Text Mining with sparklyr

Gaurav Tople, Sai Vineeth

4/19/2020

```
library(sparklyr)
library(dplyr)
```

Task 1:

a. Establish spark connection in RStudio (libraries: sparklyr, dplyr)

```
spark_install(version = "2.1.0")

sc <- spark_connect(master = "local", version = "2.1.0")
```

b. Load the text file "My_old_man.txt" into spark

```
file_path <- paste0(getwd(), "/My_old_man.txt")

myoldman <-  spark_read_text(sc, "myoldman", file_path)
```

c. Remove empty lines

```
myoldman<- myoldman %>%
  filter(nchar(line) > 0)
```

d. Remove punctuation

```
myoldman <- myoldman %>%
  mutate(line = regexp_replace(line, "[_\"\'():;,.!?\\-]", " "))
```

e. Separate each word using Spark API ft_tokenizer

```
word_list <- myoldman %>%
  ft_tokenizer(input_col = "line",
               output_col = "word_list")
```

f. Remove stop words (e.g., I, me, my, .)

```
wo_stop <- word_list %>%
  ft_stop_words_remover(input_col = "word_list",
                        output_col = "wo_stop_words")
```

g. Unnesting the tokens into their own row using explode; filtering the result with ncahr(word) > 1

```
exploded <- wo_stop %>%
  mutate(word = explode(wo_stop_words))

all_words <- exploded %>%
  filter(nchar(word) > 1)
```

h. Cache the result into Spark memory using compute()

```
all_words <- all_words %>%
  compute("all_words")
```

Task 2:

a. Generate a list of (word, count) in descending order of count

```
word_count <- all_words %>%
  group_by(word) %>%
  tally() %>%
  arrange(desc(n))
```

b. Create a list of the first 20 words with counts

```
first_20_word_count <- head(word_count,20)
print(first_20_word_count)
```

```
## # Source:     spark<?> [?? x 2]
## # Ordered by: desc(n)
##    word       n
##    <chr>  <dbl>
##  1 old       74
##  2 man       69
##  3 going     34
##  4 around    33
##  5 like      27
##  6 get       25
##  7 back      25
##  8 big       23
##  9 one       22
## 10 went      21
## # ... with more rows
```

c. How many distinct words are there in the list?

```
distinct_word_count <- all_words %>%
     select(word) %>%
     distinct() %>%
     count()

print(distinct_word_count)
```

```
## # Source: spark<?> [?? x 1]
##       n
##   <dbl>
## 1   933
```

Task 3: a. The code (your code should be tested in RStudio before submission)

b. The results: The list of the first 20 words with counts and the total number of the distinct words in the list. Ans: List of the first 20 words with count

```
print(as.data.frame(first_20_word_count))
```

```
##       word  n
## 1      old 74
## 2      man 69
## 3    going 34
## 4   around 33
## 5     like 27
## 6      get 25
## 7     back 25
## 8      big 23
## 9      one 22
## 10    went 21
## 11    came 21
## 12 looking 20
## 13   horse 19
## 14     got 19
## 15     way 19
## 16    kzar 19
## 17    said 18
## 18      go 17
## 19   right 16
## 20  george 16
```

Total number of distinct words in the list:

```
print(distinct_word_count)
```

```
## # Source: spark<?> [?? x 1]
##       n
##   <dbl>
## 1   933
```