Frankenstein Wordcloud

William Taylor Bickelmann

November 8, 2017

Abstract

This PDF will contain a wordcloud and title of the book 'Alice in Wonderland' by Lewis Carroll.

Alice in Wonderland

1 packages

This section will contain the packages which will then be used to load 'Alice in Wonderland', manipulate string and form wordclouds.

```
package<-c('dplyr')
library(tidytext)
library(tm)

## Loading required package:  NLP

library(wordcloud)

## Loading required package:  RColorBrewer

library(stringr)
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

library(knitr)
library(gutenbergr)
```

The first step is to determine the id of Alice in Wonderland:

```
gutenberg_works()%>%
  select(gutenberg_id,title,author)%>%
  filter(author=='Carroll, Lewis')

## # A tibble: 18 x 3
##    gutenberg_id
##           <int>
##  1           11
##  2           12
##  3           13
##  4          620
##  5          651
##  6         4763
##  7        19002
##  8        28696
##  9        28885
## 10        29042
## 11        29888
## 12        33582
## 13        35497
## 14        35535
## 15        36308
## 16        38065
## 17        48630
## 18        48795
## # ... with 2 more variables: title <chr>, author <chr>
```

In the resulting tibble from the code above, one can pick out the id of Alice; 11.

2 Chapter 1

Here I want to isolate the 'chapter 1' block of text

```
library(stringr)
df <- gutenberg_download(11)

## Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest
## Using mirror http://aleph.gutenberg.org

head(df[str_detect(df$text, '^CHAPTER'),],n=1)$text

## [1] "CHAPTER I. Down the Rabbit-Hole"
```

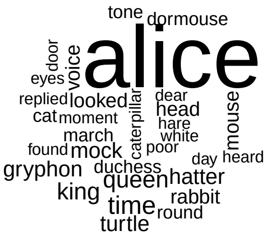
3 The Wordcloud

Next the wordcloud package will be used to form a wordcloud

```
words_df<-df%>%
  unnest_tokens(word,text)

words_df

## # A tibble: 26,694 x 2
##    gutenberg_id       word
##           <int>      <chr>
##  1           11    alice's
##  2           11 adventures
##  3           11         in
##  4           11 wonderland
##  5           11      lewis
##  6           11    carroll
##  7           11        the
##  8           11 millennium
##  9           11    fulcrum
## 10           11    edition
## # ... with 26,684 more rows
```

Using dplyr, we can remove stop words and insignificant

```
words_df<-words_df%>%
  filter(!(word %in% stop_words$word))
words_free <- words_df%>%
  group_by(word)%>%
  summarise(count = n())%>%
  arrange(-count)

wordcloud(words_free$word, words_free$count, min.freq = 25)
```


tone
dormouse
door
voice
alice
eyes
replied
looked
caterpillar
dear
head
mouse
cat
moment
hare
march
white
found
mock
poor
day
heard
gryphon
duchess
queen
hatter
king
rabbit
time
round
turtle