A Guide for Tidying Text Data in R

Hello! In order to analyze subject line keywords, I have put together a simple guide on how to separate and organize the keywords in R.

1. First, install packages 'tidytext' and 'dplyr' in R and open the libraries in R:

```
> library(tidytext)
> library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

2. Read in data:

```
> library(readxl)
> text_analysis <- read_excel("C:/Users/Heidi/Desktop/SMMC Data/text analysis
.xlsx")
> View(text_analysis)
```

3. Use tibble() function to only include the factors you want to look at (e.g. open% and subject line):

```
> employeetext<-tibble(open=text_analysis$`open %`,text=text_analysis$`subjec
t line`)
```

4. Separate the subject lines into separate words by using the unnest_tokens() function:

```
> unnested<-employeetext %>%
+ unnest_tokens(word,text)
```

5. Use the group_by() function to group the words that are the same and use the summarise(mean()) function to average out the open rates for the words:

```
> View(unnested)
> averaged<-unnested %>%
+ group_by(word) %>%
+ summarise(averageopen=mean(open))
> View(averaged)
```

6. Order by average open rate (highest to lowest):

```
text_order<-averaged[order(-averaged$averageopen),]
```

7. (optional) plot:

```
library(ggplot2) # Basic barplot
```

```
ggplot(text_order[1:10,],aes(x=word,y=averageopen)) +
+ geom_bar(stat="identity", color="deeppink1",fill="red")
```

Here is a website that goes more into depth with tidying text data:

https://www.tidytextmining.com/