# Text Mining the Hitchhiker Trilogy

*Greg Johnson*

## Contents



## Reading and Tidying Our Data

Let's read in Doug's first book in his "trilogy," The Hitchhiker's Guide to the Galaxy. We will read in the text file as a character vector in which each element is a line from the book.

```
# setwd('~/Documents/Analytics/R/DouglasAdams')
book1 <- readLines("data/HitchHiker.txt")
book1 %<>% tibble(line = 1:length(.), text = .)
c(" ", head(book1$text, 20), " ") %>% kable
```

---

Douglas Adams

The Hitch Hiker's Guide to the Galaxy

Book 1

for Jonny Brock and Clare Gorst
and all other Arlingtonians
for tea, sympathy, and a sofa

Far out in the uncharted backwaters of the unfashionable end of the
western spiral arm of the Galaxy lies a small unregarded yellow sun.
Orbiting this at a distance of roughly ninety-two million miles is an
utterly insignificant little blue green planet whose apedescended life forms are so amazingly primitive that they still think digital watches are

---

Our data are in! But they're not tidy in the sense that we want one token per line. For a first look at our data, we will look at words. The *unnest_tokens* function will tidy our data (by default) into word tokens. It will also strip punctuation and convert our words to lowercase.

```
book1 %<>% mutate(linenumber = row_number(), chapter = cumsum(str_detect(text,
    regex("Chapter \\d+$", ignore_case = TRUE))))

book1words <- book1 %>% unnest_tokens(word, text)
```

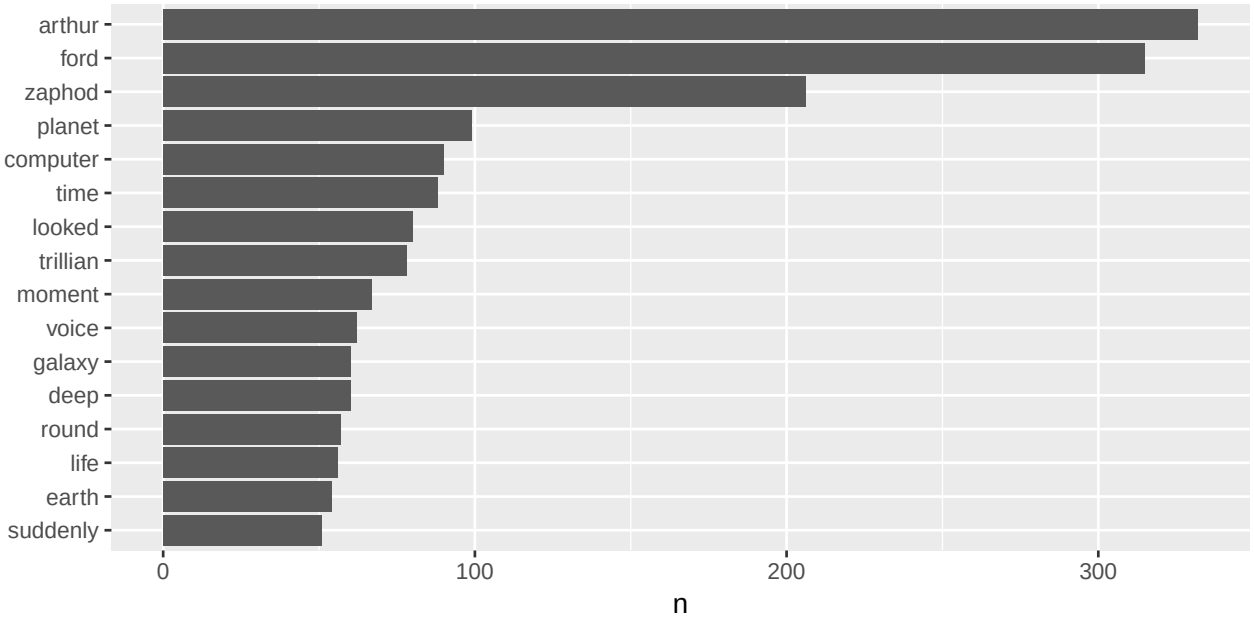
We will remove stop words from our new tidy dataset with the *anti_join* function.

```
book1words %<>% anti_join(stop_words)
```

## Word Frequency

Now we can start processing our tidy data - let's start with a simple word frequency count using *dplyr* and a visualization of frequencies using a word cloud.

```
book1count <- book1words %>% count(word, sort = TRUE)
book1count %>% filter(n > 50) %>% mutate(word = reorder(word, n)) %>% ggplot(aes(word,
    n)) + geom_col() + xlab(NULL) + coord_flip()
```




```
wordcloud(words = book1count$word, freq = book1count$n, max.words = 200, random.order = FALSE,
    rot.per = 0.35, colors = brewer.pal(8, "Dark2"))
```