Тематический анализ записей разговоров контакт-центра

Воронцов Константин Вячеславович

МФТИ • ФИЦ ИУ РАН • ШАД Яндекс • Айтея

24–30 июля 2017 • DeepHack.Turing • Москва, МФТИ

http://turing.tilda.ws

Содержание

Постановка задачи

- **Дано:**
  1) коллекция текстов разговоров
  2) семантические ядра тем (предмет разговоров известен)
  3) сегментная разметка небольшой выборки разговоров
- **Найти:**
  1) тематическая сегментация каждого разговора
  2) граф сценариев разговоров
  3) вероятность успешного исхода в любой точке разговора
  4) оценки качества работы операторов
  5) генератор онлайновых подсказок операторам
  6) рекомендации для операторов и поправки к инструкциям
- **Критерии:**
  1) точность выделения тем в разговорах
  2) точность сегментации на размеченной подвыборке

Что такое «темы» в записях разговоров контакт-центра банка

Типы тем в диалоге оператора и клиента:

- Представление
- Продукт
- Свойство продукта
- Возражение клиента
- Аргумент оператора
- Оформление заявки
- Прощание

Тинькофф
Банк

Бизнес-задачи:

- Повышение доли успешных разговоров
- Автоматизация мониторинга работы операторов

Пример темы: «Перевод баланса»

имеет смысл перевести потому что три месяца вы без процентов гасите это уже как согласитесь выигрыш но далее уже процент

с нашей помощью мы вы могли бы погасить ваши кредиты в других банках и беспроцентный период в этом случае увеличится на срок до девяноста дней

воспользоваться услугой перевод баланса услуга позволит вам погасить долг в другом банке оплачивая его вы не будить платить проценты в течении месяцев

предоставляет возможность воспользоваться услугой перевод баланса

беспроцентный период в этом случае увеличится на срок до девяноста дней

услуга бесплатная с помощью этой услуги вы можете частично или полностью закрыть действующие кредиты в других банках закрыть

Пример темы: «Льготный период»

льготный период до пятидесяти пяти дней позволяет людям отдохнуть за границей при этом вернуться и пользуясь льготным периодом восполнить средства по карте то есть не потерять на процентах

льготный период до пятидесяти пяти дней у вас практически два месяца на беспроцентное погашение

беспроцентный льготный период до пятидесяти пяти дней вы совершаете покупки в обычном режиме в магазине на заправке в аптеке и при этом не платить проценты

льготный период когда вы можете пользоваться деньгами по карте и не платить за это проценты

льготный период беспроцентный когда вы можете погашать задолженность абсолютно при этом ничего не переплачивая

Пример темы: «У банка нет отделений»

действительно в нашем банке нет отделений это дистанционный банк

потому что любой вопрос сможете решить позвонив в удобное для Вас время так как банк работает круглосуточно двадцать четыре часа в сутки семь дней в неделю

понимаю что это может вызвать недоверие однако в Тинькофф банке уже более пяти миллионов клиентов которые в основном выбрали наш банк из-за удобства

это частый вопрос от новых клиентов нашего банка многие когда слышат что нет отделений немного пугаются

а зачем нужны отделения сейчас все стараются экономить время а в отделениях банка бесконечные очереди которые забирают все силы нервы и настроение именно в нашем банке можно производить все необходимые действия дистанционно что намного удобнее

Пример темы: «Оформление заявки»

для этого нужно сначала оформить заявку

мы в телефонном режиме оформляем заявку буквально десять минут вашего времени

если заявку на сайте сделать равно вам необходима помощь операторов в том моменте что там бывают такие нюансы что ну правильно заполнение то есть без оператора

предлагаю только составить заявку

в заявке указываете в кредит который вы желаете да по своим возможностям по своим по своим потребностям

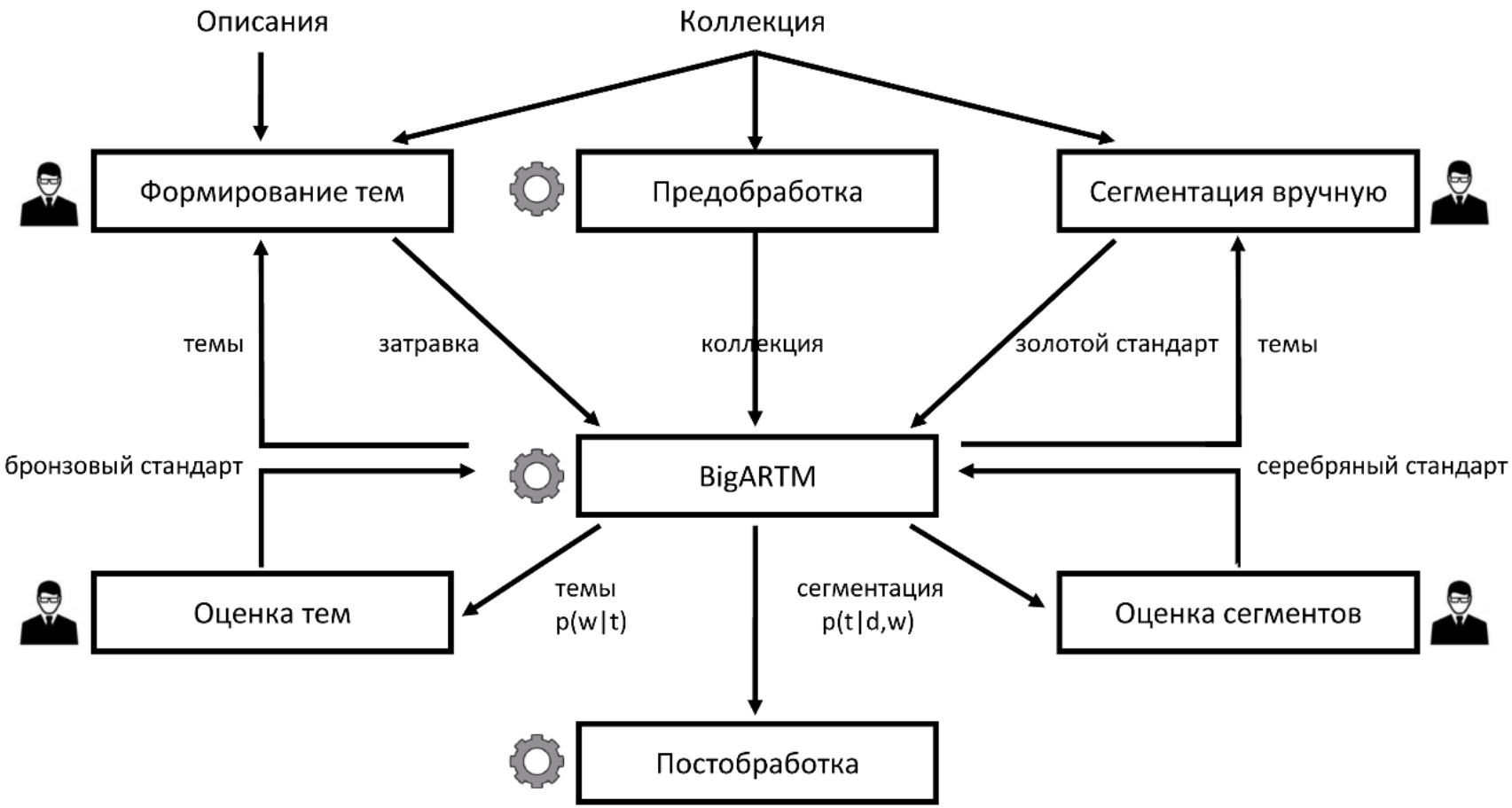
готовы сейчас оформить заявку

оформить заявку на получение кредитной карты тинькофф платинум

для получения нашей кредитной карты предлагаю сейчас заполнить заявку для этого потребуется пять семь времени

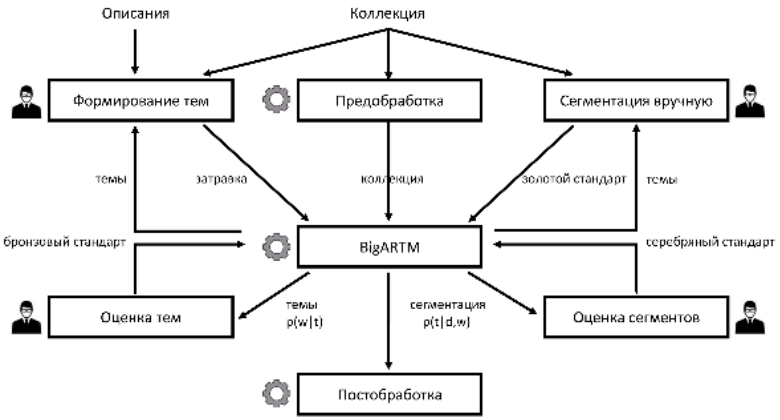
вы можете оставить заявку

Процессы обработки данных, моделирования и оценивания



Преимущества модульной архитектуры

- контроль качества модели на каждом шаге
  - качество формирования тем
  - качество сегментации разговоров
- возможность отключения модулей и упрощения продукта
  - баланс: качество — скорость внедрения
  - баланс: качество — объём ручной разметки
- возможность кастомизации под любой тип разговоров



Пример. Предварительная обработка текста

Исходная реплика:

смотрите мы предлагаем с первоначальным кредитным лимитом до трёхсот тысяч рублей можно указать желаемую беспроцентный период на покупки по карте до пятидесяти пяти дней это значит что когда вы расплачиваетесь картой потраченную сумму возвращаете беспроцентный период что вы ничего не переплачиваете ни каких процентов также по карте есть бонусная программа работает тогда расплачиваетесь картой получаете за то бонусные баллы можете экономить на другие и при этом можно не личные средства хранить также расплачиваясь получать баллы от покупок оформление для вас совершенно бесплатно и дома первой расходной операции ничего абсолютно никаких средств вас не взимается то есть вы можете получить у вас ознакомительное предложение оно что вы активировали

Пример. Предварительная обработка текста

После лемматизации:

смотреть мы предлагать с первоначальный кредитный лимит до number рубль можно указывать желать беспроцентный период на покупка по карта до number date_time это значить что когда вы расплачиваться карта потратить сумма возвращать беспроцентный период что вы ничто не переплачивать ни какой процент также по карта быть бонусный программа работать тогда расплачиваться карта получать за то бонусный балл мочь экономить на другой и при это можно не личный средство хранить также расплачиваться получать балл от покупка оформление для вы совершенно бесплатно и дома number расходный операция ничто абсолютно никакой средство вы не взиматься то быть вы мочь получать у вы ознакомительный предложение оно что вы активировать

Пример. Предварительная обработка текста

После выделения коллокаций и именованных сущностей:

смотреть мы предлагать с первоначальный кредитный_лимит до number_тысяча_рубль можно указывать_желать беспроцентный_период на покупка по карта до number date_time это_значить что когда вы расплачиваться_карта потратить_сумма возвращать беспроцентный_период что вы ничто не переплачивать ни какой процент также по карта быть бонусный_программа работать тогда расплачиваться карта_получать за то бонусный_балл мочь _экономить на другой и при это можно не личный_средство хранить также расплачиваться получать_балл от покупка оформление для вы совершенно_бесплатно и дома number расходный_операция ничто абсолютно никакой средство вы не взиматься то быть вы мочь получать у вы ознакомительный_предложение оно что вы активировать

Этапы автоматической обработки данных

- **Предварительная обработка текста**
  - расстановка точек и запятых (CRF или LSTM)
  - лемматизация (pymorphy)
  - выделение коллокаций и именованных сущностей
  - построение синтаксических деревьев (SyntaxNet)
- **Тематическое моделирование** (BigARTM)
  - модель дистрибутивной семантики (WNTM)
  - частичное обучение тем по семантическим ядрам
  - выделение слов общей лексики в фоновые темы
  - тематическая сегментация с учётом синтаксиса
  - модальности коллокаций и именованных сущностей
- **Постобработка**
  - построение графа сценариев (Sankey diagram)
  - оценивание эффективности ветвей сценария
  - выявление новых тем для дополнения скриптов и тренингов

Необязательные этапы ручной разметки данных

В зависимости от типа разговоров и бизнес-задач можно задействовать лишь некоторые из четырёх этапов:

- **Формирование тем** — «затравка»
  - *Вход:* разговоры, скрипты, грубая тематизация
  - *Выход:* отобранные темы, их семантические ядра
- **Сегментация вручную** — «золотой стандарт»
  - *Вход:* небольшая выборка разговоров, темы
  - *Выход:* границы тематических сегментов
- **Оценивание сегментов** — «серебряный стандарт»
  - *Вход:* сегменты разговоров
  - *Выход:* для каждого сегмента: тема / не однороден
- **Оценивание тем** — «бронзовый стандарт»
  - *Вход:* частотные словари тем
  - *Выход:* белые и чёрные списки терминов в темах

Задача тематического моделирования

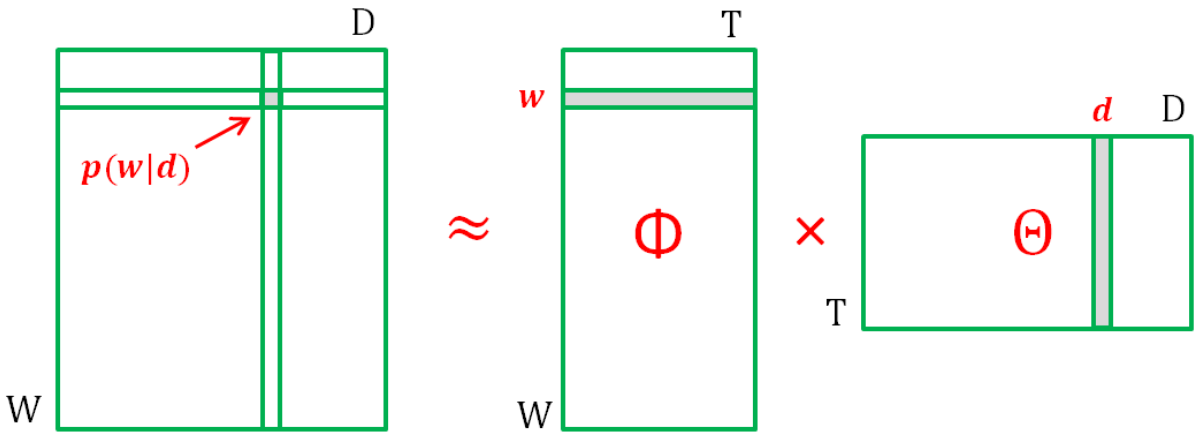
Дано: коллекция текстовых документов

- n_{dw} — частоты слов в документах, $p(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

- $\phi_{wt} = p(w|t)$ — вероятности слов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар (1865–1963)

Задача стохастического матричного разложения является *некорректно поставленной* — её решение не единственно:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

для невырожденных $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Регуляризация — дополнительные ограничения на Φ, Θ.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.

Предположения, формализуемые с помощью регуляризаторов

- Фиксация тем с помощью частичного обучения
  - задание обучающих документов из ключевых фраз темы
  - задание белых и чёрных списков слов по теме
- Разделение тем на предметные и фоновые
  - выведение слов общей лексики из всех тем в фоновую тему
  - декоррелирование тем для повышения их различности
- Использование коллокаций и именованных сущностей
  - введение модальностей и подбор их весов
- Построение первичных тем без обучающих данных
  - учёт локальной совстречаемости слов
  - использование моделей дистрибутивной семантики
- Тематическая сегментация
  - тематика слов, стоящих рядом, скорее всего, близка
  - выделение границ сегментов в местах резкой смены тем

Регуляризаторы на основе кросс-энтропии

1. разреживание предметных тем $S \subset T$:

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in S}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in S} \alpha_t \ln \theta_{td} \to \max$$

2. сглаживание фоновых тем $B \subset T$:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \to \max$$

3. частичное обучение по подмножествам $W_t \subset W$, $D_t \subset D$:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W_t} \ln \phi_{wt} + \alpha_0 \sum_{t\in T}\sum_{d\in D_t} \ln \theta_{td} \to \max$$

4. удаление неинформативных тем:

$$R(\Theta) = -\tau \sum_{t\in S} \ln p(t) \to \max, \qquad p(t) = \sum_{d\in D} \theta_{td} p(d)$$

Регуляризатор декоррелирования тем

Цель — выделить *лексическое ядро* каждой темы — множество характерных слов, отличающих её от других тем.

Минимизируем ковариации между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2}\sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W}\phi_{wt}\phi_{ws} \to \max.$$

Подставляем в формулу M-шага:

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\Big(n_{wt} - \tau\phi_{wt}\sum_{s\in T\setminus t}\phi_{ws}\Big).$$

Эффект разреживания (контрастирования) строк матрицы Φ: наименьшие ϕ_{wt} в строках могут обращаться в нуль.
Повышается различность тем как столбцов матрицы Φ.

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. 2010.

Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

Максимизация суммы log правдоподобий с регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(\sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right). \end{cases}$$

Модель совстречаемости слов в коротких текстах

Битерм — пара слов, встречающихся рядом:
в одной реплике / предложении / окне $\pm h$ слов.

Тематическая модель битермов (Biterm topic model):

$$p(u,w) = \sum_{t\in T} p(w|t)p(u|t)p(t) = \sum_{t\in T} \phi_{wt}\phi_{ut}\pi_t,$$

где $\phi_{wt} = p(w|t)$, $\pi_t = p(t)$ — параметры модели.

Критерий максимума логарифма правдоподобия:

$$\sum_{u,w} n_{uw} \ln \sum_t \phi_{wt}\phi_{ut}\pi_t \;\to\; \max_{\Phi,\pi},$$

$$\phi_{wt} \geqslant 0; \quad \textstyle\sum_w \phi_{wt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts // WWW 2013.

Регуляризатор битермов для обычной ΦΘ-модели

1. Регуляризатор битермов для матрицы Φ:

$$R(\Phi) = \tau \sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{wt} \to \max$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\operatorname{norm}} \Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \Big);$$

$$p_{tuw} \equiv p(t|u,w) = \underset{t \in T}{\operatorname{norm}} \big(n_t \phi_{wt} \phi_{ut} \big).$$

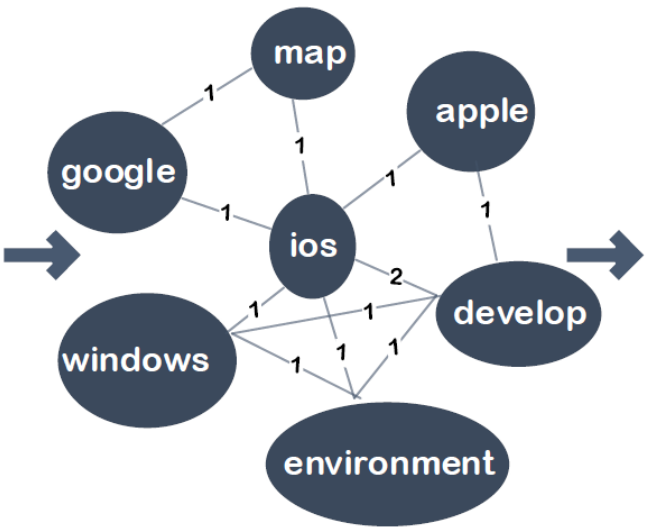
2. Регуляризатор разреживания для матрицы Θ:

$$R(\Theta) = -\tau' \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Модель сети слов WNTM для коротких текстов

Идея: моделировать не документы, а связи между словами.
d_w — псевдо-документ, объединение всех контекстов слова w.
n_{wu} — число вхождений слова u в псевдо-документ d_w.
Контекст — реплика в диалоге / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

$$p(u|d_w) = \sum_{t \in T} p(u|t)p(t|d_w) = \sum_{t \in T} \phi_{ut}\theta_{tw},$$

где d_w — псевдо-документ слова w.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{wu} \log \sum_{t \in T} \phi_{ut}\theta_{tw} \to \max_{\Phi,\Theta},$$

где n_{wu} — совстречаемость слов w, u.

Отличие от модели битермов: там $\Theta = \mathrm{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathrm{T}}$.

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription // ACM Trans., 2009.

Примеры векторных операций в задаче аналогии слов

Два подхода к синтезу векторных представлений слов:

- **word2vec**: интерпретируемые векторные операции
- **ARTM**: интерпретируемые разреженные компоненты теперь также интерпретируемые векторные операции!

Операция	Результат ARTM	Результат word2vec
king – boy + girl	*queen*, princess, lord, prince	*queen*, princess, regnant, kings
moscow – russia + spain	*madrid*, barcelona, aires, buenos	*madrid*, barcelona, valladolid, malaga
india – russia + ruble	*rupee*, birbhum, pradesh, madhaya	*rupee*, rupiah, devalued, debased
cars – car + computer	*computers*, software, servers, implementations	*computers*, software, hardware, microcomputers

Артём Попов. Регуляризация тематических моделей для векторных представлений слов. 2017. ВМК МГУ.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

BigARTM упрощает разработку тематических моделей

Для построения композитный моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

Этапы моделирования	**Bayesian TM**	**ARTM**	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- нестандартизуемые этапы, уникальная разработка для каждой задачи

-- стандартизуемые этапы

Метод тематической сегментации TopicTiling

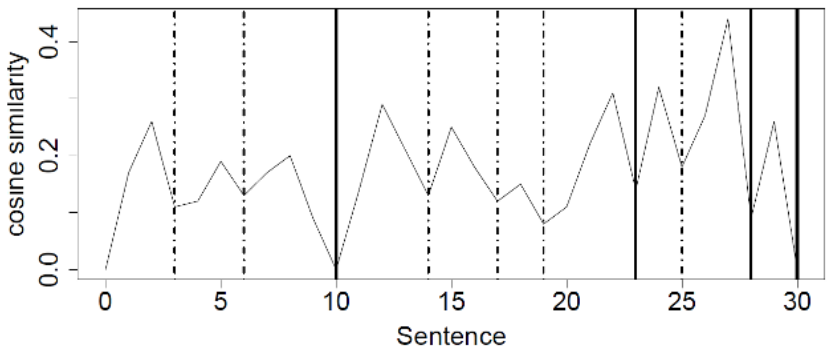
$(s_j)_{j=1}^{k_d}$ — последовательность предложений документа d

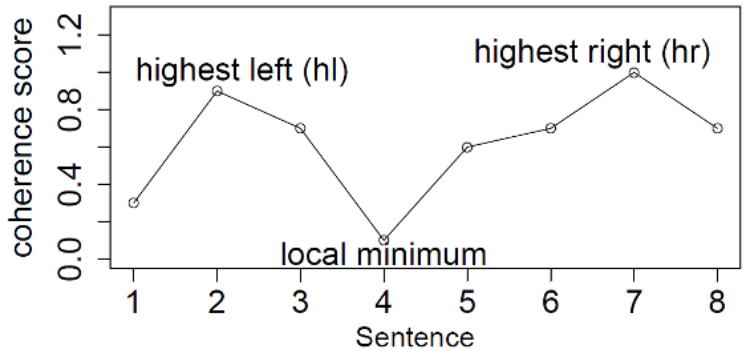
$p(t|d,s) = \frac{1}{|s|}\sum_{w\in s} p(t|d,w)$ — тематика предложения s

$p_j = \big(p(t|d,s_j)\big)_{t\in T}$ — тематический вектор предложения s_j

$c_j = \cos(p_{j-1}, p_j)$ — *coherence score*, оценка близости соседних предложений (чем глубже провал, тем чётче граница)

$d_j = \frac{1}{2}(\mathrm{hl}_j + \mathrm{hr}_j) - c_j$ — *depth score*, оценка глубины провала





Martin Riedl, Chris Biemann. Text Segmentation with Topic Models. 2012.

Второй подход: разреживание тематики предложений

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Матрица *тематики слов в документах* $p(t|d, w_i)$ размера $T \times n_d$:

Регуляризация E-шага

Трёхмерная матрица $\Pi = (p_{tdw})_{T \times D \times W}$, $p_{tdw} = p(t|d,w)$

Максимизация log правдоподобия с регуляризаторами R и $\tilde{R}$:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Pi(\Phi, \Theta)) + \tilde{R}(\Phi, \Theta) \to \max_{\Phi, \Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ & \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw} \frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w \in W}{\text{norm}}\left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \phi_{wt} \frac{\partial \tilde{R}}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial \tilde{R}}{\partial \theta_{td}}\right) \end{cases}$$

Тематическая модель мелко сегментированного текста

S_d — множество микро-сегментов документа d

n_{sw} — число вхождений слова w в сегмент s длины n_s

Тематика сегмента $s \in S_d$ — средняя тематика его слов:

$$p_{tds} \equiv p(t|d,s) = \frac{1}{n_s} \sum_{w \in s} n_{sw} p_{tdw}.$$

Кросс-энтропийный регуляризатор разреживания $p(t|d,s)$:

$$R(\Pi) = -\sum_{d \in D} \sum_{s \in S_d} \sum_{t \in T} \ln \sum_{w \in s} n_{sw} p_{tdw} \to \max.$$

Формула регуляризованного E-шага:

$$\tilde{p}_{tdw} = p_{tdw} \left(1 - \frac{\tau}{n_{dw}} \sum_{s \in S_d} \frac{n_{sw}}{n_s} \left(\frac{1}{p_{tds}} - \sum_{z \in T} \frac{p_{zdw}}{p_{zds}} \right) \right).$$

Интерпретация: если $p_{tds} < \frac{1}{|T|}$, то p_{tdw} уменьшатся $\forall\, w \in s$.

Тематика сегмента концентрируется в небольшом числе тем.

Инструмент для выполнения асессорской разметки

Assesment task #32

dialog_322.txt.

==========

O: здравствуйте меня зовут сергей банк тинькофф слушаю вас
A: а вы звонили
O: а скажите пожалуйста я разговариваю с алексеем владимировичем
A: да да да да
O: алексей владимирович удобно вам на данный момент разговаривать
A: да да говорите
O: да действительно сотрудник нашего банка пытался до вас дозвониться но к диалога у вас не получилось то есть так как у вас диалога не получилось так же хочу проинформировать вас у у нас имеется от нашего банка предложение для вас вы можете с получить нашу кредитную карту на особо выгодных весьма очень интересных условиях с высоким первоначальным кредитным лимитом беспроцентным льготным периодом на все покупки и безналичные операции сроком до пятидесяти пяти дней и совершая покупки расплачиваясь нашей картой вы всегда можете зарабатывать дополнительно и бонусы по нашей интересной бонусной программе могло ли вас заинтересовать такое предложение
A: нет мне это как бы это пока не интересно
O: а позвольте уточнить причину вашего такого решения быть может у вас имеются кредиты или кредитные карты других банков которые вы
A: нет просто здесь с кредитом связываться неохота
O: а я с вами согласен кредиты так скажем они нас обязывают выплачивать какие либо суммы постоянно ежемесячно но кредитная карта это не совсем кредит то есть данную карту можно использовать во первых как весьма очень удобное средство то есть до пятидесяти дней если вы совершали покупки расплачиваетесь картой до пятидесяти пяти дней по каждой покупке возвращаете свои задолженности в полном вы вовсе не переплачиваете никаких дополнительных денег и не выходя из дома можно оплачивать многие услуги через наш удобный интернет банк
A: нет если у меня есть карта сбербанка я также оплачиваю да и без проблем
O: то есть вы картой другого банка а вы чаще снимаете наличные деньги или расплачиваетесь ею
A: ну по разному когда как когда наличными когда когда
O: а когда какие либо бонусные программы у вас имеются по вашей карте

престиж карты
доставка карты
доставка карты
индивидуальный подход
перевод баланса
процентная ставка
связь с банком
дистанционность
общий вариант отка
данные абонента
откуда номер
точно сотрудник
желаемая сумма
снятие/партнеры
решение банка
большой процент

Topic name | Add topic

No topic

Приветствие[x] Цель звонка[x] Удобно разговаривать[x] Кредиты в других банках[x] кредитный лимит[x] беспроцентный период[x] решение о карте[x] общие вопросы клиенту[x] оформление заявки[x] наличие паспорта[x] сравнить условия[x] перезвонить[x] до свидания[x] бонусная программа[x] Использование карты[x] карту не кредит[x] с кем разговариваю[x] звонили ранее[x] есть карта другого банка[x] интернет банк[x]

Сравнение асессорской и модельной разметки (пример)

- цветом выделяются темы
- подчёркиванием выделяется асессорская разметка

Оформление заявки | Индивидуальный подход | Решение банка | Доставка
Бонусная программа | Бесплатная доставка/оформление

вот на данный момент я звоню предлагаю только составить заявку чтобы банк изучил вашу кредитную историю и подобрал под вас индивидуальный тарифный план после чего на ваш мобильный поступит уведомление в котором будет указано каким образом в случае положительного ответа будут доставлены бумаги у нас есть два способа доставки это либо курьерская доставка либо заказным письмом почтой России

ну вот бонусы значит на все абсолютно покупки один процент а если вы совершаете покупки у банка будет полный перечень магазинов у вас в личном кабинете до тридцати процентов бонусов можете то есть вот две тысячи что то купили а ориентировочно шестьсот вернулось вам на это уже плюсов согласитесь что это довольно таки это одна покупка вот так и хочу сказать что вы абсолютно ничего не теряетесь соглашаясь оформить заявку ничего за что не платите потому что вам карту выпускают доставляют абсолютно бесплатно вам либо представитель банка привозит либо по почте она приходит

Оценивание качества сегментации

Доля ошибок сегментации — доля слов, на которых тема асессора не совпадает с темой модели.

Сравниваются три модели:

1. 46.5% — частичное обучение без сегментации
2. 27.9% — сегментация без настройки параметров
3. 27.5% — сегментация с настройкой параметров

Управление качеством модели:

- оценивание качества сегментации для отдельных тем;
- либо увеличение объёма обучения по «грязной» теме;
- либо улучшение всей модели в целом

Детализация точности сегментации по темам

Назване темы	число сегментов	вклад темы в суммарную ошибку			доля ошибок сегментации		
		Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
бонусная программа	21281	2.75	0.59	0.57	28.7	6.2	5.9
беспроцентный период	18527	3.60	1.92	1.99	43.1	23.1	23.9
данные абонента	18305	3.57	1.56	1.47	43.3	19.0	17.8
использование карты	17218	6.14	4.86	4.65	79.2	62.7	60.1
перевод кредита	13905	3.00	0.96	0.84	48.0	15.3	13.4
цель звонка	13692	3.27	2.51	2.43	53.2	40.7	39.4
оформление заявки	11255	3.22	2.19	2.13	63.7	43.3	42.0
процентная ставка	10963	2.69	1.36	1.26	54.6	27.5	25.5
решение о заявке абонентом	9749	2.84	1.44	1.34	64.7	32.7	30.5
одобрение заявки	6970	1.52	0.87	0.88	48.5	27.9	28.0
кредитный лимит	6334	0.93	1.02	1.10	32.8	35.7	38.6
доставка	5983	1.14	0.61	0.61	42.3	22.6	22.5
индивидуальный подход	5498	0.91	0.48	0.50	36.7	19.2	20.2
общие вопросы клиенту	4751	1.54	1.40	1.39	71.8	65.5	64.8
бесплатное оформление	4197	0.59	0.39	0.41	31.3	20.5	21.5
представление оператора	4063	0.27	0.48	0.53	15.0	26.0	29.1
желаемая сумма	3935	0.66	0.36	0.37	37.4	20.4	20.8
дистанционность	3742	0.92	0.42	0.41	54.8	25.2	24.2
пополнение снятие партнеры	3322	0.55	0.28	0.27	37.1	18.8	18.1
связь с банком	3090	0.57	0.35	0.36	41.2	25.1	26.0
удобно разговаривать?	2618	0.23	0.40	0.44	19.6	34.0	37.2
престиж карты	2473	0.24	0.23	0.26	22.0	20.9	23.1
нужно подумать...	2433	0.66	0.32	0.31	60.6	29.0	28.0
перезвонить в другое время	2340	0.30	0.19	0.21	28.5	18.1	20.1
уже есть кредит в другом банке	2238	0.58	0.36	0.34	57.3	35.6	33.7
пользуетесь банк. продуктами?	2157	0.56	0.38	0.36	57.4	39.5	37.4

Резюме

Возможности ARTM намного шире, чем у LDA [Blei, 2003].

Возможности ARTM, используемые для анализа разговоров:

- постепенный переход от обучения без учителя к частичному обучению с контролем интерпретируемости
- модели дистрибутивной семантики (аналог word2vec): разреженные интерпретируемые векторные представления слов, предложений, фрагментов, документов
- модели сегментации в обход гипотезы «мешка слов»

Полезные ссылки:

Воронцов К. В. Обзор вероятностных тематических моделей. 2017.
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Библиотека тематического моделирования BigARTM — http://bigartm.org

Литература

- *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.Vorontsov, A.Potapenko.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. AIST 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Yanina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. 2015.
- *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
- *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov.* Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI 2016.
- *А.О.Янина, К.В.Воронцов.* Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA 2016.