Вероятностные тематические модели
Лекция 12.
Суммаризация и именование тем

К. В. Воронцов
vokov@forecsys.ru

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

МФТИ – ФИЦ ИУ РАН • 7 декабря 2023

Содержание

Задача суммаризации (реферирования, аннотирования) текста

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

Полуавтоматическая суммаризация

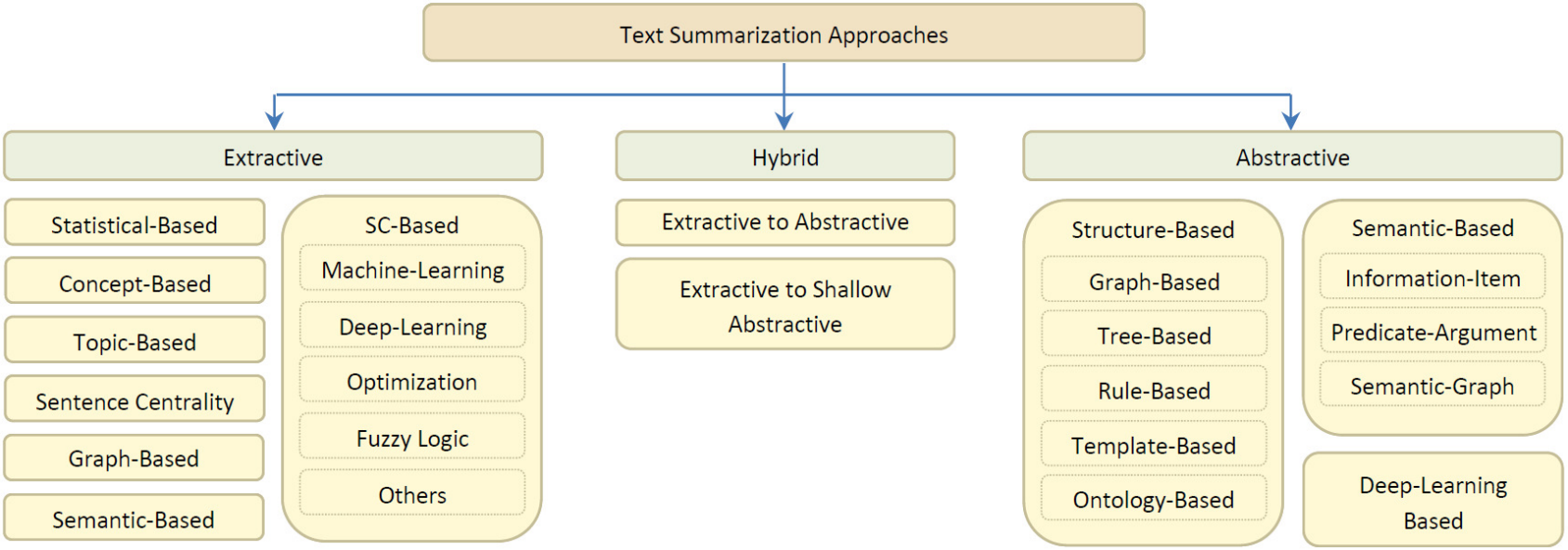
- MAHS, machine aided human summarization
- HAMS, human aided machine summarization

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 $p(d, s|t)$

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 (ранжирование) предложений
3. Отбор предложений для реферата
   — оптимизация критериев релевантности + различности
   — оптимизация последовательности предложений
   — учёт целей и особенностей прикладной задачи (новости/статьи/веб-страницы/посты/мэйлы)

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

TextRank — аналог ссылочного ранжирования PageRank

Текст — граф предложений; рёбра — похожие предложения.

Предложение $s \in S$ тем важнее,

- чем больше других предложений c, похожих на s,
- чем важнее предложения c, похожие на s,
- чем меньше предложений, на которые c также похоже.

Вероятность попасть в s, случайно блуждая по графу:

$$\mathrm{TR}(s) = (1-\delta)\frac{1}{|S|} + \delta \sum_{c \in S_s^{in}} \frac{\mathrm{TR}(c)}{|S_c^{out}|},$$

$S_s^{in} \subset S$ — множество предложений c, похожих на s,
$S_c^{out} \subset S$ — множество предложений, на которые похоже c,
$\delta = 0.85$ — вероятность продолжать блуждания (damping factor)

S.Brin, L.Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. 1998
R.Mihalcea, P.Tarau. TextRank: Bringing Order into Text. EMNLP-2004

Определение сходства предложений

- Доля общих слов в двух предложениях
- Доля общих слов, за исключением слов общей лексики
- Доля общих n-грамм в двух предложениях
- Сходство тематических распределений двух предложений
- Сходство векторных представлений двух предложений

Другое применение TextRank — *извлечение ключевых слов* (keyword extraction) из отдельных документов.

В этом случае близость между словами (n-граммами) определяется по частоте их сочетаемости в окне ширины h

Покрытие терминологии и тематики документа

S_d — множество предложений документа d
$a \subset S_d$ — искомая суммаризация

Покрытие терминологии документа (lexicon coverage):

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

Покрытие тематики документа (topic coverage):

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

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:

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

Оценка полезности предложений

Дополнительные признаки для отбора предложений:

- *SumBasic* — средняя частота слов, исключая стоп-слова
- *Centriod* — средний TF-IDF слов, превышающий порог
- *LexicalChain* — число слов сильных лексических цепочек
- *ImpactBased* — число слов из ссылающихся контекстов
- *TopicBased* — число слов из запроса пользователя

Стратегии отбора предложений:

- по одному top-предложению от каждой из top-тем
- поощрять выбор соседних предложений
- поощрять более простые (удобочитаемые) предложения
- штрафовать предложения с анафорой и эллипсисом

A.Nenkova, K.McKeown. A survey of text summarization techniques. 2012.

Тематическая модель предложений для суммаризации

S_d — множество предложений документа d;
n_{sw} — частота терма w в предложении s;
n_s — длина предложения s.

Отбор тем: $p(t|d) \to \operatorname{top}_{t \in T} k$ **и предложений:** $p(s|t) \to \max_{s \in S_d}$

Тематическая модель сегментированного текста:

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

где $p_{ws} \equiv p(w|s) = \frac{n_{ws}}{n_s}$ — частота терма w в предложении s.

Вместо ϕ_{wt} нельзя взять $p(w|t) = \sum_{d \in D} \sum_{s \in S_d} p_{ws} \psi_{st}$. Почему?

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

BSTM — Bayesian Sentence-based Topic Models

Критерий максимума регуляризованного правдоподобия:

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

- Авторы утверждают, что модель переходит в обычную $p(w|d) = \sum_t \phi_{wt}\theta_{td}$, если предложение $\equiv$ слово
- Это не так, ведь предложения уникальны: $S_d \cap S_{d'} = \varnothing$
- Модель разваливается на независимые модели документов (Litvak, 2015) такую LDA строят явно, это тоже работает!
- Но это не будет работать для multi-document summarization!
- А то, что модель «Bayesian», вообще не имеет значения ;)

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

Идея обобщения для много-документной суммаризации

Критерий максимума регуляризованного правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

$\mathrm{JS}\big(p(w|s), p(w|R)\big)$ — 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.

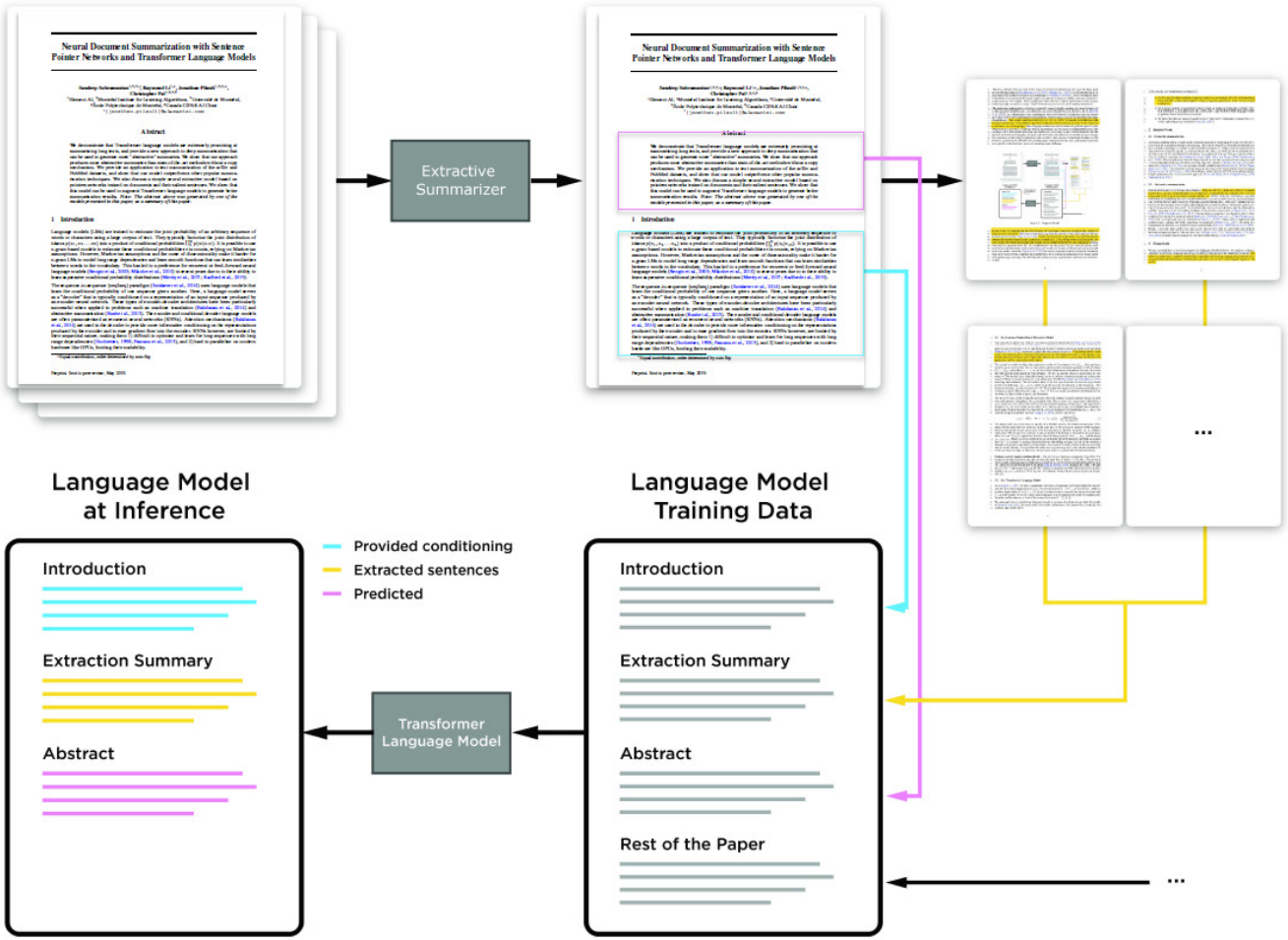
Абстрактивная суммаризация на основе трансформеров

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.

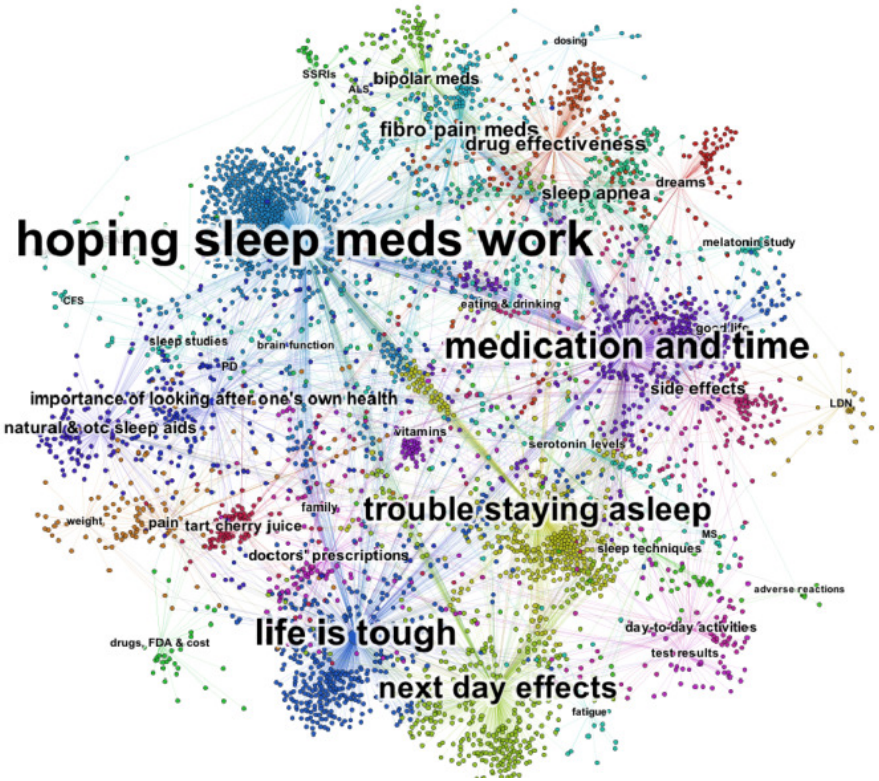
Резюме по суммаризации

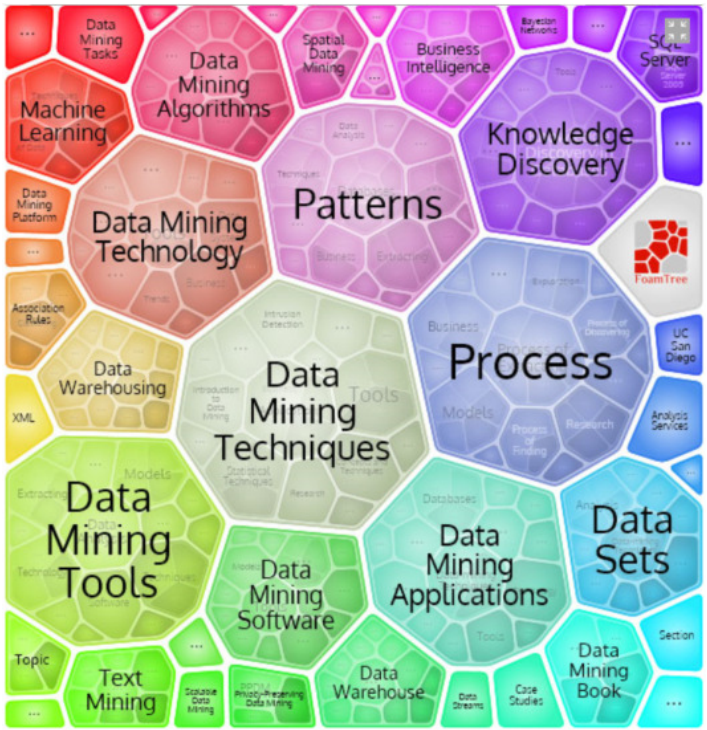
- Тематические модели используются в суммаризации, чтобы выделить и покрыть наиболее важные темы
- Суммаризация темы — открытая проблема TM, её не только не решали, но даже и не ставили! «Let the topics tell about themselves!»
- ROUGE — семейство мер качества суммаризации, характеризуют далеко не все 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

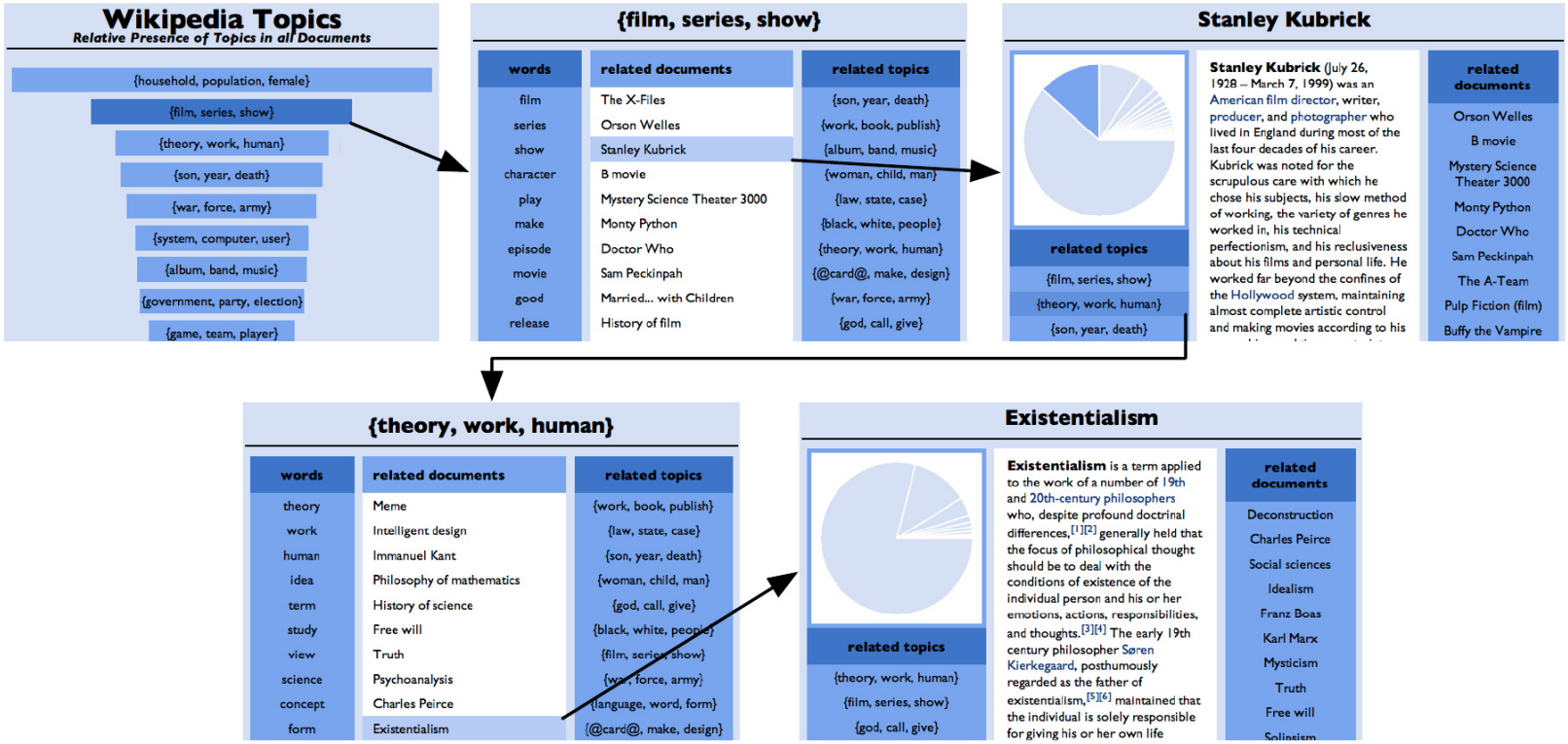
Пример 2: иерархическая карта *Data Mining*





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, включая похожие

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

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, PyTez, 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:

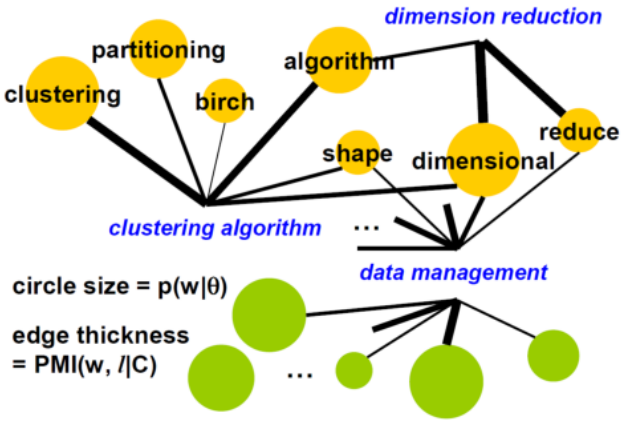
$$s(\ell, t) = \sum_{w \in C} p(w|t) \underbrace{\log \frac{p(w, \ell|C)}{p(w|C)p(\ell|C)}}_{\mathrm{PMI}(w,\ell|C)} \to \max$$

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*
но 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**

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 адекватно описывать разделение родительской темы

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.

Что точно работает в тематическом моделировании

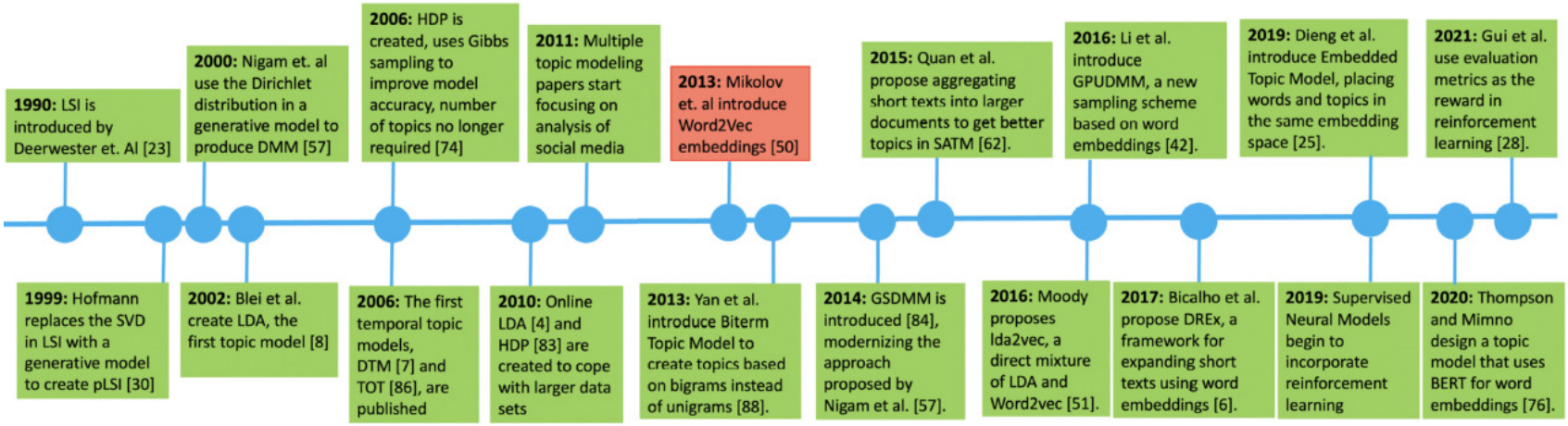
...или «узнав об этом, по другому уже не захочется»

1. многокритериальная регуляризация, ARTM и BigARTM
2. многокритериальное оценивание
3. n-граммы
4. декоррелирование
5. модальности
6. однопроходный EM-алгоритм
7. тематические иерархии
8. спектр тем
9. реализация ARTM на PyTorch
10. установка BigARTM под Python до 3.11 включительно

Открытые задачи тематического моделирования

1. ТЕМАТИЗАТОР =)
2. Проблема несбалансированности тем
3. Доля интерпретируемых тем должна быть 100%
4. Автоматическое именование и аннотирование тем
5. Автоматическое обнаружение новых тем в пакетах
6. Автоматическое разделение тем на подтемы
7. Автоматический подбор гиперпараметров, AutoML
8. Оптимизация гиперпараметров в потоковом режиме
9. Тематическая модель внимания (локальных контекстов)
10. Обеспечение полноты и устойчивости множества тем
11. Бережное слияние моделей нескольких коллекций
12. Развитие нейросетевых тематических моделей

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)

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.
He Zhao et al. Topic Modelling Meets Deep Neural Networks: A Survey. 2021.