanza_number) |>
  mutate(line_number = cumsum(!book_start & !poem_start)) |> # assign line number
  ungroup() |>
  filter(!book_start, !poem_start) |> # remove lines containing book & poem titles
  select(-ends_with("_start")) |>  # remove cols used for indexing since no longer needed
  relocate(text, .after = last_col()) |> # move text to last col
  mutate(across(ends_with("_title"), str_to_title)) # adjust capitalization of titles

# print df
df |>
  head(15) |>
  kable(
    align = "l",
    booktabs = TRUE
  )
```

book_title	poem_title	stanza_number	line_number	text
Songs Of Innocence	Introduction	1	1	Piping down the valleys wild,
Songs Of Innocence	Introduction	1	2	Piping songs of pleasant glee,
Songs Of Innocence	Introduction	1	3	On a cloud I saw a child,
Songs Of Innocence	Introduction	1	4	And he laughing said to me:
Songs Of Innocence	Introduction	2	1	'Pipe a song about a Lamb!'
Songs Of Innocence	Introduction	2	2	So I piped with merry cheer.
Songs Of Innocence	Introduction	2	3	'Piper, pipe that song again.'
Songs Of Innocence	Introduction	2	4	So I piped: he wept to hear.
Songs Of Innocence	Introduction	3	1	'Drop thy pipe, thy happy pipe;
Songs Of Innocence	Introduction	3	2	Sing thy songs of happy cheer!'
Songs Of Innocence	Introduction	3	3	So I sung the same again,
Songs Of Innocence	Introduction	3	4	While he wept with joy to hear.
Songs Of Innocence	Introduction	4	1	'Piper, sit thee down and write
Songs Of Innocence	Introduction	4	2	In a book, that all may read.'
Songs Of Innocence	Introduction	4	3	So he vanished from my sight;