

Отчет о пакете LDA* среды разработки R

© И. Р. Шаймарданов

23 апреля 2012

Аннотация

Этот обзор вводит читателя в сферу тематического моделирования и на примерах показывает как работать в системе R с пакетом LDA

Содержание

1	Описание алгоритма	2
2	Функции пакета LDA	4
3	Пример использования пакета	5

*Описание разработчиков: This package implements latent Dirichlet allocation (LDA) and related models. This includes (but is not limited to) sLDA, corrLDA, and the mixed-membership stochastic blockmodel. Inference for all of these models is implemented via a fast collapsed Gibbs sampler written in C. Utility functions for reading/writing data typically used in topic models, as well as tools for examining posterior distributions are also included.

1 Описание алгоритма

Что такое тематическое моделирование? Это область машинного обучения, оперирующая большими корпусами текстовых документов и пытающаяся выделить в них тематики. Тематики формально определены как вероятностные распределения над множеством слов. В тексте на рис. 1 можно выделить тематику, связанную с анализом данных (выделено синим) со словами «computer», «prediction» и т.д., тематику, связанную с процессом эволюции (выделено розовым) со словами «life», «organism» и т.д., тематику, связанную с генетикой (выделено желтым) со словами «genes», «sequenced») и проч.

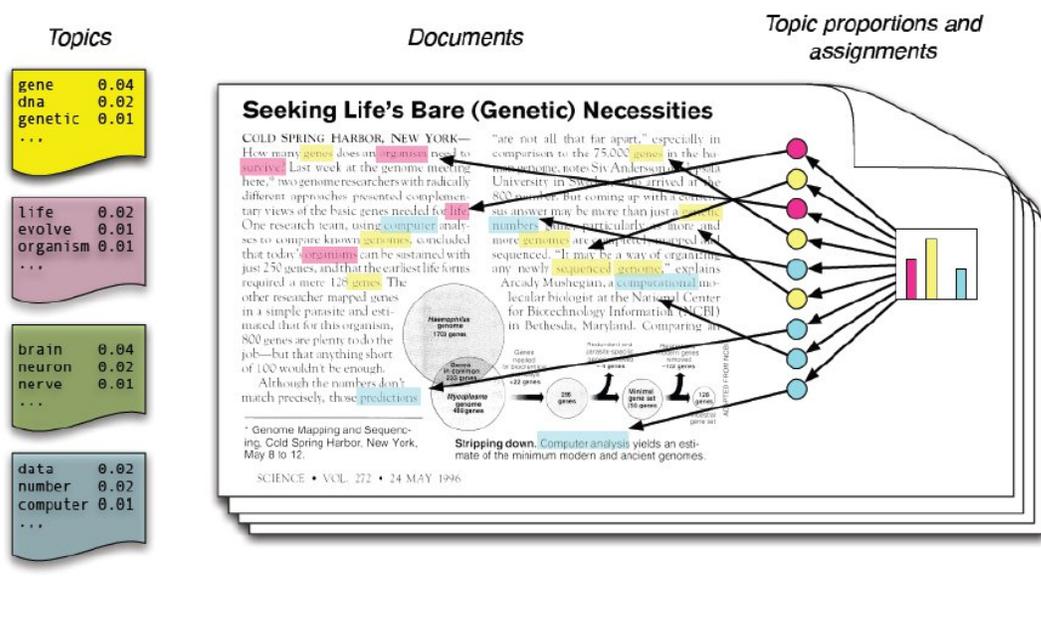


Рис. 1: Иллюстрация порождающей модели текста с помощью набора тематик

LDA — это вероятностная модель порождения текста, которая каждому документу определяет распределение над множеством тем.

Введем формальные обозначения для всех упомянутых понятий, таких как документы, тематики, слова в документе, тематики для каждого слова в документе и т.д.

$\omega \in 1, \dots, W$	номер слова в словаре
$t \in 1, \dots, T$	номер тематики
$d \in 1, \dots, D$	номер документа в корпусе
N_d	числов слов в документе
$\omega_d = [\omega_{d,1}, \dots, \omega_{d,N_d}]$	слова в документе d , $\omega_{d,n} \in 1, \dots, W$
$z_d = [z_{d,1}, \dots, z_{d,N_d}]$	тематики слов в документе d , $z_{d,n} \in 1, \dots, T$
$\theta_d = [\theta_{d,1}, \dots, \theta_{d,T}]$	вероятности тематик в документе d
$\phi_t = [\phi_{t,1}, \dots, \phi_{t,W}]$	вероятности слов в тематике t
$\Theta = [\theta_1, \dots, \theta_D]^T \in \mathbb{R}^{D \times T}$	вероятности тематик во всех документах
$\Phi = [\phi_1, \dots, \phi_T]^T \in \mathbb{R}^{T \times W}$	вероятности слов во всех тематиках
$\mathcal{W} = \omega_1, \dots, \omega_D$	набор всех слов в корпусе
$\mathcal{Z} = z_1, \dots, z_D$	разбиение всех слов по тематикам

Вероятностная модель LDA задается как

$$p(\mathcal{W}, \mathcal{Z}, \Theta | \Phi, \alpha) = \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(\omega_{d,n} | z_{d,n}, \Phi) p(z_{d,n} | \theta_d)$$

$$p(\theta_d | \alpha) = \text{Dir}(\theta_d | \alpha)$$

$$p(\omega_{d,n} | z_{d,n}, \Phi) = \Phi_{z_{d,n}, \omega_{d,n}}, p(z_{d,n} | \theta_d) = \Theta_{d, z_{d,n}}$$

Матрицы Θ и Φ неизвестны и их нужно оценить по корпусу документов. В данном пакете реализован алгоритм сэмплирования Гиббса.

Алгоритм сэмплирования Гиббса

Вход: коллекция $X = (d, w) : d \in D, w \in d$; параметры α, β, M - количество сэмплов;

Выход: оценки $\hat{p}(w|t), \hat{p}(t|d)$;

- 1: $n_{wt} := \beta; n_{td} := \alpha$ для всех $d \in D, w \in W, t \in T$;
- 2: $n_t := \beta|W|; n_d := \alpha|T|$ для всех $d \in D, t \in T$;
- 3: для $i = 1, \dots, M$ // генерация M сэмплов
- 4: для всех документов $d \in D$ и всех слов $w \in d$
- 5: $\tilde{p}(t|d, w) := (n_{wt}/n_t)(n_{td}/n_d)$ для всех $t \in T$;
- 6: если $i \geq 2$ то
- 7: $t := t_{dw}; \quad - - n_{wt}; \quad - - n_{td}; \quad - - n_t; \quad - - n_d$;

- 8: выбрать t из ненормированного распределения $\tilde{p}(t|d, w)$;
- 9: $t_{dw} := t; ++ n_{wt}; ++ n_{td}; ++ n_t; ++ n_d$;
- 10: $\hat{p}(w|t) := n_{wt}/n_t$ для всех $w \in W, t \in T$;
- 11: $\hat{p}(t|d) := n_{td}/n_d$ для всех $d \in D, t \in T$;

Продолжить знакомство с тематическими моделями можно обратившись к следующим литературным источникам [1, 2, 3]

2 Функции пакета LDA

Пакет в качестве корпуса документов принимает список матриц размера $2 \times N_d$. Каждый элемент списка соответствует одному документу. В каждой матрице первая строка соответствует номерам слов в словаре. Вторая строка соответствует количествам вхождений слов в документ. Для создания корректного корпуса документов пакет предлагает удобные функции

`lexicalize`, `read.documents`, `read.vocab`

В этот отчет, в целях краткости изложения, не вошли описания этих и некоторых других функций. Интересующийся читатель может их найти в [6].

Основная функция настройки модели `lda.collapsed.gibbs.sampler` вызывается следующим образом

```
result<-lda.collapsed.gibbs.sampler(documents, K, vocab, num.iterations, alpha,
eta, initial = NULL, compute.log.likelihood = FALSE)
```

Аргументы:

<i>documents</i>	корпус документов описанного выше формата
<i>K</i>	количество тематик
<i>vocab</i>	вектор слов корпуса(словарь)
<i>num.iterations</i>	количество проходов сэмплера по корпусу
<i>alpha</i>	параметр распр. Дирихле для распределения документов по темам
<i>eta</i>	параметр распр. Дирихле для распределения тем по словам
<i>initial</i>	иницилирующая матрица соответствий тем и слов
<i>compute.log.likelihood</i>	вычислять ли правдоподобие модели на обучающей выборке

Возвращаемые данные - список со следующими полями:

<i>assignments</i>	список длины D , содержащий вектора соответствий слов темам
<i>topics</i>	матрица размера $K \times V$, содержащая количества приписаний теме определенного слова
<i>topic_sums</i>	вектор, содержащий количества приписаний всех слов теме
<i>documents_sums</i>	матрица размера $K \times D$ содержащая количества приписаний теме слов из документа
<i>log_likelihoods</i>	матрица из двух строк, содержащая поитерационные значения правдоподобия. Первая строка содержит значения полной функции правдоподобия. Вторая строка содержит значения условной функции правдоподобия.

3 Пример использования пакета

Пакет имеет встроенную базу данных cora - корпус научных статей поисковой системы Cora. Загрузим ее в память

```
data(cora.documents)// documents' corpus
data(cora.vocab)// corpus' vocabulary
```

Определим количество тем

```
K<-10
```

Обучим тематическую модель

```
result <- lda.collapsed.gibbs.sampler(cora.documents,
                                     K, ## Num clusters
                                     cora.vocab,
                                     25, ## Num iterations
                                     0.1,
                                     0.1,
                                     compute.log.likelihood=TRUE)
```

Найдем пятерки самых встречаемых слов в темах

```
top.words <- top.topic.words(result$topics, 5, by.score=TRUE)
```

Выведем на экран тематики первых 10 документов

```
N <- 10## Количество документов для вывода
topic.proportions <- t(result$document_sums) / colSums(result$document_sums)
topic.proportions <- topic.proportions[sample(1:dim(topic.proportions)[1], N),]
topic.proportions[is.na(topic.proportions)] <- 1 / K
colnames(topic.proportions) <- apply(top.words, 2, paste, collapse=" ")
topic.proportions.df <- melt(cbind(data.frame(topic.proportions),
                                   document=factor(1:N)),
                             variable_name="topic",
                             id.vars = "document")

theme_set(theme_bw())
qplot(topic, value, fill=document, ylab="proportion",
       data=topic.proportions.df, geom="bar") +
  opts(axis.text.x = theme_text(angle=90, hjust=1)) +
  coord_flip() +
  facet_wrap(~ document, ncol=5)
```

Получили рис. 2

Список литературы

- [1] Воронцов К. В. [Тематические модели для коллекции текстов](#) — презентация лекции К. В. Воронцова (МФТИ, ВМК МГУ, ШАД Яндекс, 2011).
- [2] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). — [Latent Dirichlet allocation](#)
- [3] Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad. Knowledge discovery through directed probabilistic topic models: [Перевод на русский язык](#)
- [4] Воронцов К. В. [L^AT_EX 2_ε в примерах](#) — Москва, 2005, 55 стр.
- [5] Кнут Д. Всё про T_EX. — Протвино, RDT_EX, 1993.
- [6] [Документация разработчиков по пакету lda.](#)

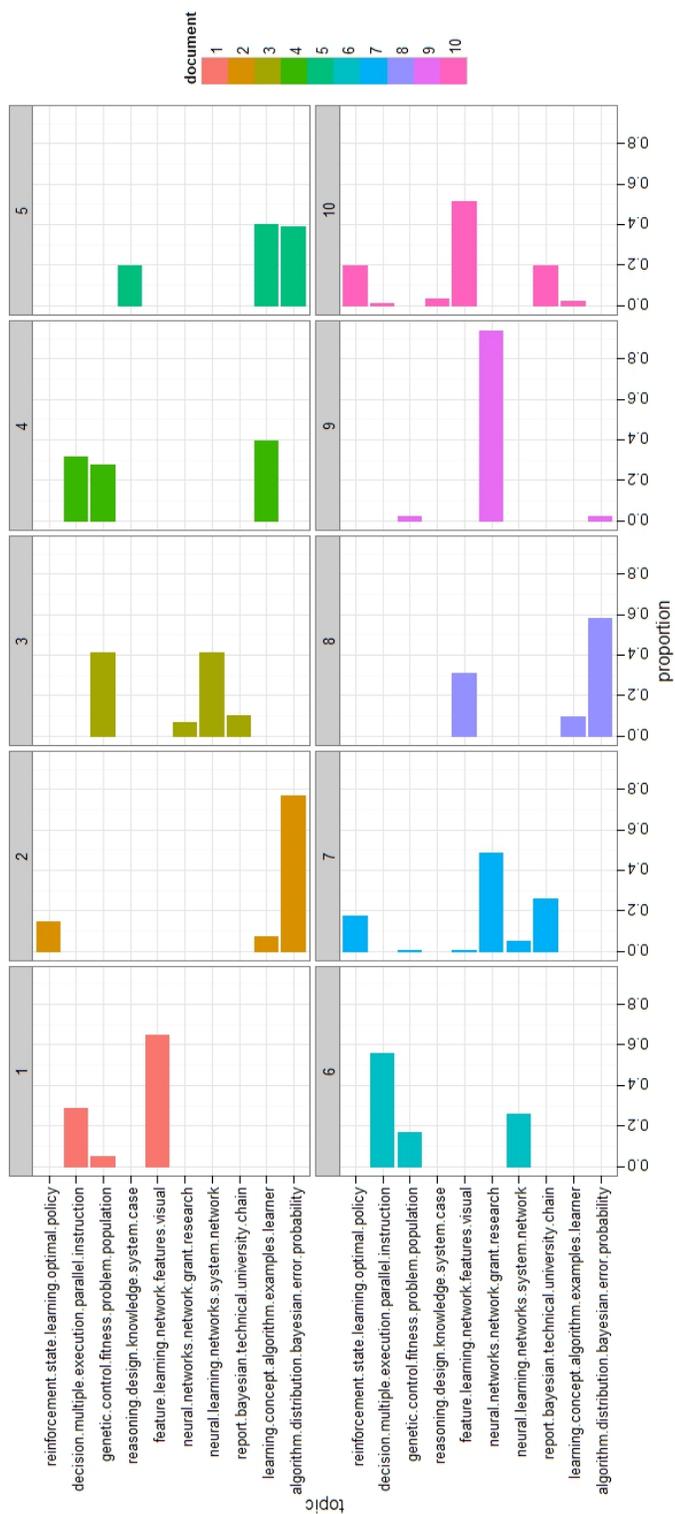


Рис. 2: Распределения первых 10 документов по темам