LDA Workshop

Denis Bulygin, Stanislav Pozdniakov, Vadim Voskresensky

04 03 2017

The research was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017 — 2018 (grant No. 17-05-0024) and by the Russian Academic Excellence Project "5-100".

Intro

```
library(quanteda)
library(tidyr)
library(topicmodels)
library(tidytext)
library(dplyr)
library(stringr)
```

```
df = read.csv("/home/voskresenskiiv/lda_lab/un_news.csv") #
#сообщения, полным текстом, датой
```

2/14

Document-feature matrix

```
stop_words = stopwords("english")
df$story = df$story %>% as.character() # делаем текстовую пе
myDfm = dfm(df$story,stem = T, removeNumbers = TRUE,
            remove = stop_words, removePunct = TRUE)
# создаем document-feature matrix,
# в которой пересечения между нашими токенами и
# документами, при
# помощи аргумента stem проводим стемминг (приводим слова к
# а при помощи аргумента remove удаляем стоп-слова
```

LDA

```
# ap_lda = LDA(myDfm,k=20,control = list(seed = 1234,verbose
# делаем тематическую модель, k отвечает за кол-во топиков
# control включает в себя ряд параметров,
# например, seed это идентификатор, который позволяет при же
# восстановить именно эту модель в будущем, а verbose отража
#основные этапы работы алгоритма

load('/home/voskresenskiiv/lda_lab/lda_model.rda')
ap_lda_td <- tidy(ap_lda)
# превращаем результаты моделирования в датасет:
# первая колонка -- номер топика (их у нас 50),
# вторая -- токен, третья -- вероятность
ap_gamma <- tidy(ap_lda, matrix = "gamma")
# делаем документ-топик датасет:
# первая колонка -- номер документа,
# вторая -- номер топика, третья -- вероятность
```

LDA

```
topic.per.word = ap_lda_td %>% spread(topic,beta)
# делаем матрицу, строчки -- токены, столбцы -- топики,
# на пересечении -- вероятность
vocabulary = topic.per.word$term
vocabulary = as.character(vocabulary)
rownames(topic.per.word) = topic.per.word$term
topic.per.word = select(topic.per.word,-1) %>% t() %>% as.ma
# теперь у нас есть матрица, в которой строчки -- топики,
# столбцы -- токены, на пересечении -- вероятность
```

LDA

```
topic.per.doc = ap_gamma %>% spread(topic, gamma)
rownames(topic.per.doc) = topic.per.doc$document
topic.per.doc = select(topic.per.doc,-1) %>% as.matrix()
# а это вторая важная матрица для нас, в которой строчки --
# столбцы -- топики, пересечение -- вероятность

wordcounts = myDfm %>% tidy() %>% group_by(term) %>% filter(
doc.length = df %>% unnest_tokens(word,story) %>% group_by(i
doc.length = doc.length$n
```

6/14

LDAvis

```
# визуализируем наш LDA
# размер кружка показывает, сколько прцоентов
#текстового корпуса на него приходится
# дистанция также важна; чем ближе топики,
#тем они более похожи друг на друга
#topic.per.word[topic.per.word==0] = 0.00000000000000001 #та
#library(LDAvis)
#library(servr)
#json <- createJSON(phi = topic.per.word, theta=topic.per.do
#doc.length=doc.length, vocab=vocabulary, term.frequency=wor
#serVis(json, out.dir="lda100", open.browser=TRUE)
```

Интерпретация

```
#ap_gamma %>% filter(topic == 6) %>% arrange(-gamma) %>% top
#df$story[470]
#df$story[698]
#ap_gamma %>% filter(topic == 15) %>% arrange(-gamma) %>% to
#df$story[739]
```

8/14

Интерпретация

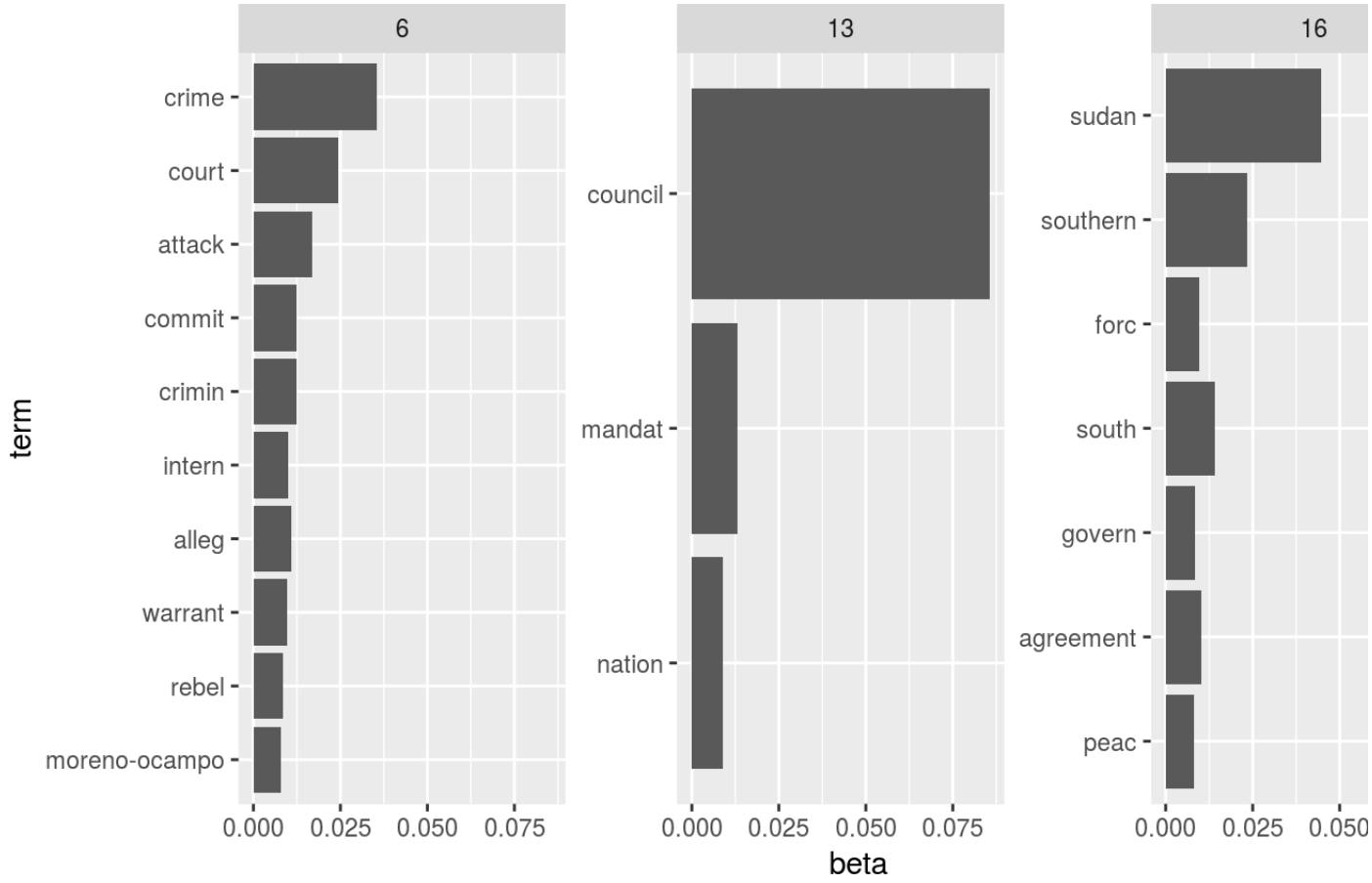
```
topic.words = ap_lda_td %>% group_by(topic) %>% top_n(20)
doc.topics = ap_gamma %>% group_by(document) %>% top_n(3)
```

```
library(ggplot2)
p4 = ap_lda_td[ap_lda_td$topic==c(16,13,6),] %>%
  mutate(term = reorder(term, beta)) %>% top_n(20) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free_y") +
  coord_flip()
```

9/14

Рисуем

p4



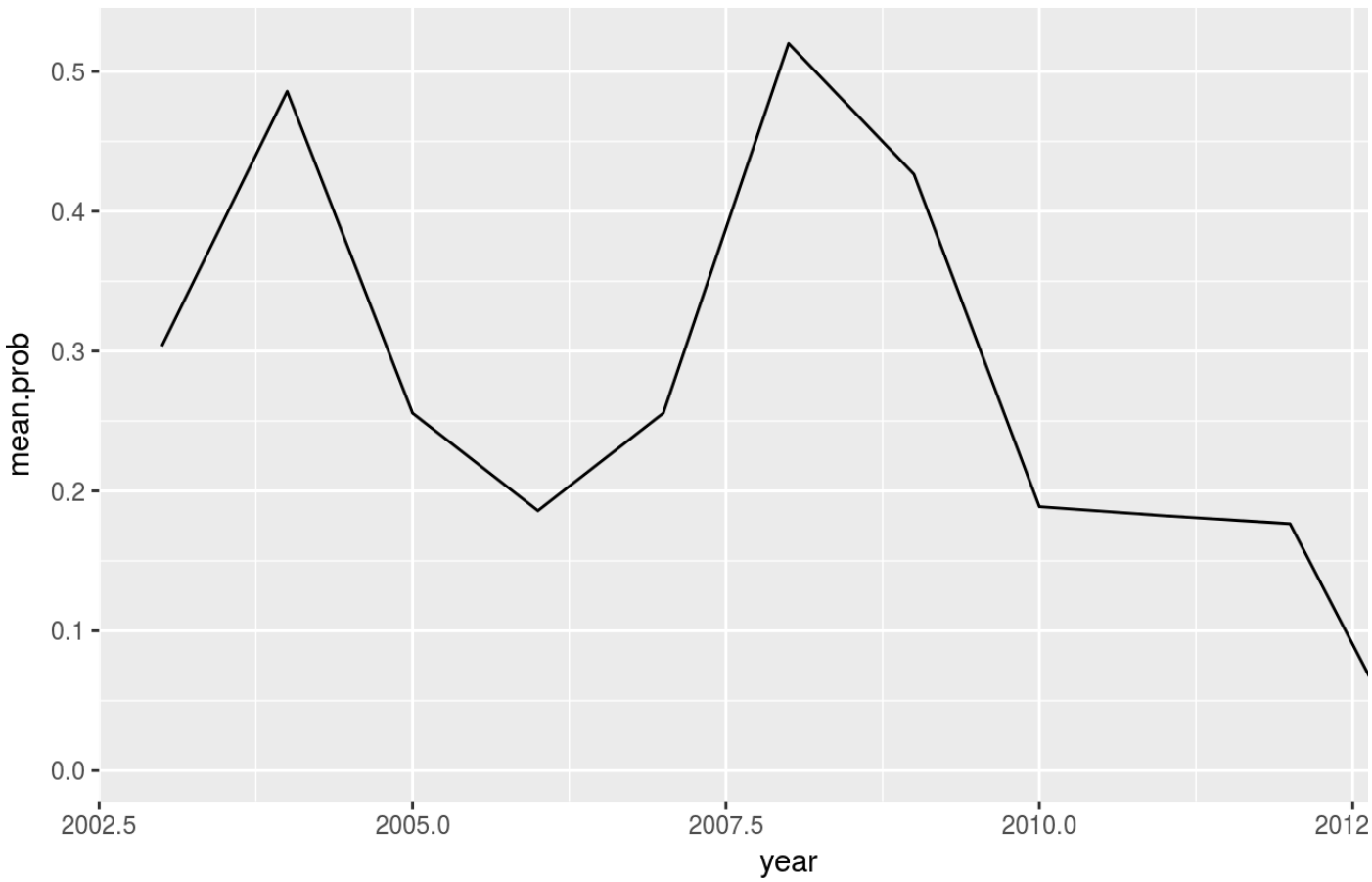
10/14

Добавим некоторые метаданные и порисуем снова

```
df$document = str_c("text",df$id)
doc.topics = left_join(doc.topics,select(df,year, document))
doc.topics = doc.topics %>%  group_by(topic,year) %>% summar
doc.topics$year = doc.topics$year %>% as.integer()
```

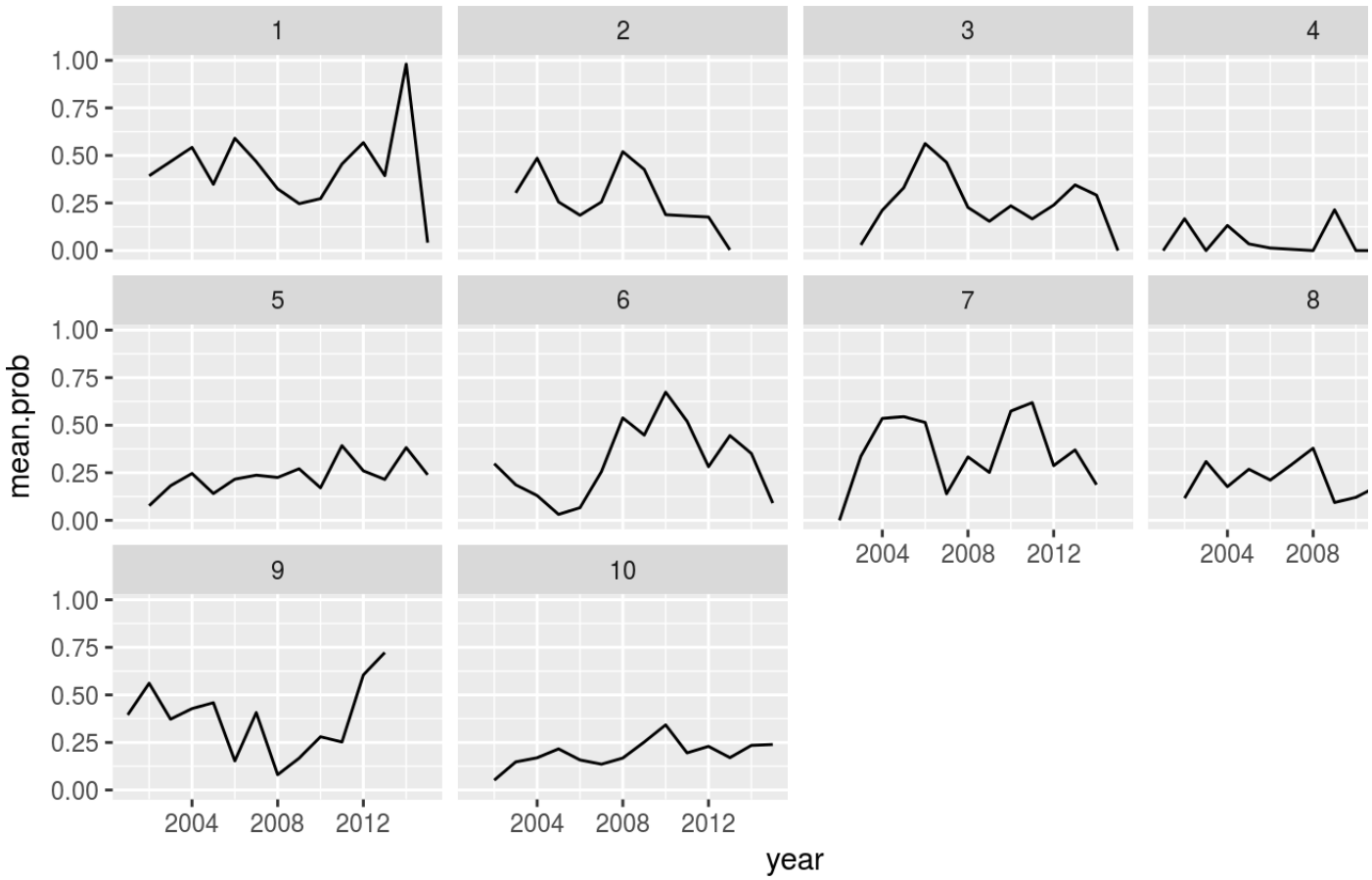
11/14

```
ggplot(data=doc.topics[doc.topics$topic == 2,]) + geom_line(
```



12/14

```
ggplot(data=doc.topics[doc.topics$topic < 11,]) + geom_line(
```



13/14

```
#ggplot(data=doc.topics[doc.topics$topic < 11,]) + #geom_lin
```

14/14