Ejercicios

Funes, el Memorioso (parte 1)

1. Separa el texto usando el carácter de espacio (" "). **Explora el parámetro simplify de la función str_split**
2. Cuenta el largo de cada una de los *strings* resultantes
3. Cuenta la cantidad de *strings* resultantes
4. Calcula el promedio del largo de los *string

```
library(readr)
funes <-  read_csv("data/funes_editado.csv")

words <- funes$text %>%
  str_split(pattern = " ", simplify = T)

lenght_words <- words %>%
  str_length()

summary(lenght_words)
```

Funes, el Memorioso (parte 2)

A partir de los *strings* separados:

- Extrae el primer caracter de cada *string*
- Extrae los 2 últimos caracteres

```
first_char <-  words %>%
  str_sub(start = 1, end = 1)

last_2 <- words %>%
  str_sub(start = -2, end = -1)
```

Validación de datos

```
datos <- tribble(~run, ~correo,
        "17.456.987-1", "roberto.bolaño@123cl",
        "15.246123-k",  "parranicanor@hola.cl",
        "14436.987-2",  "woolf_virginiagmail.cl",
        "18453986-9",   "nonafernandez@hotmail.com",
        "20.456.987-6",  "alejozambra@gmail.com"
        )
```

- Crear una nueva variable que contenga el dígito verificador del run (str_split)
- Eliminar todos los puntos de la columna *run* (str_remove)
- Validar que el correo tenga la siguiente estructura: X@X.X

```r
# Opción 1 para separar
datos$dv <- map_chr(str_split(datos$run, "-"), 2)
datos$run1 <- map_chr(str_split(datos$run, "-"), 1)
```

```r
# Opción 2 para separar
datos <- datos %>%
  separate(col = run, into = c("run1", "dv"), sep = "-", remove = F)
```

```r
# Removemos caracteres molestos en el run
datos <- datos %>%
  mutate(
    run_editado = str_remove_all(run1, "\\.|-"))
```

```r
# Validar el correo
datos <- datos %>%
  mutate(comprobar_correo = str_detect(correo, ".+@.+\\.+."))
```

Qué vergüenza

Comparemos los 2 cuentos de Paulina Flores

Para cada uno de los 2 textos, obtén los siguientes datos:

- Cantidad de palabras, excluyendo los signos de puntuación
- Cantidad de palabras únicas, excluyendo los signos de puntuación
- Cantidad de adjetivos
- Cantidad de sustantivos
- Cantidad de signos de puntuación

```r
# Cargar libro completo
libro <-  pdf_text("data/Qué vergüenza - Paulina Flores.pdf")

# Seleccionar algunos capítulos
que_verguenza <-  libro[4:13] %>%
  str_flatten()

talcahuano <- libro[26:43] %>%
  str_flatten()

cuentos <- c(que_verguenza, talcahuano)

# Procesar con udpipe
cuentos_procesado <- udpipe_annotate(modelo, cuentos)
cuentos_procesado_df <- as.data.frame(cuentos_procesado)  %>%
  select(doc_id:xpos)
```

```r
# Número de palabras
cuentos_procesado_df %>%
  filter(upos != "PUNCT") %>%
  group_by(doc_id) %>%
  summarise(n_palabras = n())

# Número de palabras distintas
cuentos_procesado_df %>%
  filter(upos != "PUNCT") %>%
  group_by(doc_id, token) %>%
```

```
  slice(1) %>%
  group_by(doc_id) %>%
  summarise(n_palabras = n())

# Número de signos de puntación, adjetivos y sustantivos
cuentos_procesado_df %>%
  filter(upos %in% c("PUNCT", "ADJ", "NOUN")) %>%
  group_by(doc_id, upos) %>%
  summarise(n_palabras = n())
```