NLP

Kristof Menyhert

Load some packages:

```
library(gutenbergr)
library(tidyr)
library(tidyverse)
library(stringr)
library(tidytext)
library(tm)
library(topicmodels)
library(ggplot2)
```

Download some books:

```
books <- gutenberg_works(title %in% c("Oliver Twist",
                                      "Candide", "Dracula",
                                      "The Adventures of Sherlock Holmes")) %>%
  gutenberg_download(meta_fields = "title")

unique(books$title)
```

```
## [1] "Dracula"                           "Oliver Twist"
## [3] "The Adventures of Sherlock Holmes" "Candide"
```

divide into documents each representing one chapter

```
reg <- regex("^chapter ", ignore_case = TRUE)

by_chapter <- books %>%
  group_by(title) %>%
  mutate(chapter = cumsum(str_detect(text, reg))) %>%
  ungroup() %>%
  filter(chapter > 0) %>%
  unite(document, title, chapter)
```

Split into words

```
by_chapter_words <- by_chapter %>%
  unnest_tokens(word, text)
```

find document word count

```
word_counts <- by_chapter_words %>%
  anti_join(stop_words) %>%
  count(document, word, sort = TRUE) %>%
  ungroup()
```

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

LDA on chapters

```r
chapters_dtm <- cast_dtm(word_counts, document, word, n)
```

Create k model model (now it is 4 as we have 4 books)

```r
chapters_lda <- LDA(chapters_dtm, k = 4, control = list(seed = 123))
chapters_lda
```

```
## A LDA_VEM topic model with 4 topics.
```

```r
chapter_topics <- tidy(chapters_lda, matrix = "beta")
```

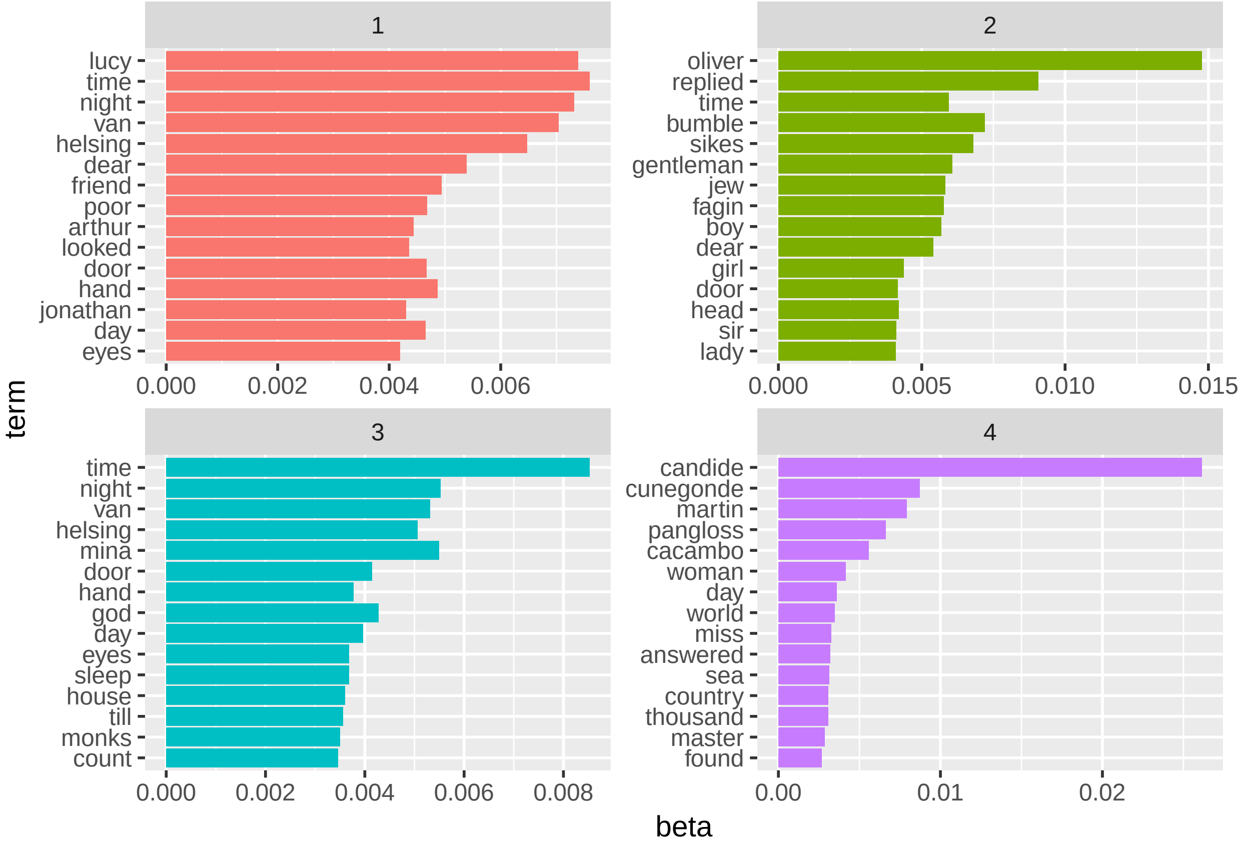
This turned the model into one-topic-per-row format. For each combination, the model computes the probability of that term being generated from that topic.

Find top top five terms within each topic:

```r
top_terms <- chapter_topics %>%
  group_by(topic) %>%
  top_n(15, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

Plot the results:

```r
top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = F) +
  facet_wrap(~topic, scales = "free") +
  coord_flip()
```



Based on these graph we can guess associate which topics goes to which books.

- 1st goes to Dracula
- 2nd goes to Oliver Twist
- 3rd goes to Sherlock Holmes
- 4th goes to Candide