УДК 519.76

# Тематическое моделирование текстовых учебных материалов по информатике средствами языка R

*Р.Е. Ерланова, А.Б. Нугуманова, Ж.З. Жантасова, Е.М. Байбурин*

Восточно-Казахстанский государственный университет им. С. Аманжолова (Усть-Каменогорск, Казахстан)

# Topic Modeling for Textual Learning Materials on Informatics Using R Language

*R.Ye. Yerlanova, A.B. Nugumanova, Zh. Z. Zhantassova, Ye. M. Baiburin*

S. Amanzholov East Kazakhstan State University (Ust-Kamenogorsk, Kazakhstan)

В работе представлены результаты тематического моделирования текстовых учебных материалов. Учебные материалы являются электронными конспектами лекций, используемых преподавателями для подготовки к занятиям по информатике. Методы тематического моделирования позволяют без дополнительной ручной работы систематизировать содержание текстовых документов, выделить в них главные темы и показать, как эти темы распределены внутри документов. Другими словами, эти методы позволяют сформировать так называемую тематическую модель, которая ставит в соответствии с заданной коллекцией документов набор тем, характеризующих содержание документов из этой коллекции. В качестве метода тематического моделирования используется латентное размещение Дирихле, а в качестве среды для реализации метода — язык R. Разработанное веб-приложение является интерактивным и предоставляет пользователю (преподавателю) набор визуальных инструментов тематического моделирования. Благодаря визуализации улучшается эргономика работы с учебными материалами, экономится время, затрачиваемое на изучение, анализ, подбор соответствующей учебной литературы.

***Ключевые слова:*** тематическое моделирование, обработка естественного языка, вероятностные языковые модели, латентное размещение Дирихле, R.

This paper presents results of topic modeling for text learning materials. Learning materials are electronic lecture notes used by teachers to prepare for computer science classes. Topic modeling methods allow users to systematize the content of textual documents without additional manual work. Main topics in documents are highlighted, and the distribution of topics in documents is demonstrated. In other words, the proposed methods provide the framework for the so-called topic model that puts a set of topics that characterize the content of documents in a given collection of documents. The latent Dirichlet allocation (LDA) is used for topic modeling. The implementation is done using the R language. The developed interactive web application provides a set of visual tools for topic modeling to a user (teacher). Visualization techniques gradually improve the ergonomics of a teacher's work with learning materials and save the time spent on studying, analyzing, and selecting relevant study materials.

***Key words:*** topic modeling, natural language processing, probabilistic language models, latent Dirichletallocation, R.

**DOI 10.14258/izvasu(2018)4-12**

**Введение.** Доля цифровых образовательных ресурсов в общем объеме образовательных программ неуклонно увеличивается. Соответственно, традиционные способы работы с такими ресурсами теряют свое значение и заменяются новыми подходами, среди которых важное место занимают методы автоматического анализа и структурирования контента, направленные на минимизацию времени, затрачиваемого на чтение и понимание текста. В работе [1] отмечается, что огромные коллекции цифровых ресурсов, с одной стороны, обеспечивают беспрецедентный доступ к источникам, с другой стороны, побуждают и даже требуют использования новых способов анализа этих источников для получения на их основе фактографической информации и формирования выводов.

Одним из таких перспективных способов семантического анализа цифровой коллекции документов является тематическое моделирование, которое позволяет выделить важнейшие темы коллекции и показать, как распределены эти темы в документах [2–6]. В данной работе в качестве метода тематического моделирования рассматривается латентное распределение Дирихле (LDA), на основе которого создается вероятностная модель, определяющая, к каким темам ($t \in T$) относится каждый документ ($d \in D$) и какие ключевые слова ($w \in W$) описывают каждую тему:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d), \qquad (1)$$

где $p(w|d)$ — известное распределение слов в документах рассматриваемой коллекции, а $p(w|t)$ и $p(t|d)$ — это неизвестные распределения слов по темам и тем по документам соответственно, $T$ — множество тем. Обозначим $\varphi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$. Тогда матрицы $\Phi = ||\varphi_{wt}||_{w \in W, t \in T}$ и $\Theta = ||\theta_{td}||_{w \in W, t \in T}$, которые являются параметрами тематической модели, находятся путем решения задачи максимизации функции (логарифма) правдоподобия:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\varphi, \theta}, \qquad (2)$$

где $n_{dw}$ — частота слова $w$ в документе $d$, $\varphi_{wt} \geq 0$, $\theta_{td} \geq 0$, $\sum_{t \in T} \varphi_{wt} = 1$, $\sum_{t \in T} \theta_{td} = 1$.

Целью данной работы является практическая реализация метода тематического моделирования LDA на примере коллекции учебных материалов с помощью экосистемы языка R. Для этого в экосистеме R разработан целый ряд библиотек, из которых мы используем три основные группы: библиотеки функций, выполняющих процессинг естественно-языковых текстов (tm, quanteda); библиотеки функций, ответственных за построение модели LDA и ее визуализацию (lda, topicmodels, LDAvis); библиотеки, предназначенные для создания интерактивных веб-приложений на базе R (мы используем shiny).

Таким образом, в соответствии с поставленной целью были определены следующие этапы реализации:

1. Формирование коллекции учебных материалов и ее предобработка.

2. Построение тематической модели и ее визуализация.

3. Разработка веб-интерфейса.

**1. Формирование коллекции учебных материалов и ее предобработка.** В качестве входных учебных материалов были взяты три источника на английском языке, т.к. в настоящее время в вузах Республики Казахстан преподавание информационно-коммуникационных технологий (ИКТ) ведется на английском языке. Первый источник — электронный учебник «Основные понятия ИКТ» под авторством Гораны Целебич и Дарио Рендулича [7]. Второй источник — сборник тезисов лекций по дисциплине ИКТ, разработанный преподавателями кафедры компьютерного моделирования и информационных технологий ВКГУ имени С. Аманжолова и утвержденный методическим советом вуза. Третий источник — содержит материалы открытого и бесплатного онлайн-курса по основам компьютерных наук «CS301: ComputerArchitecture» на интернет-площадке SaylorAcademy [8]. Предобработка учебных материалов включала их предварительный парсинг (очистку от разметки) и лемматизацию, т.е. приведение слов в нормальную форму. Эти операции выполнялись с помощью инструментов открытой библиотеки NLTK [9].

Таблица

Описание входной коллекции учебных материалов

| № | Источник | Количество слов до и после предобработки | | Остаток от первоначального объема (%) |
|---|---|---|---|---|
| | | до | после | |
| 1 | «Основные понятия ИКТ» | 14664 | 6139 | 58 |
| 2 | «Информационно-коммуникационные технологии» | 15062 | 5857 | 61 |
| 3 | «Архитектура компьютера» | 13019 | 6805 | 48 |
| | Итого | 42745 | 18801 | 56 |

Затем осуществлялись загрузка «очищенных» источников в виде корпуса текстовых документов в среду R (было загружено 26 документов из трех источников) и формирование на основе корпуса дистрибутивной частотной матрицы документы-на-термины, с обязательным удалением из нее стоп-слов и знаков пунктуации. В результате последних двух операций количество слов корпуса значительно сократилось, примерно в 2 раза (см. таблицу).

```
corpus <- VCorpus(DirSource(c(«D:/MAG/Source/ICT_book/»,
+ «D:/MAG/Source/EMCD_ICT/»,
+ «D:/MAG/Source/Comp_Arc/»)))
stop.words <- table(read.csv(«D:/MAG/stopwords.csv»))
dfm <- dfm(corpus,remove_punct=TRUE, remove_hyphens = FALSE, remove = c(stopwords(source = smart), names(stop.words)))
```

Для визуализации результатов предобработки были использованы два популярных инструмента, представленные соответствующими функциями в R — инструмент WordCloud (облако слов) и инструмент WordKeyness (ключевое слово). Инструмент WordCloud показывает распределение слов в тексте в виде облака так, что наиболее частотные слова показываются большим шрифтом и это позволяет увидеть наиболее репрезентативные слова текста. WordKeyness позволяет сравнить два текста по отличительным словам. На рисунке 1 показан скрин-шот панели разработанного веб-приложения, где производится сравнительное распределение слов в двух источниках.

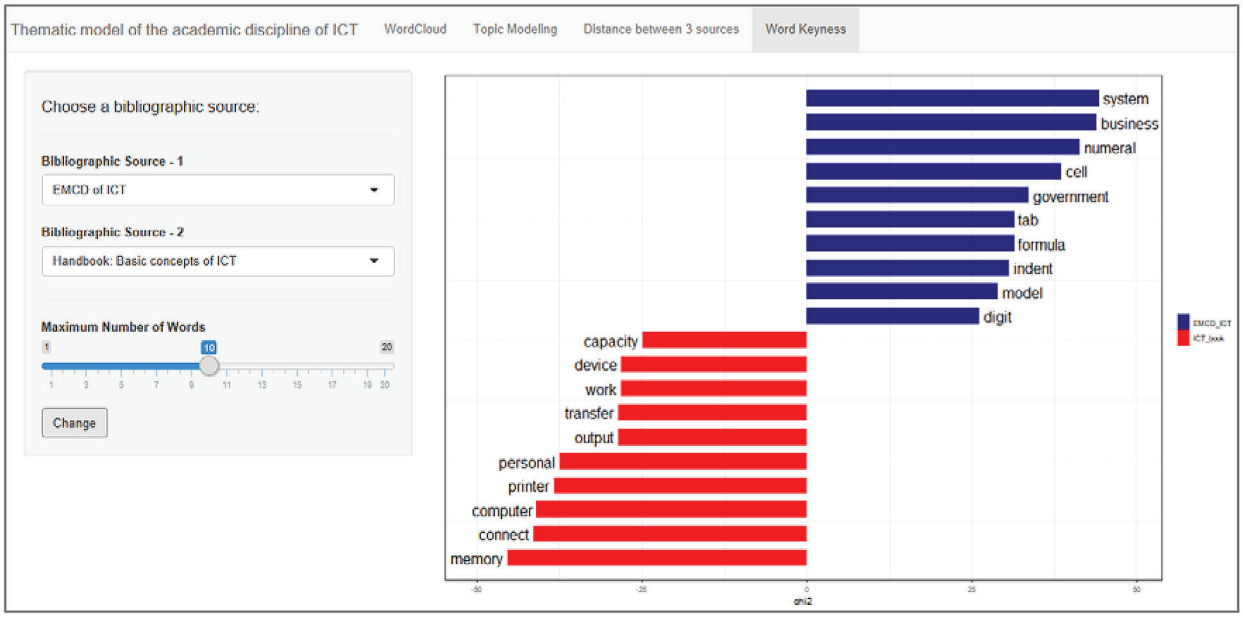


Рис. 1. Сравнение двух источников внутри корпуса

**2. Построение тематической модели и ее визуализация.** Для выполнения тематического моделирования в R предназначена функция LDA из пакета topicmodels. Функция принимает в качестве параметра матрицу документы-на-термины, число выделяемых скрытых тем *k* (определяется эмпирически) и метод оценки правдоподобия:

Res <- LDA (dfm, k, method = VEM)

Функция возвращает два основных слота, первый слот показывает распределение слов по темам, второй — распределение документов по темам.

На рисунке 2 показан фрагмент табличного представления второго слота, т.е. показаны веса каждой из семи выявленных тем в каждом из 26 документов корпуса.

Произведена сортировка по весам второй темы, и явно видно, что эта тема связана с аппаратным обеспечением. Например, документ «1.8. Legalregulations», который представляет собой восьмую главу первого источника, тесно связан с темой аппаратного обеспечения, из чего следует, что он раскрывает суть правового регулирования не в области ИКТ вообще, не в области программного обеспечения, а именно в области аппаратного обеспечения. То же самое справедливо в отношении документа «2.1. ComputerSystems». Разработанный интерактивный интерфейс для представления результатов тематического моделирования позволяет пользователю (эксперту) глубже понять тематическое представление источников и их разделов, а также выявить глубинные предметные связи между разными источниками и темами.

Более удобным средством визуализации тематической модели LDA является интерактивный инструмент LDAvis., при выборе круга справа отображается список ключевых слов, соответствующих теме (см. рисунок 3).

Как показано на рисунке 3, визуализация состоит из двух основных частей. Левая панель визуализации отвечает за отображение тем и их отношений. Каждая тема на этой визуализации представляет собой пронумерованный круг, размер круга определяется весом темы в коллекции. Более близкие темы показаны ближе друг к другу, некоторые даже пересекаются. Правая панель визуализации представлена в виде горизонтальной шкалы, которая отображает наиболее подходящие термины для объяснения выбранной темы. Диалоговая процедура позволяет пользователям изменять значение λ, параметра, который может изменить ранжирование ключевых

Document clustering | Document-Topic Matrix

Show 10 entries Search:

| Name.of.documents | Topic.1 | Topic.2 | Topic.3 | Topic.4 | Topic.5 | Topic.6 | Topic.7 |
|---|---|---|---|---|---|---|---|
| 3.1 History of computing hardware.txt | 1e-05 | 0.99997 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 |
| 1.8 Legal regulations.txt | 1e-05 | 0.99996 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 |
| 1.1 Hardware.txt | 1e-05 | 0.99994 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 |
| 3.4 Hardware and Machine Organization.txt | 6e-05 | 0.75693 | 0.24279 | 0.00006 | 0.00006 | 0.00006 | 0.00006 |
| 2.1 Computer systems.txt | 5e-05 | 0.55091 | 0.00005 | 0.29482 | 0.15405 | 0.00005 | 0.00005 |
| 3.5 Parallel and Vector Architectures.txt | 9e-05 | 0.03606 | 0.00009 | 0.00009 | 0.00009 | 0.00009 | 0.96350 |

Рис. 2. Табличное представление LDA

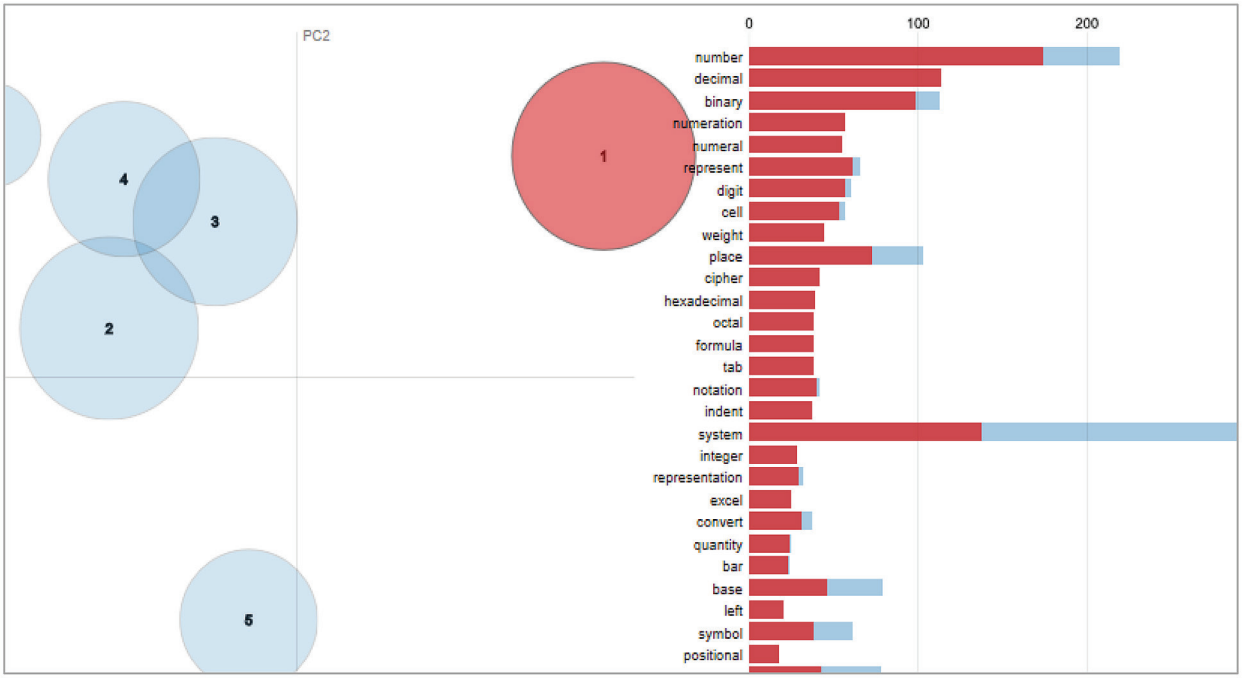


Рис. 3. Инструмент LDAvis для визуализации темы

слов темы. По умолчанию установлено значение 1. Чем ниже значение параметра, тем выше поднимаются отличительные термины темы, которые практически отсутствуют в других темах [10]. Так, например, на указанном рисунке в панели ключевых слов высокий ранг имеют слова «число», «система», «десятичная», «двоичная», «восьмеричная», «нумерация», «цифра», «шестнадцатеричная». Сочетание этих терминов позволяет определить, что речь идет о теме «Системы счисления», а доля красного и синего напротив каждого термина показывает, насколько этот термин привязан только к данной теме. Например, высокая доля синего напротив слова «система» говорит об универсальности этого термина, а высокая доля красного напротив слова «десятичный» — об узкой направленности последнего.

**3. Разработка веб-интерфейса**

Для разработки интерактивного веб-интерфейса применялся инструмент Shiny, библиотека в R. Усовершенствованные «реактивные» связи между

входными и выходные данными и предопределенный набор виджетов Shiny позволяют создавать гибкие, мощные приложения. Приложение Shiny состоит из двух компонентов: пользовательский интерфейс (ui.R); серверная часть (server.R). Первый компонент управляет компоновкой элементов управления и внешним видом приложения. Скрипт server.R содержит инструкции обработки данных. Метафора Shiny подразделяет все данные на два вида: входные данные и выходные данные. Источник данных для вычислений инициализируется при запуске приложения. Реактивные функции (в ответ на изменения пользователя) используются для повторной загрузки данных, отображаемых на экране в реальном времени. Таким образом, скрипт server.R включает в себя правила, описывающие взаимосвязь между входными и выходными параметрами. Результат выполнения скриптов экспортируется в браузер с помощью функции shiny::runApp ().

**Заключение.** Разработанная модель позволяет пользователю (учителю) визуализировать предметную область, взаимосвязь между предметными областями. Это экономит время, затрачиваемое на изучение, анализ, подбор соответствующей литературы, а также определяет наиболее подходящий контент в базовом учебном пособии. Тематическая модель учебной программы ИКТ основана на автоматическом извлечении знаний из библиографических источников с использованием статистических методов. Поэтому модель гибкая, универсальная в любой предметной области. Также очень эффективно использовать разработанную модель в процессе создания тематического плана выборных дисциплин.

## Библиографический список

1. Jockers M.L. Macroanalysis: Digital methods and literary history. — University of Illinois Press, 2013.

2. Blei D.M. Probabilistic topic models // Communications of the ACM. — 2012. — Т. 55. — №. 4.

3. Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. — 2013. — Т. 1. — № 6.

4. Коляда А.С., Яковенко В.А., Гогунский В.Д., Яковенко В.О., Гогунський В.Д. Применение латентного размещения Дирихле для анализа публикаций из наукометрических баз данных // Pratsi. — 2014. — № 1 (43).

5. David M. Blei, Andrew Y.Ng, Michael I. Jordan. Latent Dirichlet Allocation // Journal of Machine Learning Research. — Stanford, 2003. — 1/03.

6. Минаев В.А., Королев И.Д., Кисленко И.А. Методы выявления латентной и негативной информации в текстовых документах // Технологии техносферной безопасности. — 2016. — №. 5.

7. Celebic G., Rendulic D. Basic Concepts of Information and Communication Technology // Handbook [Electronic resourse]. — URL: http://www.itdesk.info/handbook_basic_ict_concepts.pdf (дата обращения: 19.05.2018).

8. Computer Architecture. Online open course [Electronic resourse]. — URL: https://learn.saylor.org/course/view.php?id=71 (дата обращения: 19.05.2018).

9. Manning C. et al. The Stanford CoreNLP natural language processing toolkit // Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. — 2014.

10. Chen F. Topic Modeling of Document Metadata for Visualizing Collaborations over Time / P. Chiu, S. Lim // Proc. of the Int. Conf. on Intelligent User Interfaces (IUI). — 2016.