Вероятностные тематические модели
Лекция 11.
Суммаризация и именование тем

Константин Вячеславович Воронцов
k.v.vorontsov@phystech.edu

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

МФТИ.ФПМИ.ИС.ИАД — ФИЦ ИУ РАН • 2025-11-27

Содержание

Основные открытые проблемы тематического моделирования

1. Гарантированная интерпретируемость тем «из коробки»
   - тематические модели внимания последовательного текста
   - моделирование коллекций с дисбалансом темами
   - измерение и оптимизация внутритекстовой когерентности
2. Каждая тема должна уметь рассказать о себе
   - автоматическое именование и суммаризация тем
   - гибридная экстрактивно-абстрактивная суммаризация
   - ранжирование фраз темы для сценария суммаризации
   - пересказ темы с помощью LLM
3. Проект «Тематизатор»: User-Friendly Topic Modeling
   - перестроение модели по экспертной разметке тем
   - авто-оптимизация числа тем и других гиперпараметров
   - обнаружение новых тем и трендов в потоке текстов
4. Проект «Мастерская знаний»:
   - визуализация тематики подборки, «карты знаний»
   - выявление трендов, вех и междисциплинарных связей

Задача суммаризации (реферирования, аннотирования) текста

Автоматическая суммаризация — *краткий* текст, построенный по одному или нескольким документам и *наиболее полно* передающий их содержание.

Основные типы задач суммаризации:

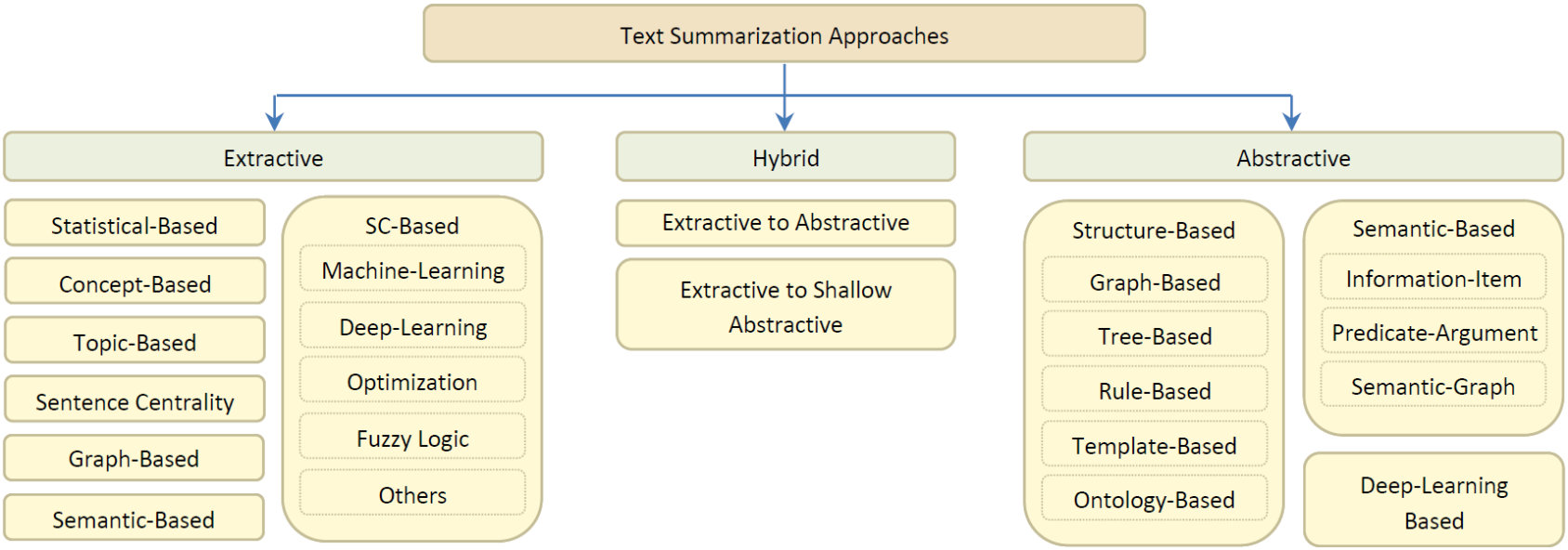
- *one-document* — на входе один документ $d \in D$
- *multi-document* — на входе набор документов $D' \subseteq D$
- ⊕ *topic* — на входе набор сегментов темы $p(d, s|t)$

Суммаризация с участием человека:

- Query-Based Text Summarization — человек задаёт запрос
- MAHS, machine aided human summarization
- HAMS, human aided machine summarization

H.P.Luhn. The automatic creation of literature abstracts. 1958
Juan-Manuel Torres-Moreno. Automatic Text Summarization. 2014

Основные подходы и методы суммаризации



Основные подходы к суммаризации:

- *extractive* — выбор некоторых предложений целиком
- *abstractive* — генерация текста на естественном языке

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, Hoda K. Mohamed. Automatic text summarization: A comprehensive survey. 2021

Основные этапы выборочной (extractive) суммаризации

1. Внутреннее представление текста
   — граф/кластеризация/тематизация предложений в тексте
   — вычисление важности и других признаков предложений
2. Оценивание полезности предложений для ранжирования
3. Возможно, поиск готовых суммаризаций или фрагментов
4. Отбор предложений для реферата
   — оптимизация критериев релевантности + различности
   — оптимизация критериев ранжирования предложений
   — учёт целей и особенностей прикладной задачи
   (статьи/патенты/новости/веб-страницы/посты/мэйлы)
5. Преобразование в связный текст (теперь с помощью LLM)

D.Das, A.Martins. A survey on automatic text summarization. 2007
A.Nenkova, K.McKeown. A survey of text summarization techniques. 2012
Y.Desai, P.Rokade. Multi document summarization: approaches and future scope. 2015
M.Gambhir, V.Gupta. Recent automatic text summarization techniques: a survey. 2016

Задача покрытия терминологии и тематики документа

Дано: коллекция, S_d — множество предложений документа d
Найти: суммаризацию как подмножество $a \subset S_d$

Критерии:

1) покрытие терминологии документа (lexicon coverage):

$$\text{WCov}(a) = \text{KL}\big(p(w|d)\|p(w|a)\big) \to \min_{a \subset S_d}$$

2) покрытие тематики документа (topic coverage):

$$\text{TCov}(a) = \text{KL}\big(p(t|d)\|p(t|a)\big) \to \min_{a \subset S_d}$$

3) избыточность суммаризации (redundancy):

$$\text{Red}(a) = \sum_{s,s' \in a} B_{ss'} \to \min_{a \subset S_d}, \qquad B_{ss'} = \text{sim}\big(p(w|s), p(w|s')\big),$$

где sim — одна из мер сходства: cos, JS, Jaccard и т.п.

Marina Litvak et al. Improving summarization quality with topic modeling. 2015.

Задача многокритериальной дискретной оптимизации

Метод релаксации: вместо $a \subset S_d$ ищем $\pi_s = p(s|a)$, где $s \in S_d$.

В релаксированной задаче:

$$p(w|a) = \sum_{s \in d} p(w|s)p(s|a) = \sum_{s \in d} \frac{n_{ws}}{n_s} \pi_s$$

$$p(t|a) = \sum_{s \in d} p(t|s)p(s|a) = \sum_{s \in d} \theta_{ts} \pi_s$$

Максимизация правдоподобия с регуляризацией:

$$\mathrm{WCov}(a) + \tau_1 \mathrm{TCov}(a) + \tau_2 \mathrm{Red}(a) =$$

$$\sum_{w \in d} n_{dw} \ln \sum_{s \in d} \frac{n_{ws}}{n_s} \pi_s + \tau_1 \sum_{t \in T} \theta_{td} \ln \sum_{s \in d} \theta_{ts} \pi_s - \tau_2 \sum_{s,s' \in d} B_{ss'} \pi_s \pi_{s'} \to \max_{\{\pi\}}$$

Можно добавить регуляризатор разреживания:

$$R(\pi) = -\tau_3 \sum_{s \in S_d} \ln \pi_s \to \max_{\{\pi\}}$$

Тематическая модель предложений для суммаризации

Дано: коллекция, S_d — множество предложений документа d;
n_{ws} — частота терма w в предложении s;
n_s — длина предложения s.

Найти: top-k тем $p(t|d)$ и top-k_d предложений $p(s|t)$ из S_d

Тематическая модель предложений: $p(s|d) = \sum_t p(s|t)p(t|d)$,

$$p(w|d) = \sum_{s \in S_d} p(w|s) \sum_{t \in T} p(s|t)p(t|d) = \sum_{t \in T} \sum_{s \in S_d} p_{ws}\psi_{st}\theta_{td},$$

где $p_{ws} \equiv p(w|s) = \frac{n_{ws}}{n_s}$ определяется из данных, $\psi_{st} \equiv p(s|t)$

Согласование тем в документах (гип. условной независимости):

$$p(w|t,d) = p(w|t) \quad \Rightarrow \quad \sum_{s \in S_d} p_{ws}\psi_{st} \approx \phi_{wt}$$

Dingding Wang, Shenghuo Zhu, Tao Li, Yihong Gong. Multi-document summarization using sentence-based topic models // ACL-IJCNLP 2009.

Согласованная ТМ для много-документной суммаризации

Критерий: строим две модели с общей матрицей Θ:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_{d,w} n_{dw} \ln \sum_{t \in T} \sum_{s \in S_d} p_{ws}\psi_{st}\theta_{td} + R \to \max_{\Phi,\Psi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$\begin{cases} p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ p_{stdw} \equiv p(s,t|d,w) = \underset{(s,t) \in S_d \times T}{\text{norm}}\left(p_{ws}\psi_{st}\theta_{td}\right) \end{cases}$$

M-шаг:
$$\begin{cases} \phi_{wt} = \underset{w \in W}{\text{norm}}\left(\sum_d n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \psi_{st} = \underset{s \in S_d}{\text{norm}}\left(\sum_w n_{dw} p_{stdw} + \psi_{st}\frac{\partial R}{\partial \psi_{st}}\right) \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(\sum_w n_{dw} p_{tdw} + \sum_w n_{dw} \sum_{s \in S_d} p_{stdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Нельзя ли сделать проще?

1. Построить обычную тематическую модель, а ещё лучше — тематическую модель локальных контекстов (см. Лекцию №2)

2. Определить *тематическую модель предложения s* через тематические векторы слов $p(t|w)$, $w \in s$:

$$p(s|t) = p(t|s)\frac{p(s)}{p(t)} = \sum_{w \in s} p(t|w)p(w|s)\frac{p(s)}{p(t)} = \sum_{w \in s} \phi_{wt}\frac{p(s)}{p(w)}p_{ws}$$

Но есть сомнение: тогда темы не согласуются с гипотезой, что в каждом предложении все слова порождаются из общей темы

Открытый вопрос: какой способ отбора предложений лучше?

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, Yihong Gong. Integrating clustering and multi-document summarization to improve document understanding. CIKM 2008

Dingding Wang, Shenghuo Zhu, Tao Li, Yihong Gong. Multi-document summarization using sentence-based topic models. ACL-IJCNLP 2009

Задача суммаризации темы

Имея top-k тем $p(t|d)$ и top-k_d предложений $p(s|t)$ из S_d можно решать три задачи:

1. суммаризация документа d, покрытие основных его тем
2. многодокументная суммаризация, покрытие основных тем
3. *многодокументная суммаризация заданной темы*

Задача покрытия лексики темы $p(w|t)$ предложениями $p(s|t)$:

Дано: коллекция D, предложения S_d, $d \in D$ матрица $\Phi = (\phi_{wt})$
Найти: подмножество предложений для суммаризации темы, в релаксированной форме: $\pi_{st} = p(s|t)$, где $s \in \sqcup_d S_d = S$
Критерий покрытия $\mathrm{KL}\big(p(w|t) \| \sum_s p(w|s)p(s|t)\big) \to \min$:

$$\sum_{t\in T}\sum_{w\in W} \phi_{wt} \ln \sum_{s\in S} p_{ws}\pi_{st} \to \max_{\Pi}$$

В такой форме задача суммаризации тем ещё не ставилась!

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

$r \in R$ — множество рефератов, написанных людьми

s — суммаризация, построенная системой

Чем больше, тем лучше — для всех метрик семейства ROUGE

Доля n-грамм из рефератов, вошедших в суммаризацию s:

$$\text{ROUGE-}n(s) = \frac{\sum\limits_{r \in R} \sum\limits_{w} [w \in s][w \in r]}{\sum\limits_{r \in R} \sum\limits_{w} [w \in r]}$$

Доля n-грамм из самого близкого реферата, вошедших в s:

$$\text{ROUGE-}n_{\text{multi}}(s) = \max_{r \in R} \frac{\sum\limits_{w} [w \in s][w \in r]}{\sum\limits_{w} [w \in r]}$$

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. 2004.

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

$r \in R$ — множество рефератов, написанных людьми
s — суммаризация, построенная системой
Чем больше, тем лучше — для всех метрик семейства ROUGE

ROUGE-L(s) максимальная общая подпоследовательность s, r
ROUGE-W(s) штрафует за пропуски в подпоследовательности
ROUGE-S(s) аналог ROUGE-2(s) для биграмм с пропусками
ROUGE-SU-m(s) для биграмм с пропусками не длиннее m

Йенсен-Шеннон $\mathrm{JS}\big(p(w|s), p(w|R)\big)$ лучше всего коррелирует с экспертными оценками качества суммаризации (Lin, 2006).

Готовые пакеты для вычисления метрик: pyRouge и др.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. 2004.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, Jian-Yun Nie. An Information-Theoretic Approach to Automatic Evaluation of Summaries. 2006.

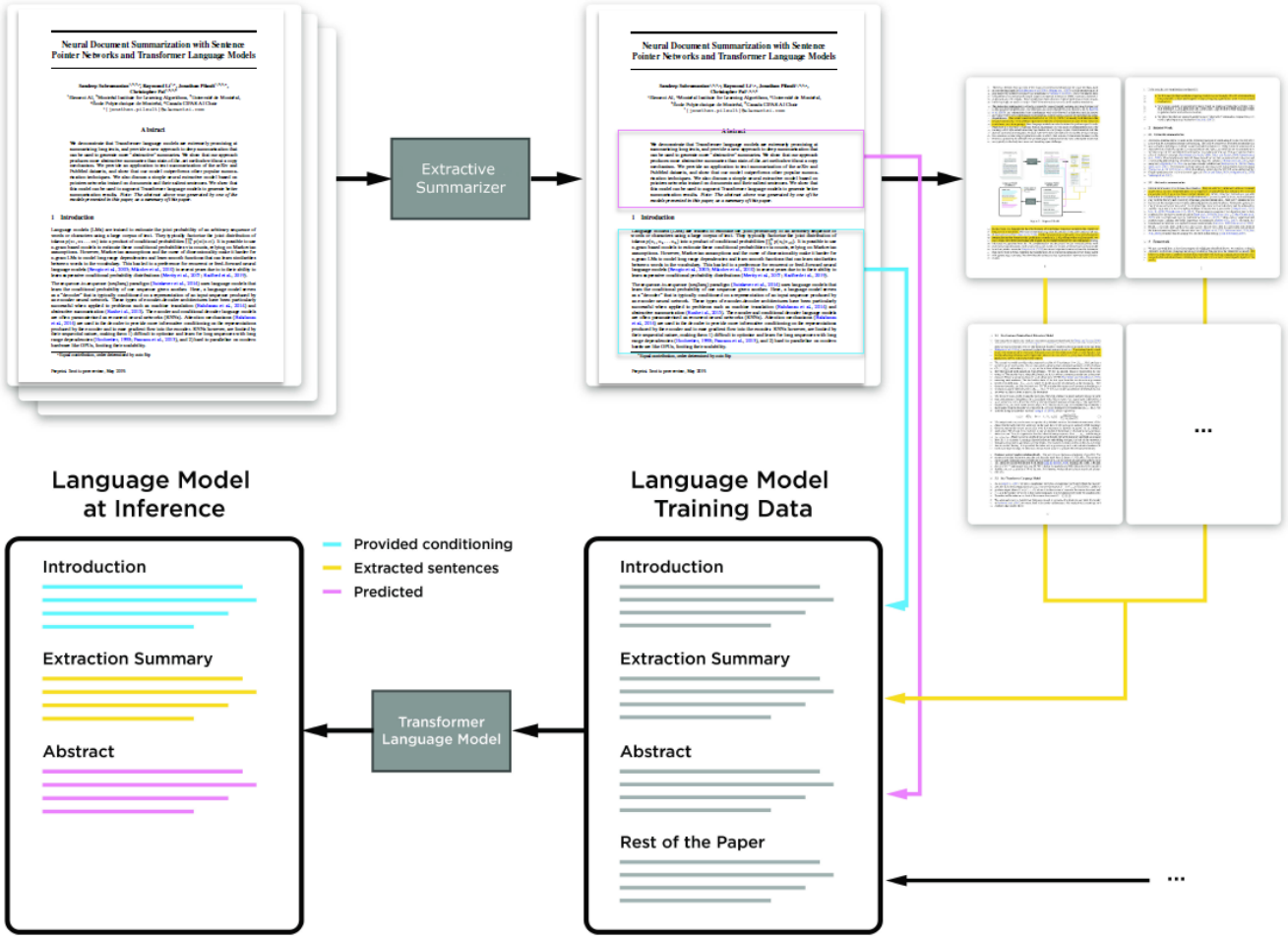
Абстрактивная суммаризация на основе трансформеров

Abstract

We present a method to produce abstractive summaries of long documents that exceed several thousand words via neural abstractive summarization. We perform a simple extractive step before generating a summary, which is then used to condition the transformer language model on relevant information before being tasked with generating a summary. We show that this extractive step significantly improves summarization results. We also show that this approach produces more abstractive summaries compared to prior work that employs a copy mechanism while still achieving higher rouge scores. *Note: The abstract above was not written by the authors, it was generated by one of the models presented in this paper.*

S.Subramanian, R.Li, J.Pilault, C.Pal. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. 2019.

Абстрактивная суммаризация использует экстрактивную



S.Subramanian, R.Li, J.Pilault, C.Pal. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. 2019.

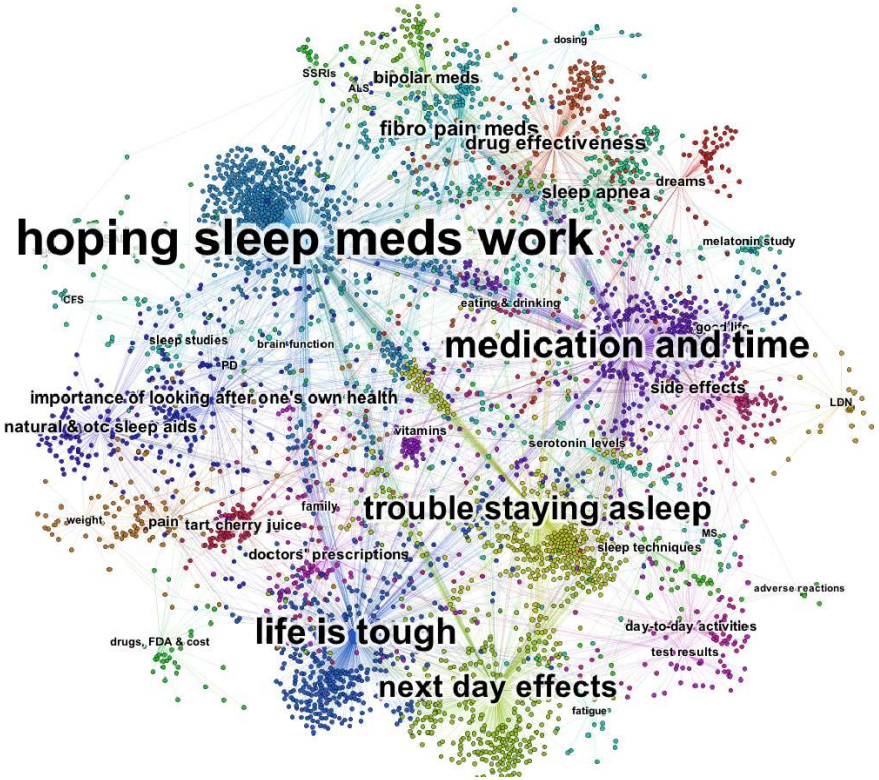
Резюме по суммаризации

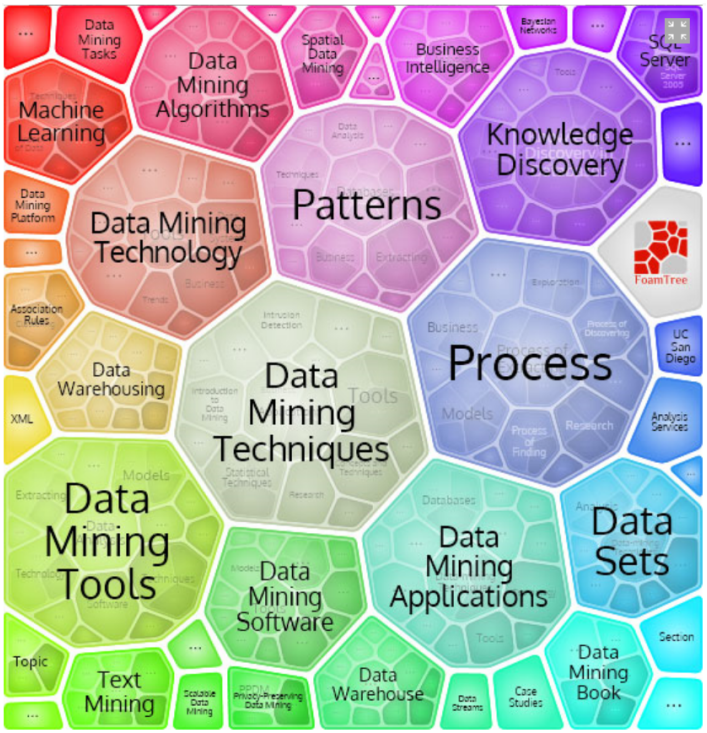
- Тематические модели в экстрактивной суммаризации — для выделения и покрытия наиболее важных тем
- *Суммаризация темы* — открытая проблема TM, до появления LLM её не решали, и даже не ставили! «Let the topics tell about themselves!»
- ROUGE — семейство мер качества суммаризации, характеризуют далеко не все аспекты качества
- BLUE — аналогичные метрики, но precision-based
- Для визуализации нужны суммаризация и именование тем

Автоматическое именование тем для визуализации

Пример 1: тематика обсуждений на www.PatientsLikeMe.com

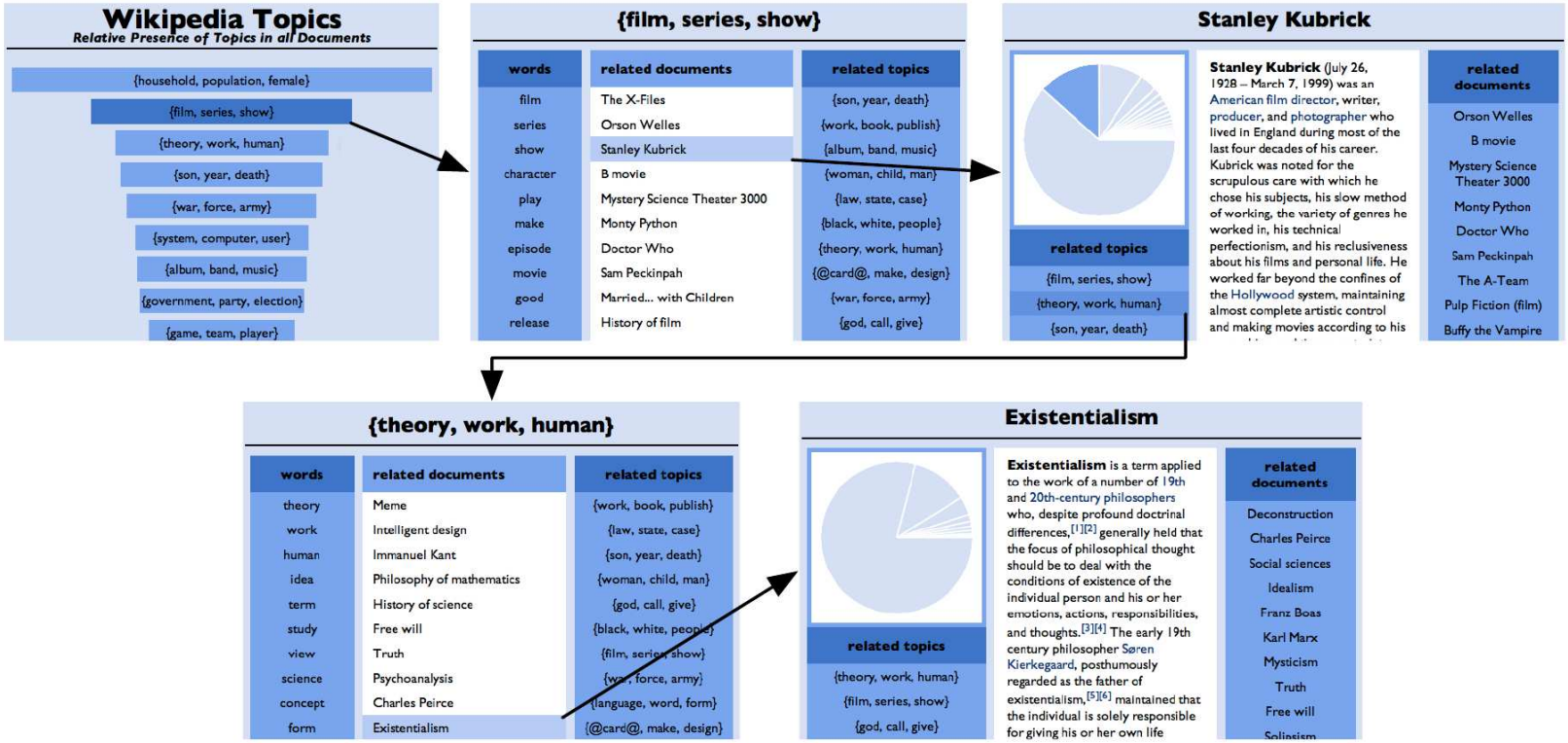
Пример 2: иерархическая карта *Data Mining*





Система TMVE — Topic Model Visualization Engine

Три топовых слова темы — самая простая модель именования:



https://github.com/ajbc/tmv

Chaney A., Blei D. Visualizing Topic Models // Frontiers of computer science in China, 2012. — 55(4), pp. 77–84.

Задача автоматического именования тем (topic labeling)

Требования к *названию темы* (topic label):

- интерпретируемость и грамматическая корректность
- точность представления семантики темы
- полнота представления семантики темы
- непохожесть на названия других тем, даже похожих

Гипотеза: все названия уже придуманы, осталось их найти.

Подзадачи

- формирование названий-кандидатов $\ell_1, \ldots, \ell_m$
- построение (обучение) функции релевантности $s(\ell, t)$
- выбор названия с учётом названий похожих тем

Qiaozhu Mei (ЦяоЧжу Мэй), Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Способы формирования названий-кандидатов

Отбираются вероятно специфичные для данной темы:

- топовые n-граммы данной темы
- синтаксические ветки наиболее тематичных предложений
- тематичные именные группы (вырезанные OpenNLP chunker)
- тематичные фразы «объект, субъект, действие»
- заголовки тематичных документов или их фрагменты
- метаданные (теги, категории) тематичных документов

Общие для всех тем:

- n-граммы из внешней коллекции, например, Википедии
- заголовки статей или категорий Википедии
- термины из внешних тезаурусов:
  WordNet, РуТез, Викисловарь, и др.

Функция релевантности (relevance score)

Релевантность нулевого порядка:

$$s(\ell, t) = \sum_{w \in \ell} \log \frac{p(w|t)}{p(w)} \to \max$$

Релевантность первого порядка: слова темы t неслучайно часто появляются рядом (в одном контексте C) с названием ℓ:

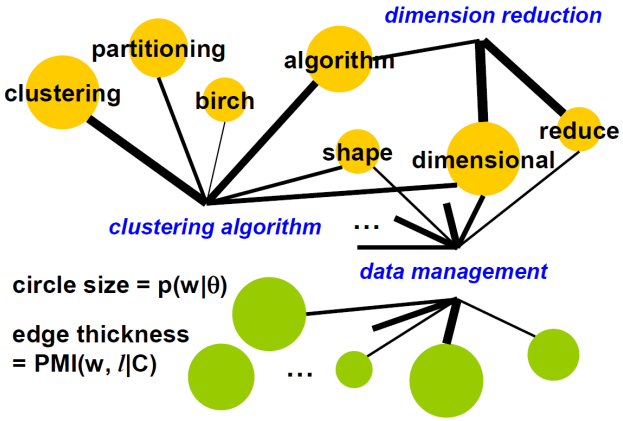
$$s(\ell, t) = \sum_{w \in C} p(w|t) \underbrace{\log \frac{p(w, \ell|C)}{p(w|C)p(\ell|C)}}_{\mathrm{PMI}(w,\ell|C)} \to \max$$

где C — релевантный теме контекст, в котором ожидается появление как слов темы t, так и названия ℓ целиком (Например, статья или категория Википедии).

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Проблема названий, подходящих для нескольких тем

Пример: оранжевая тема покрывается двумя названиями:
— *clustering algorithm*
— *dimension reduction*
третье название *data management* неудачно, конкурирует с другой темой



Выбирать каждое следующее название, чтобы оно было

- максимально релевантно, $s(\ell, t) \to \max$,
- максимально не похоже на названия ℓ' остальных тем:

$$s(\ell, t) + \lambda \max_{\ell'} \mathrm{KL}(\ell' \| \ell) \to \max$$

где параметр λ подбирается эмпирически.

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Максимизация различности названий различных тем

Модифицированная функция релевантности $s'(\ell, t)$:

- максимизирует релевантность своей темы, $s(\ell, t) \to \max$
- минимизирует релевантность других тем, $s(\ell, t') \to \min$

$$s'(\ell, t) = s(\ell, t) - \mu \sum_{t' \in T \setminus t} s(\ell, t') \to \max$$

где параметр μ подбирается эмпирически.

Методика оценивания качества именования тем:

- 3 асессора, каждый асессор видит для каждой темы:
  — список топ-слов темы, список топ-документов темы
  — варианты названия, сгенерированные разными методами
- асессор ранжирует методы $0, 1, 2, \ldots$ (чем выше, тем лучше)

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Оценивание качества именования тем

Две коллекции: научная (SIGMOD), новостная (Assoc.Press)
Автоматические и асессорские названия тем, SIGMOD:

Auto Label	**clustering algorithm**	**r tree**	**data streams**	**concurrency control**
Man. Label	**clustering algorithms**	**indexing methods**	**Stream data management**	**transaction management**
θ	clustering clusters video dimensional cluster partitioning quality birch	tree trees spatial b r disk array cache	stream streams continuous monitoring multimedia network over ip	transaction concurrency transactions recovery control protocols locking log

Победил выбор n-грамм по релевантности 1-го порядка,
но он всё ещё заметно хуже человеческого именования тем:

Baseline v.s. Zero-order v.s. First-order				
Dataset	#Label	Baseline	Ngram-0-B	Ngram-1
SIGMOD	1	0.76	0.75	**1.49**
SIGMOD	5	0.36	1.15	**1.51**
AP	1	0.97	0.99	**1.02**
AP	5	0.85	0.66	**1.48**

System v.s. Human			
Dataset	#Label	Ngram-1	Human
SIGMOD	1	0.35	**0.65**
SIGMOD	5	0.25	**0.75**
AP	1	0.24	**0.76**
AP	5	0.21	**0.79**

Qiaozhu Mei et al. Automatic labeling of multinomial topic models. KDD 2007.

Резюме по автоматическому именованию тем

- *Automatic Topic Labeling* — очень узкое направление, около 50 статей начиная с 2007 г.
- Важно для автоматизации создания приложений
- Близко к задаче суммаризации темы
- Для иерархических моделей добавляется специфичное требование *полноты*: названия дочерних тем должны отражать специфику их различий в родительской теме

Alex Yoo. Automatic topic labeling in 2018: history and trends.
https://medium.com/datadriveninvestor/automatic-topic-labeling-in-2018-history-and-trends-29c128cec17

A.Gourru et al. United we stand: Using multiple strategies for topic labeling. 2018.

Ciprian-Octavian Truicam And Elena-Simona Apostol TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition. 2021.

Supriya Kinariwala, Sachin Deshmukh Onto_TML: Auto-labeling of topic models. 2021.

M.Allahyari, S.Pouriyeh, K.Kochut, H.R.Arabnia. A knowledge-based topic modeling approach for automatic topic labeling. 2017.

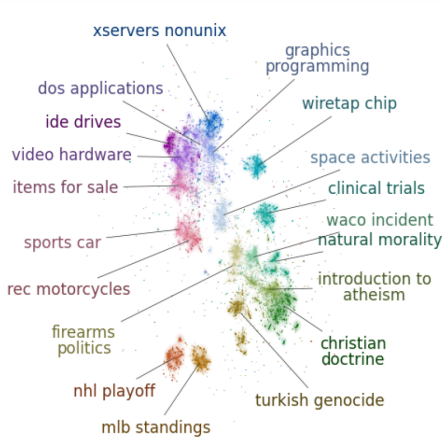
Начинаем подводить итоги по всему курсу

Что точно работает в тематическом моделировании
...или «узнав об этом, по другому уже не захочется»

1. лемма о максимизации функции на единичных симплексах
2. аддитивная регуляризация вместо байесовского обучения
3. библиотека BigARTM: скорость + функциональность
4. однопроходный EM-алгоритм, тематическое внимание
5. многокритериальное оценивание
6. словари терминов n-грамм
7. декоррелирование тем, выделение фоновых тем
8. мультимодальные модели
9. тематические иерархии
10. спектр тем (ранжирование тем по взаимной близости)

Нейросетевая тематическая модель Contextual-Top2Vec

Вместо PTM — сумма технологий:



1. векторизация токенов (Sentence-BERT)
2. векторизация предложений скользящим окном в 50 токенов (mean pooling)
3. понижение размерности векторов (UMAP)
4. иерархическая кластеризация (hDbscan) с автоматическим определением числа тем
5. иерархичекое укрупнение тем слиянием мелких кластеров с ближайшими соседями (Top2Vec)
6. разбиение документа на монотематические сегменты
7. $p(t|d)$ = доля векторов данной темы в документе
8. именование тем: поиск фраз, ближайших к центроиду темы

Dimo Angelov. Top2vec: Distributed representations of topics. 2020.
D.Angelov, D.Inkpen. Topic modeling: contextual token embeddings are all you need. 2024.

Нейросетевая тематическая модель Contextual-Top2Vec

Достоинства:

- модель BERT предобучена по большим внешним данным, поэтому качество тем не зависит от размера коллекции
- документ разбивается на монотематические сегменты
- автоматически определяется число тем (hDbscan, Top2Vec)
- тема описывается фразами, а не отдельными словами

Недостатки:

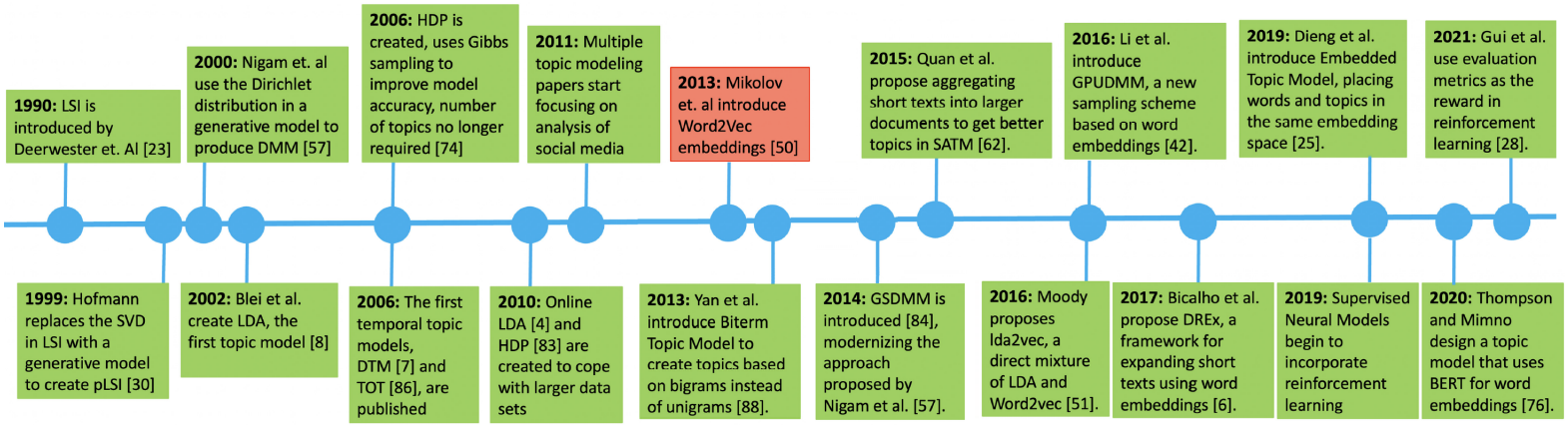
- это работает долго, особенно на больших коллекциях
- инкрементное добавление документов не предполагается

Сходство локализованного E-шага:

- обработка локальных контекстов скользящим окном
- возможно разреживать $p(t|C_i)$ до монотематичности

Dimo Angelov. Top2vec: Distributed representations of topics. 2020.
D.Angelov, D.Inkpen. Topic modeling: contextual token embeddings are all you need. 2024.

Обсуждение. Эволюция тематического моделирования



Neural Topic Models — поток публикаций начиная с 2016

Как «объединить лучшее от двух миров»?

- **Neural:** качество, универсальность, генеративность
- **Topic:** скорость, интерпретируемость, простота

Что объединяет: векторизация, оптимизация, регуляризация, гомогенизация, локализация (контекст и внимание)

X.Wu et al. A survey on neural topic models: methods, applications, and challenges. 2023.
Rob Churchill, Lisa Singh. The evolution of topic modeling. 2022.
He Zhao et al. Topic modelling meets deep neural networks: a survey. 2021.

Задания по курсу

Задача-минимум: научиться решать задачи NLP
с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

виды деятельности	оценка
теоретические задания	$\sum_i X_i$
решение прикладной задачи	$5X$
обзор по последним NeuralTM	$5X$
интеграция ARTM в pyTorch	$5X$
участие в одном из проектов	$10X$
работа над открытой проблемой	$10X$

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min\big(10, \lfloor \text{score}/5 \rfloor\big)$ по 10-балльной шкале.

Теоретическое задание к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$
Найти параметры модели ξ_{dw}.

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d
Найти параметры модели ξ_w.

Подсказка: применить условия KKT или основную лемму.

3. Творческое задание (возможны разные решения)
Предложите модель, определяющую роли слов в текстах:
— тематические слова
— специфичные слова документа (шум)
— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т}, \text{ш}, \text{ф}\}$.
Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.
Подсказка 3: можно использовать документную частоту слов.

Теоретическое задание к лекции 2

4. Запишите критерий логарифма правдоподобия с регуляризацией для тематической модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, используя исходные данные $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw}. Выведете из него EM-алгоритм, докажите его эквивалентность обычному EM-алгоритму для ARTM.

5. Запишите критерий логарифма правдоподобия для локализованной тематической модели $p(w|C_i) = \sum_t \phi_{wt} p(t|C_i)$. Выведете из него EM-алгоритм с локализованным E-шагом.

Какие приближения пришлось сделать в процессе вывода? Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw}?

6. Творческое задание (возможны разные решения) Предложите «какую-нибудь разумную» параметризацию для тематической модели внимания. Используя «основную лемму», получите уравнения для новых параметров модели.

Исследовательское задание к лекции 2

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный E-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i$, $\overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайнового EM-алгоритма
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на E-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}$, $\overleftarrow{\theta}_{ti}$»

Теоретическое задание к лекции 3

7. Выведете формулы EM-алгоритма в случае, когда логарифм в функции потерь заменяется гладкой монотонно возрастающей функцией ℓ:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ell\left(\sum_{t\in T}\phi_{wt}\theta_{td}\right) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

Подумайте, какие замены логарифма полезны, и почему.

8. Замените ln гладкой монотонно возрастающей функцией μ в регуляризаторе сглаживания–разреживания (модель LDA):

$$R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W}\beta_w\mu(\phi_{wt}) + \sum_{d\in D}\sum_{t\in T}\alpha_t\mu(\theta_{td}).$$

Как изменится M-шаг и воздействие регуляризатора на модель?

9. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \operatorname*{norm}_w\big(n_{wt}[n_{wt} > \gamma n_t]\big)$$

Теоретическое задание к лекции 3

Аналитик построил тематическую модель Φ^0, Θ^0
и отметил среди столбцов матрицы Φ^0 темы двух типов:
удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

10. Предложите регуляризаторы для этого.

11. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi^0_{wt}$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

12. Предложите способ инициализации Φ для новой модели.

Исследовательские задания к лекции 4

- Проблема несбалансированности тем
  - генераторы синтетических несбалансированных коллекций
  - модели локального контекста лишены этой проблемы?
  - регуляризаторы декоррелирования + семантической однородности
- Семейство средневзвешенных статистик
  - генераторы синтетических коллекций, удовлетворяющих гипотезе условной независимости
  - как (и нужно ли) определять пороги для построения статистических тестов условной независимости?
  - как ослабить проверку гипотезы условной независимости в модели локального контекста?
  - как перестраивать несогласованные темы?
- Критерий внутритекстовой когерентности
  - найти лучший вариант критерия с помощью калибровки по размеченным тематическим цепочкам
  - вычисление критерия должно естественным образом встраиваться в модель локального контекста

Теоретическое задание к лекции 5

13. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = \big(p(s|t)\big)$, гарантирующий, что
1) у каждой родительской темы будет хотя бы одна дочерняя;
2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу М-шага для матрицы Ψ.

14. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s, то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$, то есть тема t на данном уровне не расщепляется на подтемы.

15. Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum_{s \in S} p(t|s)p(s|d)$ с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$.

Исследовательское задание к лекции 5

Участие в проекте «Мастерская знаний»

Дано:

- подборки, сгенерированные SciRus по одной статье
- асессорская разметка статей подборки по релевантности
- несколько вариантов токенизации
  — в том числе с автоматическим выделением терминов

Найти:

- тематическую модель
- модель ранжирования подборки по релевантности
- оптимальные: токенизацию, число тем, регуляризаторы
- распределение терминов по тематичности

Критерий:

- качество ранжирования
- (визуально) интерпретируемость тем
  — в том числе автоматического именования тем

Теоретическое задание к лекции 9

16. Выведите EM-алгоритм для тематической модели битермов (Biterm Topic Model) из предыдущей лекции.

17. Выведите EM-алгоритм для тематической модели с гладким регуляризатором $R(\Phi, \Theta)$ и суммой L_1-регуляризаторов

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) - \sum_{j \in J} \lambda_j \big| R_j(\Phi, \Theta) \big|.$$

Подсказка: ввести дополнительные неотрицательные переменные, чтобы избавиться от негладкой функции модуля, затем применить условия Каруша–Куна–Таккера.

18. Запишите формулы M-шага для частного случая — L_1-сглаживания $p(i|t) = \phi_{it}$ в соседних интервалах i, $i-1$:

$$R_{\text{сгл}}(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \big| \phi_{it} - \phi_{i-1,t} \big| \to \max.$$

Примеры датасетов для практических заданий по курсу

- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Научные статьи: eLibrary, Semantic Scholar, arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- TechCrunch (английский)
- Данные социальных сетей: VK, Twitter, Telegram,...
- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

---

http://bigartm.org
http://drive.google.com/drive/folders/1PPnw6aZOJAJoLRYuwdGm437RssV-XQx0

Проекты

- «Тематизатор» для социо-гуманитарных исследований
  — пользователь задаёт грубый фильтр текстового потока
  — задача: «классифицировать иголки в стоге сена»,
  — разделив темы на информативные и мусорные,
  — выделив аспекты и тональности в каждой теме
  — конечная цель: q&q аналитика проблемной среды
- «Мастерская знаний» для научного поиска
  — пользователь строит тематические подборки статей,
  — поисковая выдача формируется моделью SciRus.
  — задача: показать пользователю тематику подборки
  — понадобится автоматическое выделение терминов,
  — выделение тематических фраз из документов,
  — автоматическое именование и суммаризация тем
  — конечная цель: ускорить понимание предметной области

Открытые проблемы тематического моделирования

1. Проблема несбалансированности тем в коллекции
2. Обеспечение 100%-й интерпретируемости тем
3. Тематические модели внимания последовательного текста
4. Обнаружение новых тем или трендов в потоке текстов
5. Автоматическое именование и аннотирование тем
6. Подбор архитектуры нейросетевых тематических моделей
7. Обеспечение полноты и устойчивости множества тем
8. Автоматический подбор гиперпараметров, AutoML
9. Оптимизация гиперпараметров в потоковом режиме
10. Проблема несбалансированности текстов по длине
11. Бережное слияние моделей нескольких коллекций
12. Гиперграфовые тематические модели в RecSys