

Вероятностные тематические модели

Лекция 11.

Визуализация и суммаризация

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)»

ВМК МГУ • 16 мая 2019

1 Визуализация тематических моделей

- Визуализация матричного разложения
- Карты, динамика, иерархия, сегментация
- Проект VisARTM

2 Суммаризация текстов

- Оценивание и отбор предложений для суммаризации
- Тематическая модель предложений для суммаризации
- Метрики качества суммаризации

3 Автоматическое именоване тем

- Формирование названий-кандидатов
- Максимизация функции релевантности
- Максимизация покрытия и различности

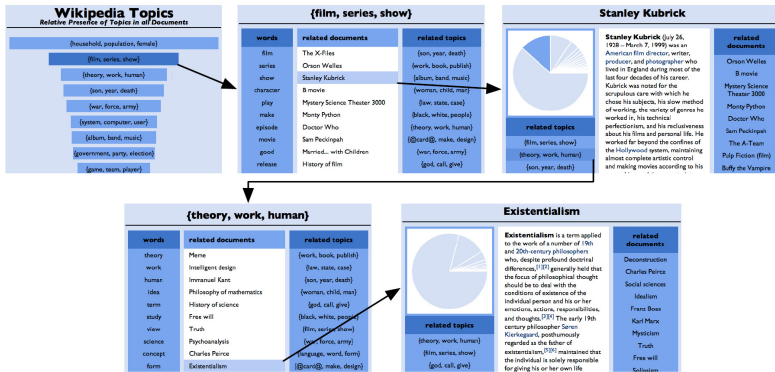
Что можно визуализировать

Одна из целей тематического моделирования — систематизация результатов информационного поиска.

- текстовое представление темы: название, топ-слова, топ-термы, топ-документы, аннотация, близкие темы
- масштабируемая тематическая карта коллекции
- иерархия тем
- граф связей между темами
- текст документа: темы слов или термов, сегментация
- графическая тематическая сегментация документа
- динамика тем во времени: временные ряды, реки тем
- иерархия + динамика

Система TMVE — Topic Model Visualization Engine

Тематический навигатор с веб-интерфейсом:

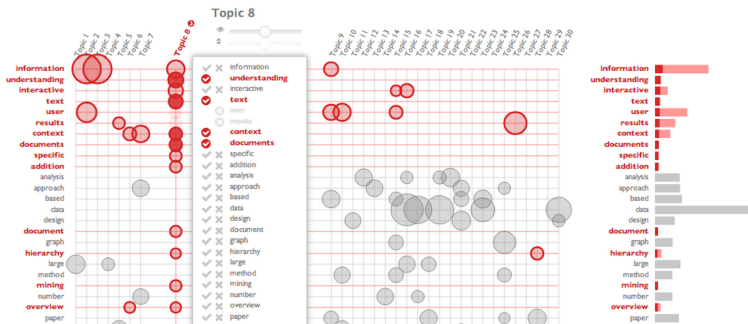


<https://github.com/ajbc/tmv>

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

Система Termite

Интерактивная визуализация матрицы Φ и сравнение тем:

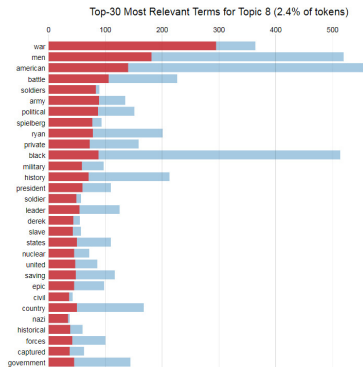
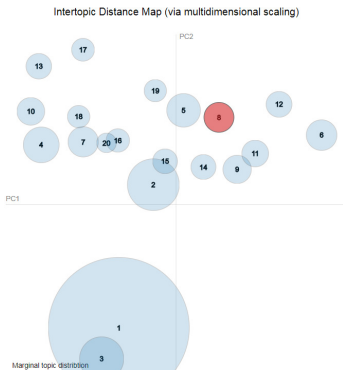


<https://github.com/uwdata/termite-visualizations>

Chuang J., Manning C., Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models. IWCAVI 2012.

Система LDAvis

Карта сходства тем и сравнение $p(w|t)$ с $p(w)$:

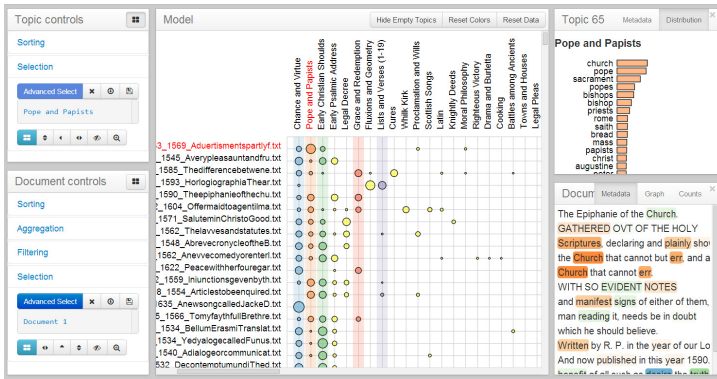


<https://github.com/cpsievert/LDAvis>

C.Sievert, K.Shirley. LDAvis: A method for visualizing and interpreting topics. 2014.

Система Serendip

Визуализация матриц Φ , Θ и тематики слов в текстах:

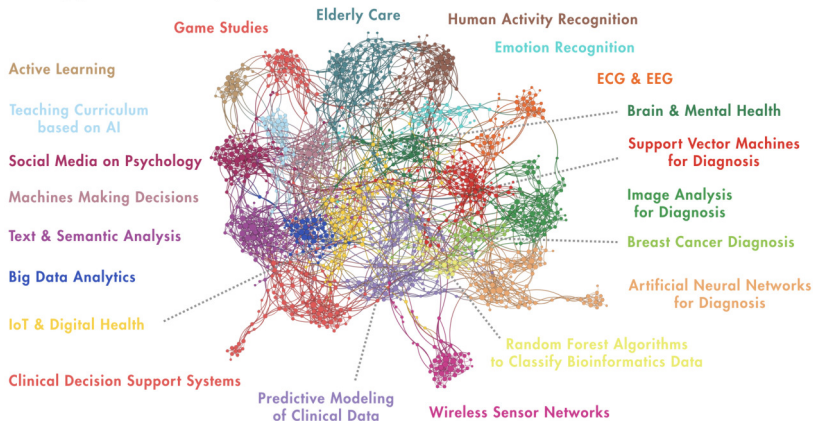


<http://vep.cs.wisc.edu/serendip>

E.Alexander, J.Kohlmann, R.Valenza, M.Witmore, M.Gleicher. Serendip: Topic Model-Driven Visual Exploration of Text Corpora. IEEE VAST 2014.

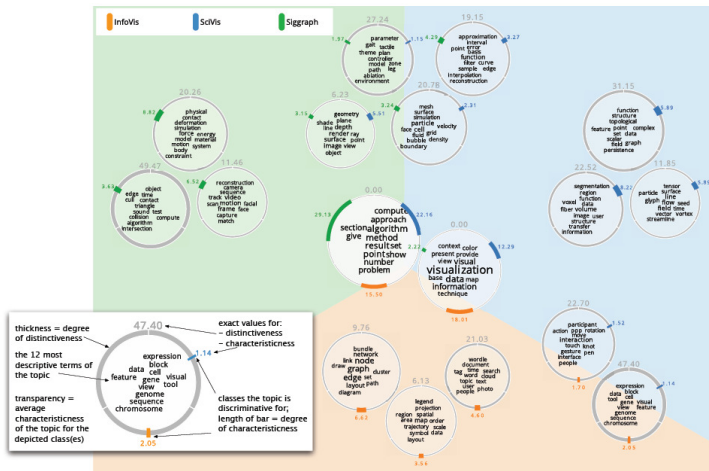
Ещё одна карта тематической кластерной структуры

Academic papers on AI in Healthcare published in 2016



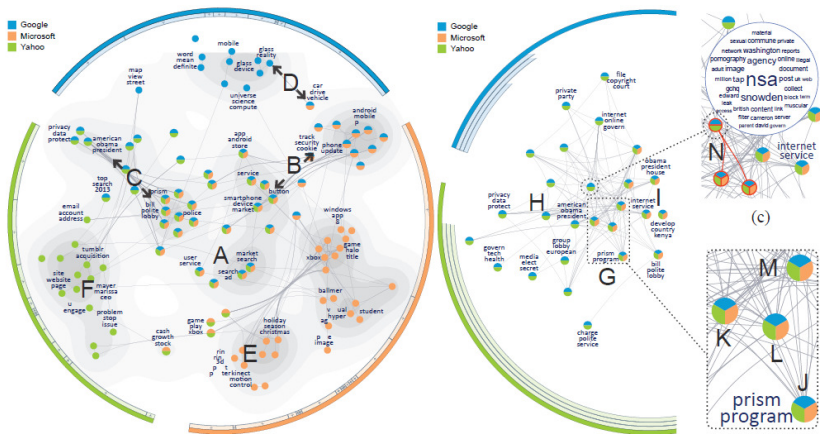
C.Folgar, J.McCuan. The 3 most-cited studies in healthcare and AI. Quid, 2017.

Тематический анализ источников



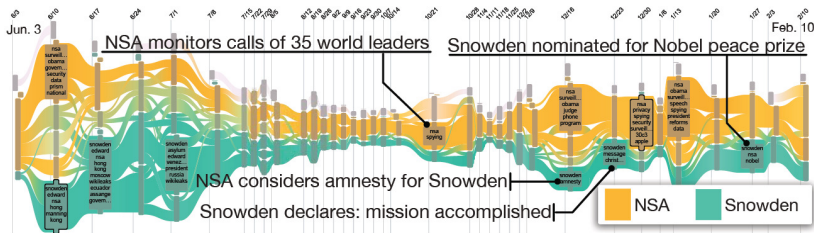
Oelke D., Strobelt H., Rohrdantz C., Gurevych I., Deussen O. Comparative exploration of document collections: a visual analytics approach. EuroVis. 2014.

Тематический анализ источников



Shixia Liu, Xiting Wang, Jianfei Chen, Jun Zhu, Baining Guo. TopicPanorama: a full picture of relevant topics. IEEE VAST, 2014.

Динамика тем: эволюция предметной области



Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- генерирует отчёт.

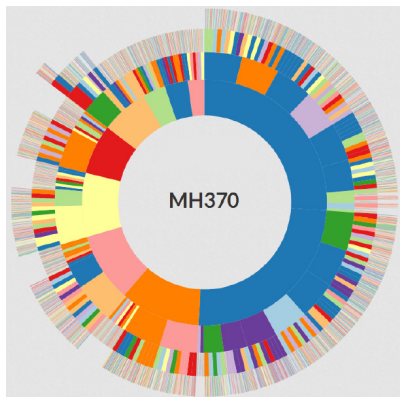
Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

Тематическая иерархия: структура предметных областей



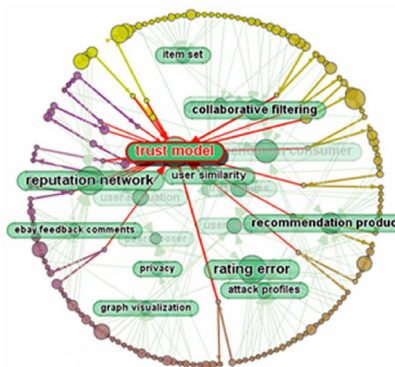
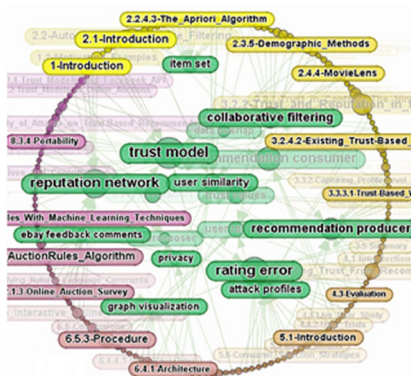
<https://carrotsearch.com/foamtree>

Тематическая иерархия: альтернативное представление



Smith A., Hawes T., Myers M. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

Тематическая сегментация документа запроса



Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

<http://textvis.lnu.se>

Интерактивный обзор 430 средств визуализации текстов



Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // JMLDA, 2015.

VisARTM: визуализация для BigARTM

- Web-приложение для визуализации ARTM моделей
- Открытый код: <https://github.com/bigartm/visartm>
- Автоматическое перестроение моделей через BigARTM
- Текстовые интерактивные визуализации документов, тем, термов, модальностей
- Графическая визуализация иерархических моделей
- Графическая визуализация темпоральных моделей
- Тематические спектры
- Сбор ассессорских оценок

Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация. МФТИ, 2017.

VisARTM: Визуализация документа

Химические коммуникации планктона

Эколог Егор Задерев о типах химических сигналов, миграциях зоопланктона и образовании покоящихся яиц

Text Bag of words

Что исследователи знают о химической коммуникации планктона в воде? Какими сигналами обменивается зоопланктон? Как размножается зоопланктон? Об этом рассказывает кандидат биологических наук Егор Задерев.

Планктон — это организмы, местоположение которых в водной толще в основном определяется течениями. То есть это что-то маленькое, то, что переносится течениями. Планктон делится на фитопланктон (это водоросли) и зоопланктон. Мы будем говорить про зоопланктон — это рачки. То, как водные объекты между собой коммуницируют с помощью химических сигналов, исследовано довольно плохо. В наземных экосистемах, мы знаем, есть феромоны, различные сигнальные системы, которые хорошо исследованы. Мы используем их для создания ловушек, например, для вредителей — феромонные ловушки. Вода — это среда, которая благоприятна для химической коммуникации.

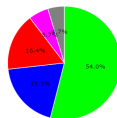
[post id="33793"]

Химические сигналы от хищников заставляют зоопланктон мигрировать. Это одно из самых масштабных на планете перемещений биомассы, которые ежегодно происходят в океанах, морях и озерах. Зоопланктон ночью поднимается к поверхности, а днем уходит на глубину. Днем свет сверху помогает хищникам ловить животных, и животные уходят на глубину, а ночью поднимаются к поверхности, чтобы есть. Было показано, что эти вертикальные миграции регулируются двумя факторами. Первый — это освещенность. Очевидно, что, если не будет света, не будет сигнала. А второй — это химия, которую выделяют хищники.

В 2006 и 2009 годах выходили хорошие обзоры по химическим коммуникациям. То есть а) это очень маленькие молекулы, и б) они работают в очень низких концентрациях. Это до сих пор удивляет и поражает, потому что сообщества зоопланктона и вообще планктона в водных экосистемах — это сотни видов водорослей, рачков, которые живут в озерах, в морях, взаимодействуют между собой. А между ними есть очень сложная, судя по тому, что мы получаем в лаборатории, и разветвленная сеть химических сигналов и коммуникаций, которые влияют на разные поведенческие, физиологические и продуктивные функции. И эта сложная цель, сеть взаимодействий до сих пор слабо исследована.

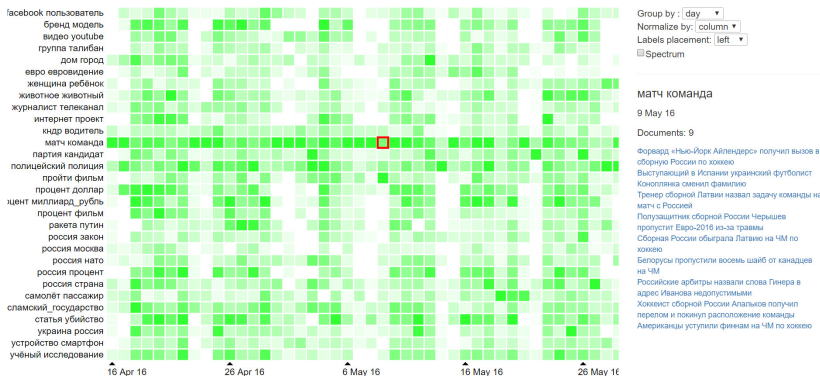
Dataset: postnauka
Time: Dec. 14, 2014, 3 p.m.
[View original](#)
index_id: 1866
text_id: 36719.txt
Terms count: 0
Unique terms count: 0
Model: [flat-20 ▾]
Highlighting: [Words ▾]

Topic distribution

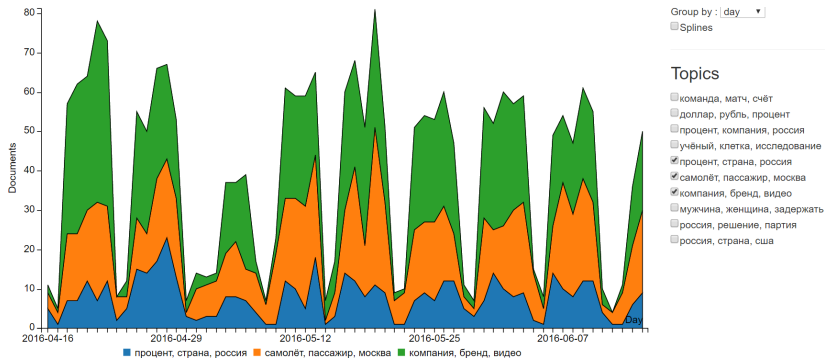


- земля, микроорганизм, вид
- вид, эволюция, ген
- материал, квантовый, структура
- город, социальный, пространство
- Other

VisARTM: Визуализация темпоральной модели



VisARTM: Визуализация темпоральной модели



VisARTM: Визуализация иерархической модели



Тексты научно-просветительского ресурса Postnauka.ru:
2976 документов, 43196 слов, 1799 тэгов

Belyy A.V., Seleznova M.S., Sholokhov A.K., Vorontsov K.V. Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue 2018.

Задача суммаризации (аннотирования, реферирования) текста

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

Полуавтоматическая — 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 фрагментов темы $p(d, s|t)$

Основные подходы к суммаризации:

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

H.P.Luhn. The automatic creation of literature abstracts. 1958.

Juan-Manuel Torres-Moreno. Automatic Text Summarization. 2014.

Основные этапы выборочной (extractive) суммаризации

- 1 Внутреннее представление текста
 - выявление тематики текста и отдельных предложений
 - вычисление признаков предложений
- 2 Оценивание полезности (ранжирование) предложений
- 3 Отбор предложений для реферата
 - оптимизация критериев информативности и 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.

Yogita Desai, Prakash Rokade. Multi Document Summarization: Approaches and Future Scope. 2015.

Mahak Gambhir, Vishal Gupta. Recent automatic text summarization techniques: a survey. 2016.

Покрывание терминологии и тематики документа

S_d — множество предложений документа d

$a \subset S_d$ — искомая суммаризация

Покрывание терминологии документа (lexicon coverage):

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

Покрывание тематики документа (topic coverage):

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

Избыточность суммаризации (redundancy):

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

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

Marina Litvak, Natalia Vanetik, Chunlei Liu, Lemin Xiao, Onur Savas.
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:

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

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

Сумма трёх критериев $WCov(a) + \tau_1 TCov(a) + \tau_2 Red(a)$:

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

Максимизация покрытия — это максимизация правдоподобия!

Можно добавить регуляризатор разреживания:

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

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
- 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 термина w в предложении s ;

n_s — длина предложения s .

Отбор предложений для суммаризации: $p(s|t) \rightarrow \max_{s \in S_d}$

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 термина 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:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{s \in S_d} p_{ws} \sum_{t \in T} \psi_{st} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

- Авторы утверждают, что модель переходит в обычную $p(w|d) = \sum_t \phi_{wt} \theta_{td}$, если предложение \equiv слово
- Это не так, т.к. предложения уникальны: $S_d \cap S_{d'} = \emptyset$
- 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:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & \begin{cases} p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ p_{stdw} \equiv p(s, t|d, w) = \operatorname{norm}_{s, t \in S_d \times T}(p_{ws} \psi_{st} \theta_{td}) \end{cases} \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \psi_{st} = \operatorname{norm}_{s \in S_d} \left(\sum_{w \in S_d} n_{dw} p_{stdw} + \psi_{st} \frac{\partial R}{\partial \psi_{st}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(\sum_{w \in d} n_{dw} p_{tdw} + \tau \sum_{w \in d} \sum_{s \in S_d} n_{dw} p_{stdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

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 :

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

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

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

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 длинее m

$JS(p(w|s), p(w|R))$ — 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.

Резюме

- Для удобства систематизации нужна визуализация
- Для визуализации нужны суммаризация и именование тем
- Качественная суммаризация — открытая проблема NLP
- Суммаризация темы — открытая проблема ТМ
- Тематические модели суммаризации развиты слабо
- ROUGE — семейство мер качества суммаризации, характеризуют далеко не все аспекты качества
- BLUE — аналогичные метрики, но precision-based

Задача автоматического именованя тем (topic labeling)

Требования к *названию темы* (topic label):

- релевантность названия теме
- интерпретируемость и грамматическая корректность
- непохожесть на названия похожих тем

Гипотеза 1: тройка топ-слов — плохое название.

Гипотеза 2: все названия уже придуманы, осталось их найти.

Подзадачи

- формирование названий-кандидатов l_1, \dots, l_m
- построение (обучение) функции релевантности $s(l, t)$
- выбор названия с учётом названий похожих тем

Qiaozhu Mei, 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, PyТез, Викисловарь, и др.

Функция релевантности (relevance score)

Релевантность нулевого порядка:

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

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

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

где C — релевантный теме контекст, в котором ожидается появление как слов темы t , так и названия ℓ целиком.

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

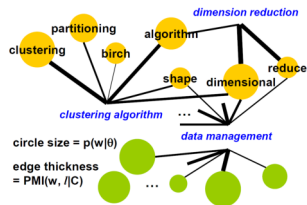
Выбор нескольких названий для темы

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, $s(\ell, t) \rightarrow \max$,
- максимальное не похоже на уже выбранные названия ℓ' :

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

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 $s'(l, t)$:

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

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

где параметр μ подбирается эмпирически.

Методика оценивания качества именованя тем:

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

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

Оценивание качества именованя тем

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

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 именованя 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

- *Automatic Topic Labeling* — очень узкое направление, всего 20–30 статей за 10 лет
- Важно для автоматизации создания приложений
- Близко к задаче суммаризации темы
- Для иерархических моделей добавляется требование *полноты*: названия дочерних тем должны адекватно описывать разделение родительской темы

Wanqiu Kou, Fang Li, T.Baldwin. Automatic Labelling of Topic Models using Word Vectors and Letter Trigram Vectors. 2015.

S.Bhatia, Jey Han Lau, T.Baldwin. Automatic Labelling of Topics with Neural Embeddings. COLING-2016

Xiaojun Wan, Tianming Wang. Automatic Labeling of Topic Models Using Text Summaries. 2016.

M.Allahyari, S.Pouriyeh, K.Kochut, H.R.Arabnia. A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling. 2017.