

LDA в исследовании блогосферы

Границы применимости и проблемы интерпретации

К. А. Маслинский

Высшая школа экономики Санкт-Петербург
Лаборатория интернет-исследований

24 апреля 2012, Москва, ВШЭ

Outline

Понятие темы в LDA

Данные: ввод и вывод

Подготовка текста

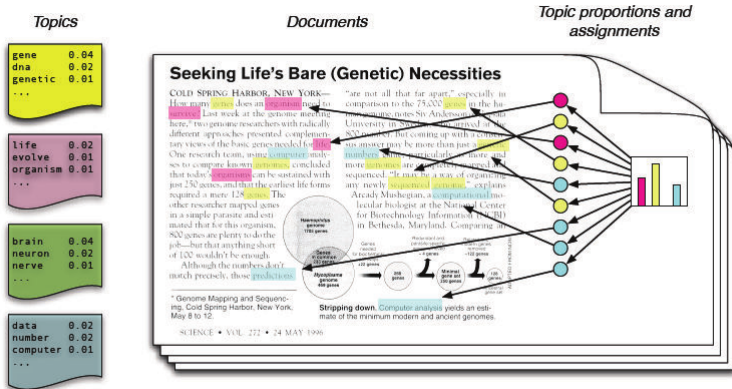
Результат работы LDA

Интерпретация тем

Типология тем

Выводы

Модель порождения документов и слов



Slide by David Blei.

Распределение тем в тексте

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Slide by David Blei.

Дистрибутивная семантика

Дистрибутивная семантика (distributional semantics) — определение семантических свойств слова по лексическому составу контекстов, в которых оно встречается.

Основные факторы, влияющие на вычленяемые в LDA темы:

- частотность терминов;
- совместная встречаемость терминов в документах.

Outline

Понятие темы в LDA

Данные: ввод и вывод

Подготовка текста

Результат работы LDA

Интерпретация тем

Типология тем

Выводы

Удаление разметки

После удаления разметки:

Оригинал взят у LJUSERkuteeva в HREF А скоро новый год
SMILEA Пора задуматься о праздничной сервировке стола. IMG
Подробности — <http://kuteeva.livejournal.com/25417.html>

- удаление всей html-разметки (шрифтовые выделения, таблицы и т.п.)
- Специальная обработка:
 - Гиперссылки → HREF
 - Изображения → IMG
 - Смайлики → SMILEA, SMILEU
 - Специальные элементы ЖЖ → LJrepost LJuser LJembed LJcut

Лемматизация

До лемматизации:

Обожаю всякое время типа 4:44. Постоянно обращаю внимание на часы, когда на них 5:55 или 12:34 или еще какая-нибудь последовательность.

После лемматизации:

обожать всякое время тип постоянно обращать внимание на час когда на они или или еще какой-нибудь последовательность

- лемматизатор mystem
- разрешение омонимии: автоматический выбор более частотной либо первой леммы

Фильтрация редких/частотных терминов

После фильтрации:

дом никто дверь сортир запирасть закрывать ванная

Исходный текст:

А вы, когда дома никого нет, дверь в сортир запираете? А закрываете? А в ванную?

Stanford Topic Modeling Toolbox

1. Удаление 100 наиболее частотных терминов
2. Удаление терминов, встречающихся менее чем в 5 документах
3. Удаление документов длиной менее 5 терминов (после чистки)

Outline

Понятие темы в LDA

Данные: ввод и вывод

Подготовка текста

Результат работы LDA

Интерпретация тем

Типология тем

Выводы

Топ-20 слов каждой темы

Topic 040 23451

сергей 518

удальцов 494

суд 390

сутки 372

судья 292

арест 232

москва 209

взять 179

задерживать 175

оригинал 174

декабрь 173

сегодня 166

навальный 158

тверской 150

час 140

митинг 130

сейчас 129

свобода 128

Topic 052 40304

животное 1116

собака 1040

кошка 703

кот 445

жить 306

птица 297

слон 291

вид 286

хозяин 246

зверь 217

крыса 197

зоопарк 186

дикий 185

котенок 183

корова 175

черный 174

обезьяна 173

самец 168

Распределение тем по документам

Какие новости? Что в стране происходит? Что делать? Кто виноват? Кто пидарас?

Topic 1 0.2 полиция 1426 полицейский 832 сотрудник 776
задерживать 696 сообщать 458 милиция 371
находиться 322 происходить 320 несколько 304
рассказывать 282 молодой 274 наркотик 247 около
246

Topic 90 0.6 понимать 1711 надо 1616 потому 1604 делать 1602
т 1571 вообще 1457 нужно 1209 тоже 1185 что-то
1176 думать 1135 почему 1088 про 1059 какой-то
1051 сейчас 983

Topic 98 0.2 эфир 623 программа 560 канал 545 журналист 461
немцов 442 передача 374 вести 369 телеканал 330
разговор 295 новость 289 интервью 273 тв 264
LJembed 251 радио 246 телевидение 222 москва

Распределение тем по документам

новость[98] происходить[1] делать[90] виноватый[90]
пидарас[90]

Topic 1 0.2 полиция 1426 полицейский 832 сотрудник 776
задерживать 696 сообщать 458 милиция 371
находиться 322 происходить 320 несколько 304
рассказывать 282 молодой 274 наркотик 247 около
246

Topic 90 0.6 понимать 1711 надо 1616 потому 1604 делать 1602
т 1571 вообще 1457 нужно 1209 тоже 1185 что-то
1176 думать 1135 почему 1088 про 1059 какой-то
1051 сейчас 983

Topic 98 0.2 эфир 623 программа 560 канал 545 журналист 461
немцов 442 передача 374 вести 369 телеканал 330
разговор 295 новость 289 интервью 273 тв 264
LJembed 251 радио 246 телевидение 222 москва

Outline

Понятие темы в LDA

Данные: ввод и вывод

Подготовка текста

Результат работы LDA

Интерпретация тем

Типология тем

Выводы

Язык

Topic 014 16249 the 772 of 592 a 536 to 520 and 475 in 461 by
267 is 226 I 220 The 206 s 193 on 176 with 138 for 133 be 130 at
123 it 105 A 104 are 101 de 99

1.0

LJraw> LJcut> 1.21 JiggaWatts! "love that movie SMILEA " by
MikeBehnken

href read more at href Michael de'Oz - Лента сообщений

href rss2lj

Язык

Торіс 017 24244 і 1261 з 933 що 668 як 389 про 249 це 230 алло
214 від 183 є 163 його 145 час 123 рок 120 вже 118 під 117 ми
116 які 112 коли 111 було 105 він 103 тако 102

Событие (новость)

Topic 050 14195 убивать 279 мирзаев 248 удар 206 девушка 190
русский 183 парень 175 иван 146 студент 112 расул 110
кавказец 109 агафонов 105 убийство 98 голова 97 убийца 95
чемпион 90 летний 87 случай 82 клуб 80 сила 68 момент 67

В столичном УВД расследуется уголовное дело против 19-летнего Ивана Агафонова, скончавшегося после удара Расула Мирзаева. Убитый чемпионом был под следствием за разбой [HREF](#) Подробнее

Известно что Майк Тайсон избил Дона Кинга. Но убить не смог. Силы удара не хватило. Валуев избил пенсионера охранника. Но не убил. Силы удара не хватило. А вот Мирзаев якобы с одного удара убил. Хотя легче Тайсона и Валуева вполтину. Может у бойцов по боям без правил удар сильнее чем у боксеров?

Событие (новость)

Topic 030 49951 ливий 1317 каддафи 1087 нато 667 ливийский
610 война 522 триполи 502 повстанец 427 военный 392 против
378 сила 362 власть 257 оон 256 убивать 253 сирия 249 режим
248 событие 223 западный 209 правительство 208
международный 205 полковник 204

0.95

Генеральный секретарь НАТО заявил: мы останемся в Ливии, пока в ней существует угроза мирному населению. Между тем логичнее было бы сказать: пока мы останемся в Ливии, в ней существует угроза мирному населению.

Событие (новость)

Topic 015 20873 сша 1009 американский 598 сентябрь 394
америка 375 штат 337 мир 294 американец 273 теракт 259
война 225 нью-йорк 210 башня 158 здание 156 террорист 148
президент 143 обама 139 мировой 136 погибнуть 135 событие
132 вашингтон 126 афганистан 119

1.0

10 лет назад, 11 сентября 2001 года в результате захвата пассажирских самолётов и тарана ими зданий Всемирного торгового центра и Пентагона в Нью-Йорке и Вашингтоне и на юго-западе Пенсильвании погибло почти три тысячи человек. Этот день... [HREF Читать далее](#)

Предметная область

Topic 001 25477 женщина 1678 мужчина 1290 девушка 714 секс
499 муж 361 друг 304 жена 268 женский 267 отношение 256
любовь 255 сексуальный 250 любить 212 свадьба 208 молодой
198 мужской 181 невеста 162 парень 159 половой 140 подруга
137 тело 135

0.888888888889

- Вы рождены в законном браке? - Да, наполовину. -
Наполовину? Как это понять? - Мой отец был женат, а мать
нет...

Предметная область

Topic 004 18875 машина 925 автомобиль 683 дорога 382
водитель 282 авто 147 w 144 с 140 скорость 136 руль 131 колесо
120 движение 112 гонка 109 км 106 ездить 104 двигатель 101
трасса 99 задний 96 гибдд 95 пробка 91 транспорт 87

0.714285714286

Можно ли в шабат ездить на машине с акпп, если зажигание
включит человек из другого сословия?

Между событием и предметной областью

Topic 002 17666 церковь 874 храм 502 православный 364
монастырь 329 **охлобыстин 261** святой 240 бог 214 священник
199 церковный 180 христос 172 иван 158 иоанн 146 вера 144
божий 136 александр 129 духовный 113 религия 111
христианин 105 русский 103 слово 101

0.555555555556

«... Как повелось от корней Ратную службу несут Всяк на
своём рубеже: Инок, воин и шут!» (К. Кинчев) Известный
киноактёр и запрещённый в служении священник Иван
Охлобыстин заранее привлёк внимание к своей программе,
озвученной 10 сентября в... [HREF Читать далее](#)

Регистр (дискурсивная тема)

Topic 006 82864 делать 895 сейчас 764 надо 742 работа 726
много 657 нужно 629 работать 598 потому 589 потом 557 дом
548 вообще 531 деньги 527 нужный 517 конечно 511 жить 498
думать 484 понимать 452 тоже 435 любить 425 ничто 408

1.0

Кажется интернет начал работать стабильно. Хорошо.

1.0

А я как в детстве - хочу сам не знаю чего... не знаю с кем, не
знаю, как, не знаю, надолго ли, но точно знаю две вещи: хочу
сейчас и очень. HREF IMG

Регистр (прагматика)

Topic 032 45268 понимать 652 про 640 надо 566 ли 484 потому
455 вопрос 421 почему 415 думать 404 писать 398 делать 372
ваш 321 ничто 319 слово 315 должный 312 написать 306
спрашивать 293 никто 290 рассказывать 281 вообще 277 сейчас
251

1.0

IMG если к этой картинке вам еще подпись нужна, значит она
точно не для вас

1.0

Интересно получается когда, например, то что ты рассказывал
другим пытаешься принять как услышанное тобой. Фу. Занудой
себе кажешься.

Регистр

Topic 000 70194 вопрос 686 ли 666 интересный 622 сейчас 615
про 581 понимать 571 думать 534 вообще 528 потому 443
почему 428 много 426 несколько 407 видеть 400 какой-то 393
конечно 379 сегодня 375 т 371 что-то 368 именно 354 тоже 352

1.0

HREF IMG Может кто то хотел задать вопрос, да не собрался
это сделать в каком то посте не по теме? Спрашивайте тут все,
что хотели у меня узнать. Отвечу. SMILEA

Регистр (книжный стиль)

Торіс 039 97278 более 683 являться 661 именно 539 случай 475
иметь 464 образ 435 лишь 409 однако 408 отношение 370 слово
361 происходит 358 существовать 353 собственный 334
некоторый 316 мир 312 общество 310 либо 309 часть 306
известный 300 причина 298

0.833333333333

Искажение в связи с проекцией — тенденция бессознательно
полагать, что другие люди разделяют те же, что и субъект,
мысли, верования, ценности и позиции (с) Вот чем я часто
грешу. Ожидаю от людей одного, а получаю совсем-совсем
другое...

Регистр (нецензурная лексика)

Topic 046 19356 хуй 234 гавно 187 мужик 184 жопа 162 блядь
153 блятить 148 сука 143 тупой 136 пиздец 134 надо 116 хуйня
112 тип 111 получать 106 нахуй 96 пацан 82 ребенок 77 брать
76 бля 75 че **хипстер 71**

1.0

- Шеф, сука, заставил сисадмина сделать доступ к
"Вконтакте"и Facebook с 8 до 9 утра! ! Якобы чтобы народ
приходил вовремя... - И что? - И ведь приходят

Между регистром и предметной областью

Торіс 023 28958 имя 333 великий 328 русский 323 война 248
история 218 александр 198 память 192 герой 181 родина 179
народ 150 слава 138 царь 130 век 126 дед 123 помнить 118
старый 109 фамилия 106 князь 103 писать 102 сей 100

0.8

HREF Башкирская сказка "Сапожник и царь"(аудио)

Сегодня 70 лет со дня начала блокады Ленинграда. Скорбная дата, которая явилась настоящей трагедией. Но героизм Защитников города является одной из памятных страниц нашей истории и моего города в частности. Я просто не мог не отметить. IMG

Между регистром и предметной областью

Topic 025 20910 LJembed 1300 игра 538 видео 524 ролик 403
смотреть 323 LJlike 322 LJrepost 287 SMILEA 249 интересный
191 посмотреть 190 про 177 показывать 165 SMILEAA 139 ru
131 канал 126 эфир 119 здесь 115 играть 107 снимать 106
прямой 98

1.0

Здесь вот можно HREF Jimm скачать для мобильного, так что
заходим, смотрим и скачиваем.

Жанр (текстовая структура)

Topic 040 15224 москва 466 сергей 332 александр 296 олег 188
узбекистан 160 михаил 155 владимир 154 иван 152 елена 146
журналист 137 область 134 андрей 125 алексей 118 николай
115 иванович 110 Санкт-Петербург 100 татьяна 98 писатель 97
дмитрий 95 директор 92

По мере поступления подписей -

LJcut 1. Юрий Михайлович Васильев, журналист, Москва 2.
Максим Сергеевич Александров, писатель, Санкт-Петербург 3.
Тимур Алиев, госслужащий, Грозный 4. Святослава Швец,
журналист, Рим 5. Алексей Цветков, литератор, Нью-Йорк

Между жанром и предметной областью

Торіс 056 43257 масло 540 добавлять 464 минута 434 рецепт
 385 вкусный 380 вода 336 перец 298 соль 296 сахар 287 блюдо
 283 мясо 281 вкус 273 мука 244 получаться 232 тесто 230 л 228
 молоко 220 лук 218 SMILEA 196 яйцо 193

1.0

..это я сегодня SMILEA

казан-кабоб с говядиной

IMG

LJcut 500 гр говяжьих ребрышек, порубленных на некрупные
 куски 1 кг нерассыпчатого картофеля 1 луковица растительное
 масло смесь молотых зиры, кориандра, острого красного перца
 и сушеного базилика 1.5 ст л 1 ст л винного уксуса

Растиражированный текст

Торис 016 16654 земля 256 планета 245 ученый 155 звезда 144
космический 140 робот 136 приводит 135 ядерный 124 свет
118 энергия 117 число 108 луна 104 человечество 84 солнечный
83 путь 79 создавать 76 возможно 70 поверхность 69 результат
68 взрыв 67

0.992167101828

В будущем с человечеством может произойти глобальная катастрофа, которая приведёт к полному человеческому вымиранию. Мною составлен максимально полный список возможных причин такой катастрофы. Наверное, многие угрозы в него не вошли, а многие в будущем будут опровергнуты, однако сам факт большого числа рисков говорит о том, что хотя бы один может сработать, и заставляет задуматься о том, что же мы можем сделать, чтобы этого избежать.

Артефакты и шум

Topic 008 22204 com 2753 html 2060 livejournal 1823 ru 1733
www 1520 news 310 v 284 ссылка 215 youtube 211 php 196 watch
187 org 186 посмотреть 145 labazov 144 здесь 140 LJembed 139
id 137 htm 110 пост 101 neoguru 95

1.0

IMG Это превьюшка. Оригиналы в Яндекс-фотки. Кто хочет
посмотреть, вам сюда: HREF

[http://fotki.yandex.ru/users/bulochnikov-
boris/view/456126/?page=8](http://fotki.yandex.ru/users/bulochnikov-boris/view/456126/?page=8) Там есть опция подбора размеров
изображения.

Тематический состав текста

- LDA topic $\neq$ тематика текста в обыденном понимании
- Тема в LDA отражает статистические особенности распределения слов по текстам, характерные для **данной коллекции**.

Ядром темы могут быть:

- предметная область;
- регистр (прагматика текста, стиль) и жанр;
- списочные структуры, формульный текст;
- язык;
- копии одного и того же текста.

Лексический состав темы

- Термины из топа темы не обязательно характерны для самых вероятных текстах темы.
- Термины из топа темы позволяют вычленить наиболее различающиеся по сочетаемым свойствам группы лексики в коллекции.
- Присутствие слова среди лексики, принадлежащей отчетливой семантической группе, говорит о высокой частотности данного слова в текстах с данной лексикой.