Topic_modelling

Group 5

2022-11-15

Set R enviornment

```
library(magrittr)
library(gutenbergr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

```
library(tidytext)
library(dplyr)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(RColorBrewer)
library(topicmodels)
library(ggplot2)
library(dplyr)
```

Importing Data

```
gutenberg_metadata %>%
  filter(title == "On the Origin of Clockwork, Perpetual Motion Devices, and the Compass")
```

```
## # A tibble: 1 x 8
##   gutenberg_id title               author guten~1 langu~2 guten~3 rights has_t~4
##          <int> <chr>               <chr>    <int> <chr>   <chr>   <chr>  <lgl>
```

```
## 1        30001 On the Origin of C~ Price~   34299 en      Techno~ Publi~ TRUE
## # ... with abbreviated variable names 1: gutenberg_author_id, 2: language,
## #   3: gutenberg_bookshelf, 4: has_text
```

```
book <- gutenberg_download(30001)
```

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

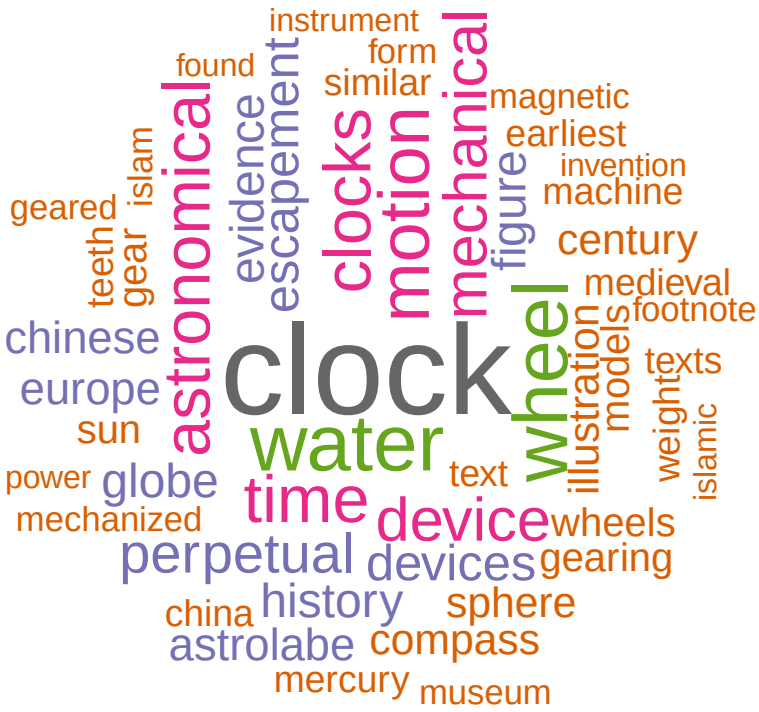
```
## Using mirror http://aleph.gutenberg.org
```

Counting the frequency of whole book's words

```
book_df <-book%>%unnest_tokens(word, text)


data(stop_words)
book_df <- book_df %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
book_df2 <- book_df
book_df<-book_df %>%
  count(word, sort = TRUE)

book_df <-filter(book_df,!(word %in% c("_ca","de","pp","vol",
                                       "a.d","1","al") ))


book_df%>% with(wordcloud(word, n, max.words = 50, random.order = FALSE, rot.per = 0.35,
                          colors = brewer.pal(8, "Dark2")))
```



We separate the words from every sentence and put them into a new data frame. Then we remove some useless words and draw a word cloud to visualize the frequency of the words. In the plot, we can see "clock" comes up the most, then "water" and "wheel", then "astronomical", "mechanical", "device", "clocks", "time", "motion". So, the word frequency can prove that the content of this book is closely related to the title.

Separate book by chapters

```
chap1num <- grep("Power and Motion Gearing", book$text, ignore.case = F)
chap2num <- grep("Mechanical Clocks", book$text, ignore.case = F)
chap3num <- grep("Perpetual Motion and the Clock before de Dondi", book$text, ignore.case = F)
chap4num <- grep("The Magnetic Compass as a Fellow-traveler from China", book$text, ignore.case = F)
end <- grep("Chronological Chart", book$text, ignore.case = F)

chap1 <- book %>% slice(chap1num:chap2num-1)
chap2 <- book %>% slice(chap2num:chap3num-1)
chap3 <- book %>% slice(chap3num:chap4num-1)
chap4 <- book %>% slice(chap4num:end-1)

chap1$chapter<-rep("Power and Motion Gearing_1",nrow(chap1))
chap2$chapter<-rep("Mechanical Clocks_2",nrow(chap2))
chap3$chapter<-rep("Perpetual Motion and the Clock before de Dondi_3",nrow(chap3))
chap4$chapter<-rep("The Magnetic Compass as a Fellow-traveler from China_4",nrow(chap4))
chap<-bind_rows(chap1,chap2,chap3,chap4)
```

We found that the body of this book has 4 chapters. Then we use "grep" to find the row numbers and

seperate them into different data frames. We number those chapters and bind them together into a new data frame, which represents the main body of this book.

Visualizing a network of bigrams with ggraph

```
book_bigrams <- chap %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  filter(!is.na(bigram))

book_bigrams %>%
  count(bigram, sort = TRUE)
```

```
## # A tibble: 8,677 x 2
##    bigram       n
##    <chr>    <int>
##  1 of the     204
##  2 in the      75
##  3 to the      65
##  4 by the      40
##  5 from the    38
##  6 that the    37
##  7 of a        36
##  8 and the     35
##  9 it is       32
## 10 on the      31
## # ... with 8,667 more rows
## # i Use 'print(n = ...)' to see more rows
```

```
bigrams_separated <- book_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)

library(igraph)
```

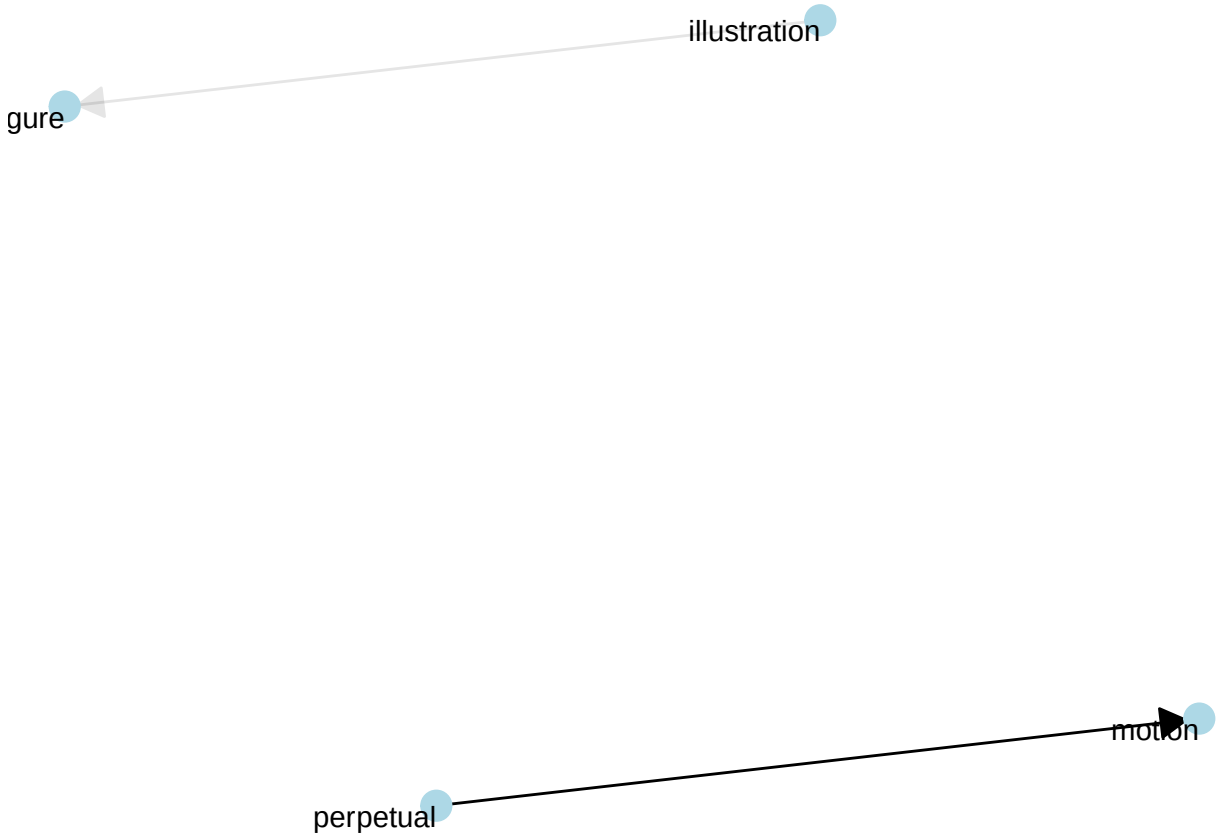
```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:purrr':
##
##     compose, simplify
```

```
## The following object is masked from 'package:tidyr':
##
##     crossing

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
library(ggraph)
bigram_graph <- bigram_counts %>%
  filter(n > 20) %>%
  graph_from_data_frame()

set.seed(2020)

a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()
```



We try to find which two words has the closest connection in this book, so we visualize the network of bigrams. In this plot, we can see, "illustration" is always followed by "figure", and "perpetual" is always followed by "motion".

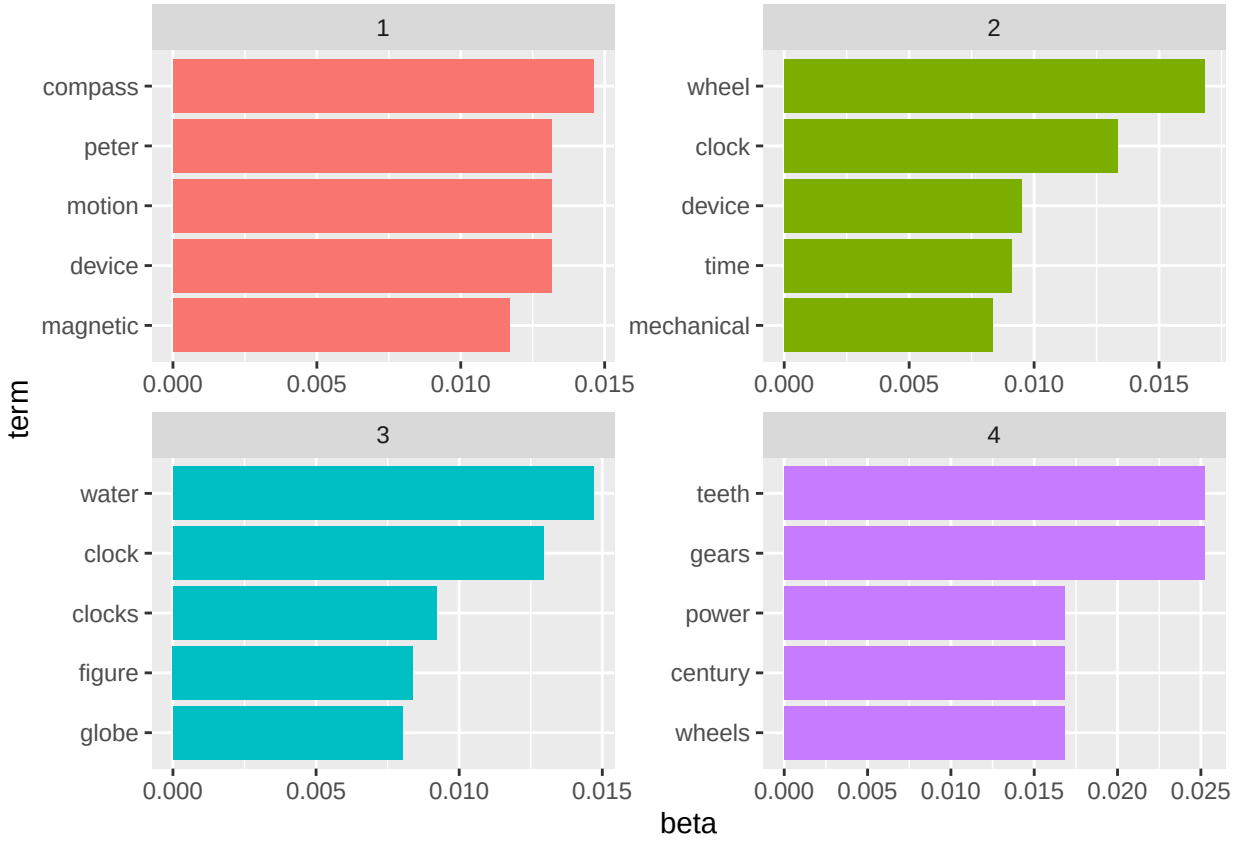
LDA on chapters

```
chap_dtm <- chap %>%
  unnest_tokens(word, text)

chap_dtm <-filter(chap_dtm,!(word %in% c("_ca","de","pp","vol",
                                         "a.d","1","al") ))
chap_dtm %<>%  anti_join(stop_words) %>%
  count(chapter, word) %>%
  cast_dtm(chapter, word, n)
```

```
## Joining, by = "word"
```

```
chap_lda <- LDA(chap_dtm, k = 4, control = list(seed = 1234))
chap_topics<- tidy(chap_lda, matrix ="beta")
chap_terms <- chap_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 5) %>%
  ungroup() %>%
```

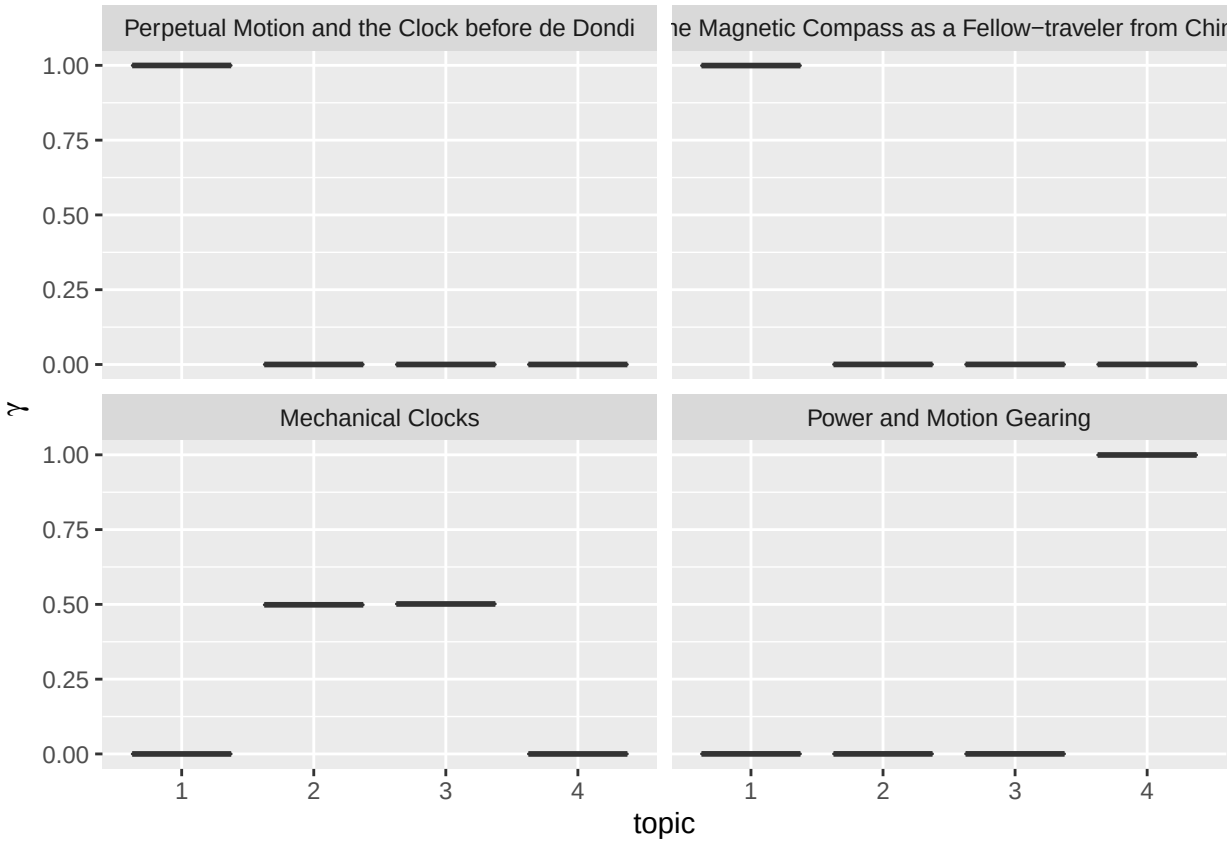
```
  arrange(topic -beta)
chap_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```



From the plot, we can observe that the model and their 4 topics of the book. Topic 1 contains 'compass' and 'magnetic', we suppose that it may be topic about the mechanism of the clock, and second topic is about the wheel and mechanism of clock. Topic 3 is about water,clock and globe, it seems to be correlated with geographic, and topic 4 is about engine and gears.

Pre_document classification

```
chap_gamma <- tidy(chap_lda, matrix = "gamma")
chap_gamma <- chap_gamma %>%
  separate(document, c("title", "chapter"), sep = "_", convert = TRUE)
chap_gamma %>%
  mutate(title = reorder(title, gamma * topic)) %>%
  ggplot(aes(factor(topic), gamma)) +
  geom_boxplot() +
  facet_wrap(~ title) +
  labs(x = "topic", y = expression(gamma))
```



It seems like the first and second parts are all about the topic 1, while the Mechanical Clocks is about topic 2 and 3 with percentage of 50% and 50%. Power and Motion Gearing is all about topic 4.

By word assignments

```
chap_classifications <- chap_gamma %>%
  group_by(title, chapter) %>%
  slice_max(gamma) %>%
  ungroup()

chap_topics <- chap_classifications %>%
  count(title, topic) %>%
  group_by(title) %>%
  slice_max(n, n = 1) %>%
  ungroup() %>%
  transmute(consensus = title, topic)

chap_classifications %>%
  inner_join(chap_topics, by = "topic") %>%
  filter(title != consensus)
```

```
## # A tibble: 2 x 5
##   title                                                chapter topic gamma conse~1
##   <chr>                                                  <int> <int> <dbl> <chr>
```

```
## 1 Perpetual Motion and the Clock before de Dondi           3     1 1.00  The Ma~
## 2 The Magnetic Compass as a Fellow-traveler from Ch~       4     1 0.999 Perpet~
## # ... with abbreviated variable name 1: consensus
```

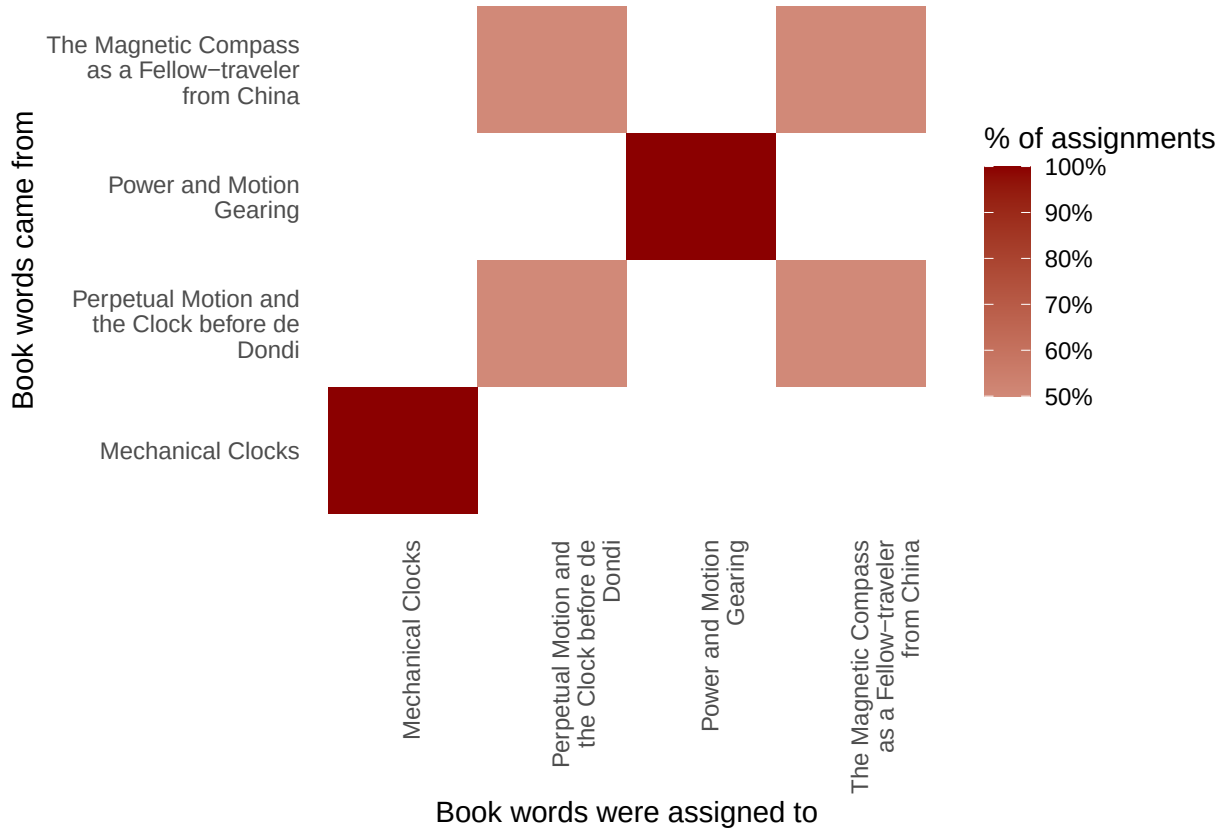
```
assignment <- augment(chap_lda, data = chap_dtm)
assignment <- assignment %>%
  separate(document, c("title", "chapter"),
           sep = "_", convert = TRUE) %>%
  inner_join(chap_topics, by = c(".topic" = "topic"))


library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
assignment %>%
  count(title, consensus, wt = count) %>%
  mutate(across(c(title, consensus), ~str_wrap(., 20))) %>%
  group_by(title) %>%
  mutate(percent = n / sum(n)) %>%
  ggplot(aes(consensus, title, fill = percent)) +
  geom_tile() +
  scale_fill_gradient2(high = "darkred", label = percent_format()) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        panel.grid = element_blank()) +
  labs(x = "Book words were assigned to",
       y = "Book words came from",
       fill = "% of assignments")
```



From the plot, we can also observe that Perpetual Motion and the Clock before de Dondi and The Magnetic Compass as a Fellow-traveler from China are in a very similar topic.