LDA Modelling of 'The Brothers Karamazov'

Casey Crary

2024-03-03

'The Brothers Karamazov' has several distinct themes throughout the book, such as a focus on God and religion, as well as the nature of morality. LDA modelling may allow us to quantify some of these themes, and see in which parts of the book they are more prevalent.

```
BrothersKaramazov_words <- BrothersKaramazov |>
  filter(book != 0) |>
  unnest_tokens(word, text, token = "words") |>
  anti_join(stop_words, join_by(word)) |>
  group_by(book, word) |>
  count() |>
  ungroup()
```

To perform LDA modelling with the text of the Brothers Karamazov, we first need to `unnest_tokens` to get the words by themselves. We chose to group by sub-books here, but we could also perhaps look at parts or chapters of the book. We also removed the first book, as it is really just the table of contents.

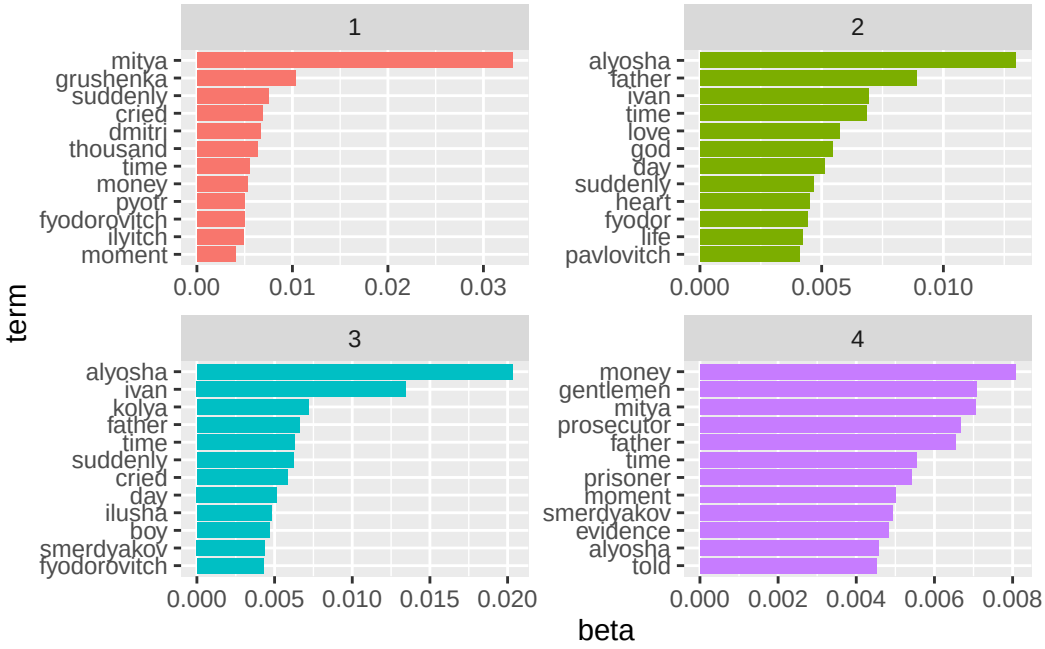
```
bk_dtm <- BrothersKaramazov_words |>
  cast_dtm(book, word, n)

k_val <- 4
bk_lda <- LDA(bk_dtm, k = k_val, control = list(seed = 325))

bk_topics <- tidy(bk_lda, matrix = "beta")

bk_top_terms <- bk_topics |>
  group_by(topic) |>
  slice_max(beta, n = 12) |>
  ungroup() |>
  arrange(topic, -beta)
```

```
bk_top_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()
```



Since there are 4 main parts to the book, having 4 topics makes a reasonable amount of sense. Topic 1 is focused on Mitya, a shortening of Dmitri, which is also present in the list. It includes the name of his lover "Grushenka," and refence to the three "thousand" Rubles he allegedly stole. It also perhaps is about the patricide he allegedly commits, considering words such as "cried" appear. Topic 2 seems to be about the time Alyosha spends in the monastery, with words like "love" and "elder" showing up, two words associated with Elder Zossima, the eldest monk at the monastery. Topic 3 seems to be about Alyosha's interactions with Kolya and Ilusha, two school children Alyosha gets involved in conversation with. Topic 4 seems to be about Mitya again, although this time about the actual trial he is on for the patricide. Words like "evidence" and "prosecutor" make this topic more clear than the others.

Notice that three of these four topics have Alyosha present in them. Although he is the main character, and so it makes sense he would appear commonly, perhaps we could get more information if we exclude him and some other common words from the modelling.

```
common_bk_words <- tibble(word = paste(c("alyosha", "ivan", "it's",
                                         "suddenly", "don't")))
BrothersKaramazov_words <- BrothersKaramazov_words |>
  anti_join(common_bk_words)
```

```
Joining with `by = join_by(word)`
```

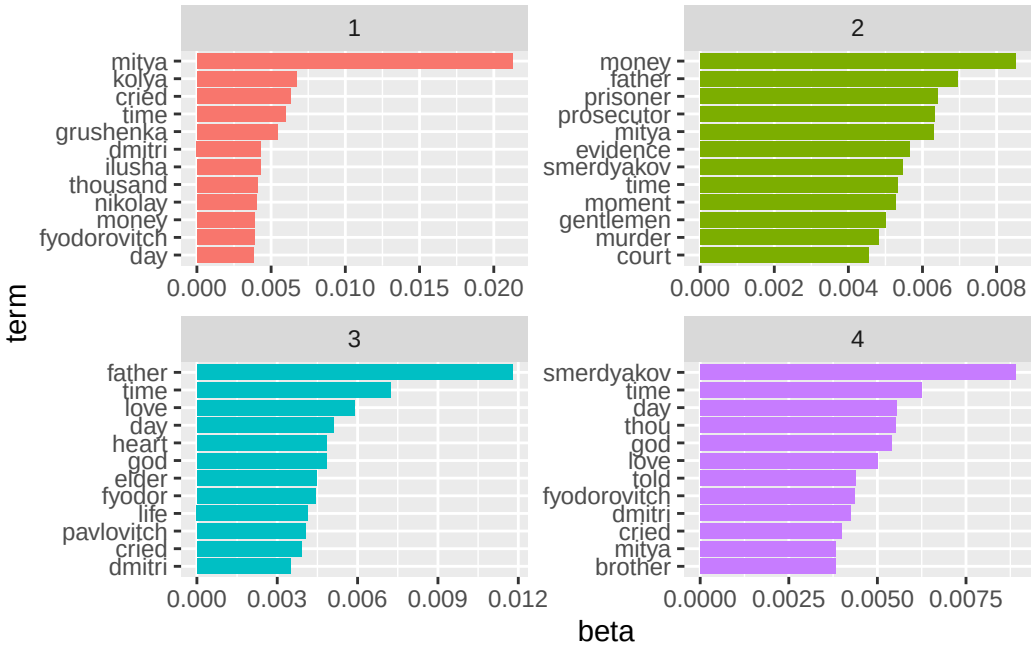
```
bk_dtm <- BrothersKaramazov_words |>
  cast_dtm(book, word, n)

k_val <- 4
bk_lda <- LDA(bk_dtm, k = k_val, control = list(seed = 325))

bk_topics <- tidy(bk_lda, matrix = "beta")

bk_top_terms <- bk_topics |>
  group_by(topic) |>
  slice_max(beta, n = 12) |>
  ungroup() |>
  arrange(topic, -beta)

bk_top_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()
```



The topics here are very similar to what they were above. Topic 1 still focuses on Mitya and his various crimes. Topic 2 seems to have swapped to what topic 4 was about, the trial. Topic 3 still focuses on the chapters about the monastery and Elder Zosima. Topic 4 is the main change in the topics after removing Alyosha. Instead of having a topic about Alysoha and the school children, there is now a topic that seems to just have a variety of popular words that don't appear in any other topic, such as the character "Smerdyakov."

```
bk_lda4 <- LDA(bk_dtm, k = 4, control = list(seed = 1821))
bk_gamma <- tidy(bk_lda4, matrix = "gamma")
bk_gamma |> filter(gamma > 0.001) |>
  arrange(readr::parse_number(document)) |>
  knitr::kable()
```

document	topic	gamma
1	3	0.9999793
2	1	0.9999902
3	1	0.0580286
3	2	0.7918868
3	3	0.1500817
4	1	0.8477667
4	4	0.1522252

document	topic	gamma
5	3	0.9999916
6	1	0.9623122
6	3	0.0376792
7	1	0.9999850
8	4	0.9999931
9	4	0.9999911
10	1	0.9999877
11	2	0.9803118
11	4	0.0196837
12	2	0.9999952

From the gamma values for each book, we can see what topic our model think it fits with best. Only gammas greater than 0.01 are shown here, so we can assume if a certain document-topic pair doesn't occur the model thinks it is very unlikely that book is in that topic. The model puts most books into just one topic with very high likelihood (above 99%). Three of the twelve books are put in two topics with the highest secondary likelihood being for book 4, where it has a gamma of .84 and .15 for topics 1 and 4, respectively. One of the books is put in three of the topics, book 3. Exploring this could be interesting.

```
bk_gamma |>
  group_by(document) |>
  summarize(sd_gamma = sd(gamma), document = first(document)) |>
  arrange(sd_gamma) |>
  knitr::kable()
```

document	sd_gamma
3	0.3665045
4	0.4049201
6	0.4752068
11	0.4869629
1	0.4999862
7	0.4999900
10	0.4999918
2	0.4999934
9	0.4999941
5	0.4999944
8	0.4999954
12	0.4999968

From the above table we can see Book 3's gammas have the lowest standard deviation, so it is the book most split between multiple topics. What book is this?

```
filter(BrothersKaramazov, book == 3) |> select(text) |> head()
```

```
# A tibble: 6 x 1
  text
  <chr>
1 Book III. The Sensualists
2 Chapter I.
3 In The Servants' Quarters
4 The Karamazovs' house was far from being in the center of the town, but
5 it was not quite outside it. It was a pleasant-looking old house of two
6 stories, painted gray, with a red iron roof. It was roomy and snug, and
```

The most ambiguous book topic-wise is "Book III. The Sensualists." This book focuses on the two characters described as "sensualists" in the work, mainly Fyodor, the father, but also Mitya, his son. It seems this is why it got placed in mainly topics 2 and 3 since they are focused on these two characters.