Вероятностные тематические модели
Лекция 2. Когерентность тем и модели парной сочетаемости слов

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)»

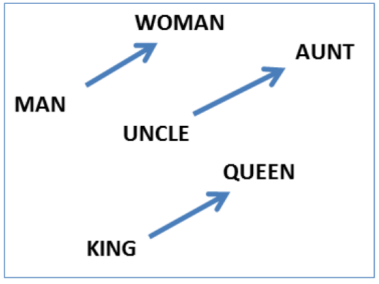
МФТИ.ФПМИ.ИС.ИАД • 2026-09-17

Напоминание. Эволюция подходов в анализе текстов

Анализ текстов 15 лет назад: пирамида NLP

- морфологический анализ, лемматизация, опечатки
- синтаксический анализ, выделение терминов, NER
- семантический анализ, выделение фактов, тем

Контекстно независимые эмбединги слов в вероятностных моделях языка на основе матричных разложений



- модели дистрибутивной семантики
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016]
- тематические модели LDA [Blei, 2003], ARTM [2014]

Контекстно зависимые нейросетевые эмбединги

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d}}\right) V$$

- рекуррентные нейронные сети: LSTM [1997]
- модели внимания и трансформеры: NMT [2015], BERT [2018], GPT-3 [2020], GPT-4 [2023]
- NEW! тематические модели внимания

Напоминание. Тематические модели «мешка термов»

Дано: коллекция текстовых документов D, словарь W;
n_{dw} — частота терма $w \in W$ в документе $d \in D$.

Найти: вероятностную языковую модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$
с параметрами $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}$$

С чем разберёмся сегодня

- ТМ строится ради получения интерпретируемых тем
  — как измерять интерпретируемость тем?
  — как её улучшать, оптимизировать?
  — как строить регуляризаторы интерпретируемости?
- Как уйти от гипотезы мешка слов?
  — смысл слова зависит от его окружения, контекста
  — как это формализовать в эмбедингах $p(t|w)$?
- Как уйти от проблемы коротких текстов?
  — когда длина документа n_d меньше размерности $p(t|d)$

Спойлер:

все эти проблемы решаются с помощью одной и той же идеи
— использования данных о парной сочетаемости слов

Содержание

1. **Когерентность как мера интерпретируемости**
   - Измерение интерпретируемости тем
   - От когерентности темы к когерентности текста
   - Задача калибровки текстовой когерентности
2. **Семантические векторные представления слов**
   - Дистрибутивная гипотеза
   - Языковые модели word2vec
   - Другие модели векторных представлений текста
3. **Тематические модели дистрибутивной семантики**
   - Тематическая модель битермов BitermTM
   - Тематическая модель сети слов WNTM
   - Регуляризаторы когерентности

Измерение интерпретируемости тем

Тема интерпретируемая, если по топовым словам темы эксперт может определить, о чём эта тема, и дать ей название.

- *Экспертные оценки*:
  — интерпретируемость темы по балльной шкале;
  — каждую тему оценивают несколько экспертов.
- *Метод интрузий* (intrusion):
  — в список топовых слов внедряется лишнее слово;
  — измеряется доля ошибок экспертов при его определении

Задача: найти внутренний критерий интерпретируемости, наиболее коррелирующий с экспертными оценками

Решение: *когерентность* (согласованность) тем (topic coherence)

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Эксперимент по поиску меры интерпретируемости

Измерялась ранговая корреляция Спирмена экспертных оценок с каждой из 15 мер интерпретируемости.

PMI — лучшая метрика.

Gold-standard — средняя корреляция Спирмена между оценками разных экспертов.

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Вывод: когерентность близка к «золотому стандарту».

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Определение когерентности

Когерентность (согласованность) темы t по k топовым словам:

$$\text{coh}_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{PMI}(w_i, w_j)$$

где w_i — i-е слово в порядке убывания ϕ_{wt},

$\text{PMI}(u, v) = \ln \frac{P_{uv}}{P_u P_v}$ — *поточечная взаимная информация* (pointwise mutual information), где:

P_{uv} — доля документов, в которых слова u, v хотя бы один раз встречаются рядом (в одном предложении или в окне 10 слов),

P_u — доля документов, в которых u встретился хотя бы 1 раз,

P_{uv}, P_u можно вычислять по другой коллекции (Википедии).

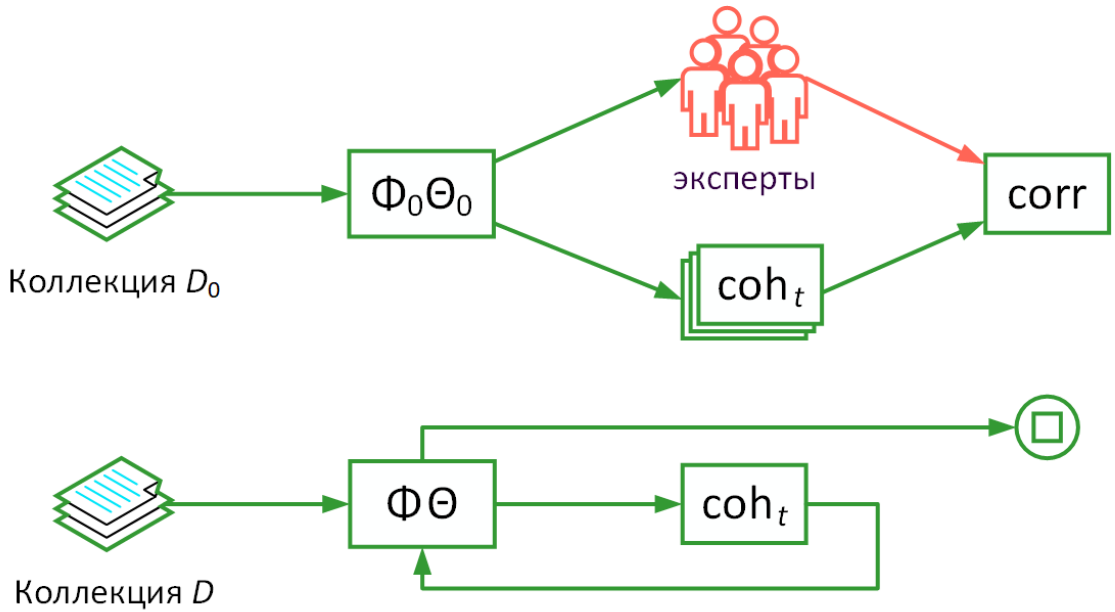
Когерентность модели = средняя когерентность всех тем.

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Схема калибровочного эксперимента Ньюмана

1. строим по коллекции D_0 тематическую модель $\Phi_0\Theta_0$
2. эксперты оценивают темы (рейтингами или интрузиями)
3. ищем критерий, коррелирующий с оценками экспертов

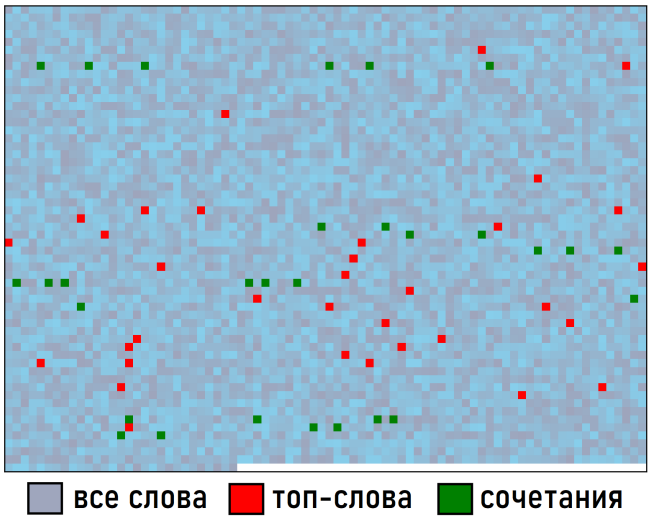
На новой коллекции D используем откалиброванный критерий (когерентность тем coh_t) для оценивания и выбора моделей $\Phi\Theta$



Недостаток когерентности

Обычно берут $k = 10..20$ топовых (самых частых) слов темы, но они занимают лишь 1–2% текста совместно по всем темам, а пары с большим N_{uv} образуются из топовых слов ещё реже!

Более 99% текста игнорируется оценкой когерентности модели, и «золотой стандарт» Ньюмана страдает тем же недостатком!



Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

first top words of *topic 3: физика* with top 10 in bold: **частица**, **электрон**, **кварк**, **атом**, **энергия**, **вселенная**, **фотон**, **физика**, **физик**, **эксперимент**, масса, теория, свет, симметрия, протон, эйнштейн, нейтрино, вещество, квантовый, ускоритель, детектор, волна, эффект, свойство, спин, гравитация, материя, адрон, поль, частота

V.A.Alekseev, V.G.Bulatov, K.V.Vorontsov. Intra-text coherence as a measure of topic models interpretability // Dialogue, 2018.

Обобщение — семейство средневзвешенных когерентностей

Средневзвешенная когерентность темы:

$$\mathrm{coh}_t = \frac{\sum_{u,v} \mathrm{rel}_t(u,v)\,\mathrm{coh}(u,v)}{\sum_{u,v} \mathrm{rel}_t(u,v)} \equiv \operatorname*{avg}_{u,v}\big(\mathrm{rel}_t(u,v), \mathrm{coh}(u,v)\big)$$

$\mathrm{coh}(u,v)$ — *сочетаемость* пары слов $(u,v) \in W^2$ в текстах
$\mathrm{rel}_t(u,v)$ — *релевантность* слов u и v теме t, в частности
$\mathrm{rel}_t(u,v) = \big[\phi_{ut}, \phi_{vt} > \mathrm{top}_k \phi_{wt}\big]$ — *когерентность Ньюмана*

Возможные модификации:

- сделать rel ненулевым для большего числа пар u, v:
  $\mathrm{rel}_t(u,v) = \sqrt{\phi_{ut}\phi_{vt}}$ или $[\phi_{ut}\phi_{vt} \geqslant \varepsilon]$
- поэкспериментировать с выбором coh:
  $\mathrm{coh}(u,v) = (\mathrm{PMI} - \delta)_+$ или $\mu\big(\frac{P_{uv}}{P_u P_v}\big)$ или $\frac{P_{uv} - P_u P_v}{\sqrt{P_{uv}}}$

Проблема: большой объём вычислений по всем парам слов

Текстовая когерентность (intra-text coherence)

Средневзвешенная когерентность темы:

$$\mathrm{coh}_t = \frac{\sum_{u,v} \mathrm{rel}_t(u,v)\,\mathrm{coh}(u,v)}{\sum_{u,v} \mathrm{rel}_t(u,v)} = \operatorname*{avg}_{u,v}\big(\mathrm{rel}_t(u,v), \mathrm{coh}(u,v)\big),$$

где суммирование по парам слов (u,v) в общих *контекстах*, например, в одном предложении или на расстоянии ± 10 слов.

Вычисление: за один проход по коллекции для каждой темы t аккумулируются суммы в числителе и в знаменателе.

Возможные модификации:

- $\mathrm{rel}_t(u,v) = \sqrt{p(t|d,u)\,p(t|d,v)}$ после E-шага
- перейти в coh от документных частот N_* к терм-парным m_*:
  $\mathrm{coh}(u,v) = \frac{m}{m_u m_v}$, $m_w = \sum_{u,v} [u=w] + [v=w]$, $m = \sum_w m_w$,

---

Василий Алексеев. Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций. МФТИ, 2018.

Что такое терм-парные частоты (term-pair frequency)

Пример: словарь $W = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$, ширина окна $h = 5$

текст: «$\boxed{\mathtt{A\ B\ C\ B\ A}}\ \mathtt{D\ A\ C\ C\ B\ D\ A}$», длина текста $n = 12$

число пар термов во всех окнах: $m = (n - h + 1)(h - 1) = 32$

частоты m_{uv} пар термов

A A	2
A B	3
A C	5
A D	2
B A	3
B B	1
B C	2
B D	2
C A	3
C B	2
C C	1
C D	2
D A	1
D B	1
D C	2
D D	0
	32

частоты n_w термов

A	4
B	3
C	3
D	2
	12

частоты m_w термов-в-парах: левые, правые, двусторонние

A∗	12	∗A	9	A	21
B∗	8	∗B	7	B	15
C∗	8	∗C	10	C	18
D∗	4	∗D	6	D	10
	32		32		64

Два варианта расчёта отношения вероятностей пары (u, v) к вероятностям термов-в-паре (на примере $u = \mathtt{A}$, $v = \mathtt{C}$) — для пар упорядоченных ($uv \neq vu$) и неупорядоченных ($uv = vu$):

$$\frac{P_{uv}}{P_{u*} \cdot P_{*v}} = \frac{P(\mathtt{AC})}{P(\mathtt{A}*) \cdot P(*\mathtt{C})} = \frac{\frac{5}{32}}{\frac{12}{32} \cdot \frac{10}{32}} = 1.33$$

$$\frac{P_{uv,vu}}{P_{u*,*u} \cdot P_{v*,*v}} = \frac{P(\mathtt{AC} \cup \mathtt{CA})}{P(\mathtt{A}) \cdot P(\mathtt{C})} = \frac{\frac{5+3}{64}}{\frac{12+9}{64} \cdot \frac{8+10}{64}} = 1.35$$

Как проверить адекватность текстовой когерентности

. . . если «золотой стандарт» Ньюмана столь же неадекватен?

Идея: размечать слова в текстах, а не слова в темах

- эксперты выделяют в текстах *тематические цепочки слов*
- тексты — научно-популярные, междисциплинарные

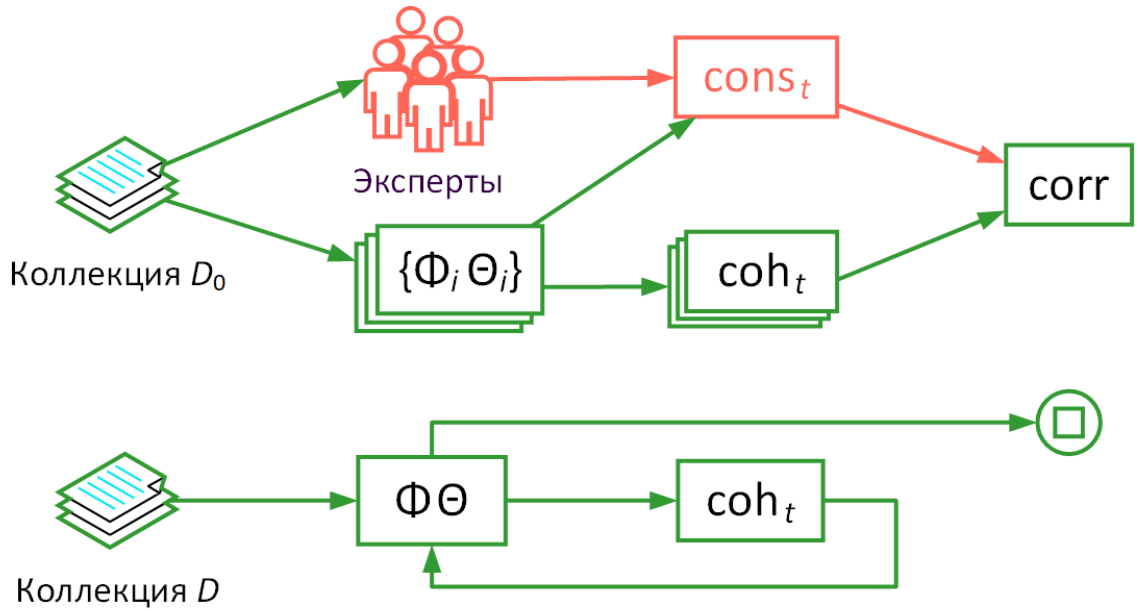
Пример: разметка текста на три тематические цепочки

транспорт	психология	общенаучная лексика	общеупотребительная лексика

В исследованиях мы действительно можем находить корреляции между стилем вождения и особенностями личности. Например, склонные к экстраверсии водители могут больше отвлекаться на внешние факторы и стимулы внешней среды и в этом отношении представляют большую опасность. В свою очередь, люди, которым требуется большее количество психических ресурсов, для того чтобы справиться с тревогой, будут вести себя осторожнее в условиях трафика. Вместе с тем есть и обратная сторона: та же характеристика интроверсии за счет высокого уровня тревожности приводит к чрезмерной осторожности. Для таких водителей характерен крадущийся тип вождения, что будет влиять на общее тревожное поведение всех участников трафика.

Схема калибровки для текстовой когерентности

1. выбираем из коллекции D_0 фрагменты для разметки
2. эксперты размечают тематические цепочки во фрагментах
3. строим по коллекции D_0 много моделей $\{\Phi_i\Theta_i\}$ с разнообразными темами, как хорошими, так и плохими
4. ищем критерий, коррелирующий с согласованностью cons_t между темами t и размеченными тематическими цепочками



Оценки согласованности тем с размеченными цепочками

Тема интерпретируема, если она согласована с разметкой,
т. е. различает в тексте цепочки, которые различили эксперты.

Тематика цепочки C, по формуле полной вероятности:

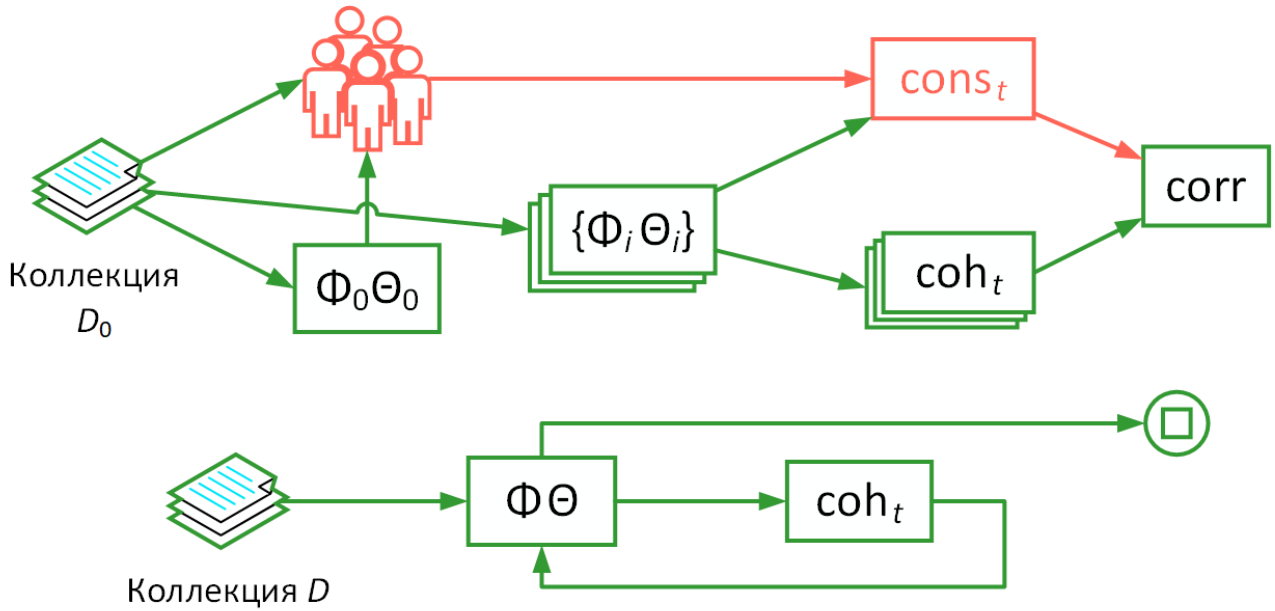
$$p(t|C) = \sum_{w \in C} p(t|w)p(w|C) = \frac{1}{|C|} \sum_{w \in C} p(t|w)$$

Вход: Φ_0, множество цепочек $\{C_{dk} : d \in D_0, k \in K_d\}$;
Выход: оценки согласованности тем cons_t, $t \in T$;
$S_t := 0$; $N_t := 0$ для всех $t \in T$;
для всех $d \in D_0$: **для всех** размеченных цепочек $k \in K_d$:
  $t := \arg\max_{t \in T} p(t|C_{dk})$ — главная тема цепочки C_{dk};
  $\Delta_t := p(t|C_{dk}) - \max_{j \neq k} p(t|C_{dj})$ — различимость темы t;
  $S_t \mathrel{+}= \Delta_t$; $N_t \mathrel{+}= 1$;
$\text{cons}_t := S_t / N_t$ для всех $t \in T$;

Число тем $|T|$ оптимизируется по максимуму среднего cons_t

Возможно ли (почти) обойтись без экспертов?

1. строим по коллекции D_0 тематическую модель $\Phi_0\Theta_0$
2. находим монотематичные фразы (предложения, абзацы)
3. эксперты проверяют, что темы и фразы интерпретируемые
4. убирая из них стоп-слова, получаем размеченные цепочки C_{dk}
5. строим по коллекции D_0 много моделей $\{\Phi_i\Theta_i\}$
6. ищем критерий, коррелирующий с согласованностью cons_t



Текстовая когерентность: преимущества и проблемы

Преимущества: теперь

- когерентность оценивается по полному массиву текста
- разметка отделена от построения тематической модели
- участие экспертов можно минимизировать или исключить

Открытые проблемы:

- провести калибровку для текстовой когерентности: подобрать наилучшее сочетание эвристик: rel, coh, окно
- как полностью отказаться от экспертных оценок? [2025]

Michael Röder et al. Exploring the space of topic coherence measures. 2015.
Alexander Hoyle et al. Is automated topic model evaluation broken? The incoherence of coherence. 2021.
Dominik Stammbach et al. Revisiting automated topic model evaluation with large language models. 2023.
Hamed Rahimi et al. Contextualized Topic Coherence Metrics. 2024.
Zhiyin Tan, Jennifer D'Souza. Bridging the evaluation gap. Leveraging large language models for topic model evaluation. 2025.

Дистрибутивная гипотеза и виды семантической близости слов

Смысл слова есть множество всех контекстов его употребления

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
сочетаемость слов в одном контексте.

здание–строитель, кран–вода, функция–точка

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте.

здание–дом, кран–смеситель, функция–отображение

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.Turney, P.Pantel. From frequency to meaning: vector space models of semantics. 2010.

Формализация дистрибутивной гипотезы в программе word2vec

Дано: частоты n_{wu} пар слов w, u в контекстном окне $\pm k$ слов
Найти: векторные представления слов x_w и предсказывающих слов-из-контекста y_u в вероятностной языковой модели

$$p(w|u) = \operatorname*{SoftMax}_{w \in W} \langle x_w, y_u \rangle = \operatorname*{norm}_{w \in W} \big(\exp \langle x_w, y_u \rangle \big)$$

Критерий 1: максимум правдоподобия языковой модели

$$\sum_{w,u \in W} n_{wu} \ln p(w|u) \;\to\; \max_{x_w, y_u}$$

Критерий 2: бинарная классификация пар слов близки/далеки

$$\sum_{w,u \in W} \left(n_{wu} \ln p(w,u) + \sum_{i=1}^{n_{wu}} \ln \big(1 - p(w,u_i)\big) \right) \;\to\; \max_{x_w, y_u}$$

$p(w,u) = \sigma \langle x_w, y_u \rangle$ — модель вероятности встретить w, u рядом
u_i сэмплируется из $p(w)^{3/4}$ (skip-gram negative sampling, SGNS)

T.Mikolov et al. Efficient estimation of word representations in vector space, 2013.

Связь word2vec с матричными разложениями

d — размерность векторов слов x_w и слов-из-контекста y_u
$X = (x_w)_{W \times d}$ — матрица векторов предсказываемых слов
$Y = (y_u)_{W \times d}$ — матрица векторов слов-из-контекста

SGNS строит матричное разложение $P \approx XY^{\mathsf{T}}$ матрицы $W \times W$
Shifted PMI (Point-wise Mutual Information):

$$P_{wu} = \ln \frac{n_{wu} n}{n_w n_u} - \ln k,$$

n_{wu} — частота пары слов w, u в контекстном окне $\pm k$ слов,
n_w, n_u — число пар с участием слова w и u соответственно,
n — число всех пар слов в коллекции.

В качестве эвристики используют также Shifted Positive PMI:

$$P^+_{wu} = \left(\ln \frac{n_{wu} n}{n_w n_u} - \ln k \right)_+ .$$

O.Levy, Y.Goldberg. Neural word embedding as implicit matrix factorization, 2014.

Модель векторных представлений FastText

Идея: векторное представление слова w определяется как сумма векторов всех его буквенных n-грамм $G(w)$:

$$u_w = \sum_{g \in G(w)} u_g$$

В Skip-gram вместо векторов слов u_w обучаются векторы u_g

Пример: $G(\text{дармолюб}) = \{<\text{да}, \text{арм}, \text{рмо}, \text{мол}, \text{олю}, \text{люб}, \text{юб}>\}$

Преимущества:

- Это решает проблемы новых слов и слов с опечатками
- Подходит для обработки текстов социальных медиа
- Словарь 2- и 3-грамм обычно меньше словаря W
- Существует много предобученных моделей

Bojanowski et al. Enriching word vectors with subword information. 2016.

Модели векторных представлений текстов и графов

word2vec: *эмбединги* (векторные представления) слов
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных *n*-грамм
`https://github.com/facebookresearch/fastText`

node2vec: эмбединги вершин графа
A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

BERT: контекстно-зависимые эмбединги от Google AI Language
J.Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

GPT-3: эмбединги, предобученные по 570Gb текстов от OpenAI
T.B.Brown et al. Language Models are Few-Shot Learners. 2020.

Сходство и отличия языковых моделей WE и PTM

Сходство: то и другое — это

- матричные разложения
- векторные представления слов $x_w \in \mathbb{R}^d$ и $p(t|w) \in \mathbb{R}^T$

Отличия WE от PTM:

- данные — частоты пар слов n_{wu}, а не слов-в-документах n_{dw}
- безусловная оптимизация SGD вместо EM-алгоритма
- векторы x_w — плотные; векторы $p(t|w)$ — разреженные

⊕ WE успешны в решении задач аналогии и близости слов

⊕ WE предобучаются по большим текстовым коллекциям

⊖ компоненты векторов не интерпретируемы

⊖ не ясно, почему XY^{T}, а не XX^{T} (обычно Y игнорируют)

Идея: обучать PTM тоже по данным о сочетании пар слов

Проблема коротких текстов в тематическом моделировании

Проблема *коротких текстов* (short text): $n_d \ll |T|$

- Twitter и другие микроблоги
- социальные медиа
- заголовки статей и новостных сообщений

Тривиальные подходы (у каждого свой набор недостатков):

- считать каждое сообщение отдельным документом
- разреживать $p(t|d)$ вплоть до единственной темы
- объединить сообщения по автору/времени/региону/и т. п.
- объединить посты с комментариями
- дополнить коллекцию длинными текстами (Википедия и др.)

Идея: вместо документов использовать данные о битермах

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

Битермы: модель сочетаемости слов в коротких текстах

Битерм — пара слов $(u, v) \in W^2$, встречающихся *рядом*, т. е. в одном коротком сообщении / предложении / окне $\pm h$ слов.

Тематическая модель битермов (Biterm Topic Model):

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

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

Критерий максимума логарифма правдоподобия:

$$\sum_{u,v} n_{uv} \ln \sum_t \phi_{ut}\phi_{vt}\pi_t \;\to\; \max_{\Phi,\pi},$$

$$\phi_{vt} \geqslant 0; \quad \textstyle\sum_v \phi_{vt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

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

Необходимые условия точки максимума правдоподобия

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{u,v} n_{uv} \ln \sum_{t} \phi_{ut}\phi_{vt}\pi_t + R(\Phi,\pi) \;\to\; \max_{\Phi,\pi},$$

где n_{uv} — частота битерма (u,v) в документах коллекции.

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tuv} \equiv p(t|u,v) = \operatorname*{norm}_{t\in T}\left(\phi_{ut}\phi_{vt}\pi_t\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{vt} = \operatorname*{norm}_{v\in W}\left(n_{vt} + \phi_{vt}\dfrac{\partial R}{\partial \phi_{vt}}\right), \quad n_{vt} = \sum\limits_{u\in W} n_{uv}p_{tuv} \\ \pi_t = \operatorname*{norm}_{t\in T}\left(n_t + \pi_t\dfrac{\partial R}{\partial \pi_t}\right), \quad n_t = \sum\limits_{u,v\in W} n_{uv}p_{tuv} \end{cases}$$

И данные, и модель симметричны: $n_{uv} = n_{vu}$, $p(u,v) = p(v,u)$.

Битермы как регуляризатор для обычной ΦΘ-модели

1. Регуляризатор битермов для матрицы Φ:

$$R(\Phi) = \tau \sum_{u,v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt} \to \max.$$

Подставляем в формулу M-шага, получаем сглаживание:

$$\phi_{wt} = \underset{w}{\operatorname{norm}} \Big(n_{wt} + \tau \sum_{u \in W} n_{uw} p_{tuw} \Big); \quad p_{tuw} = \underset{t \in T}{\operatorname{norm}} \big(n_t \phi_{wt} \phi_{ut} \big).$$

Это эквивалентно обработке *псевдо-документов* d_u, где каждый d_u объединяет все контексты слова u, причём $\theta_{tu} \propto n_t \phi_{ut}$;
n_{uw} — число вхождений слова w в псевдо-документ d_u.

2. Регуляризатор разреживания Θ для коротких текстов:

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

Модель сети слов WNTM для коротких текстов

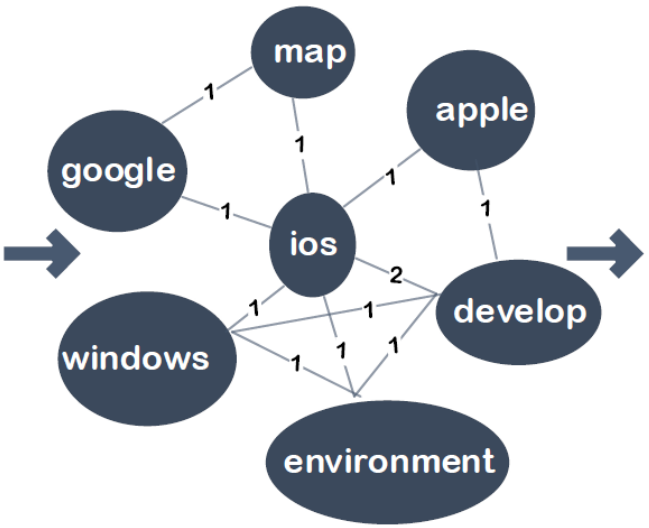
Идея: моделировать не документы, а связи между словами.

d_u — псевдо-документ, объединение всех контекстов слова u.

n_{uw} — число вхождений слова w в псевдо-документ d_u.

Контекст — короткое сообщение / 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 (Word Network) и WTM (Word Topic Model)

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

$$p(w|d_u) = \sum_{t \in T} p(w|t)p(t|d_u) = \sum_{t \in T} \phi_{wt}\theta_{tu},$$

где d_u — псевдо-документ слова u, для краткости $\theta_{tu} = \theta_{td_u}$

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} \phi_{wt}\theta_{tu} \to \max_{\Phi,\Theta},$$

где n_{uw} — частота сочетания пары слов (w, u).

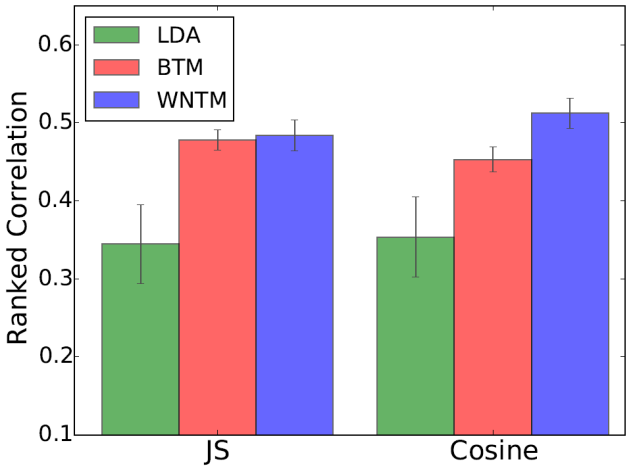
Отличие модели битермов: там $\Theta = \mathrm{diag}(\pi_1, \ldots, \pi_t)\Phi^{\mathrm{T}}$, здесь данные и модель НЕ симметричны: $n_{uw} \neq n_{wu}$, $p(u|d_w) \neq p(w|d_u)$

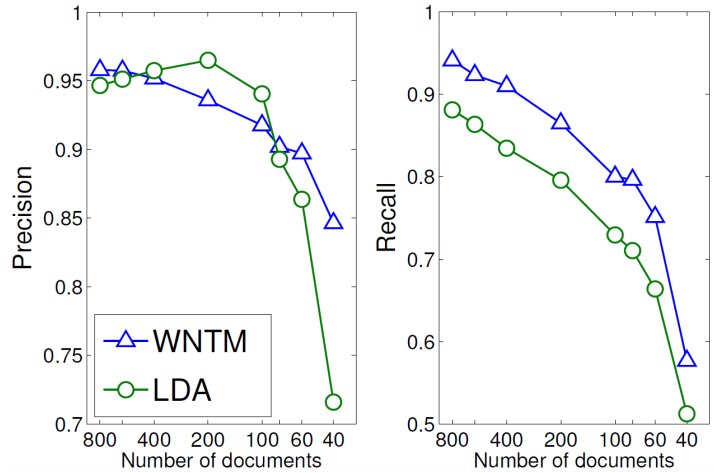
Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

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

Результаты оценивания модели WNTM

- Когерентность на коротких текстах лучше, чем у LDA и BitermTM; на длинных текстах преимуществ нет.
- *Слева:* оценивание семантической близости слов по $p(t|w)$, корреляция с 10-балльными экспертными оценками.
- *Справа:* полнота и точность распознавания новой темы в зависимости от числа документов.





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

WN-ARTM на задачах семантической аналогии слов

Сравниваются два подхода к обучению векторов слов:

- **WN-ARTM**: интерпретируемые разреженные координаты
- **word2vec**: интерпретируемые векторные операции

Операция	Результат WN-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

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Сравнение word2vec и WN-ARTM по интерпретируемости тем

SGNS (word2vec) — нет интерпретируемости:

- avg hearth soc protector decomposition whip stochastic sewer splinter accessory howie thief thermodynamic boltzmann equilibrium kingship unconscious
- rainy miocene snowy horner cfb triassic eleventh amadeus dams tenth mesozoic fourteenth thirteenth ninth diaries bight demographics seventh almanac eocene
- gnis usda bloomberg usgs regulator nhk gerd magnetism capacitor fed classifies capacitance stadt bipolar multilateral trpod kunst reciprocal smiths potassium

WN-ARTM — есть интерпретируемость:

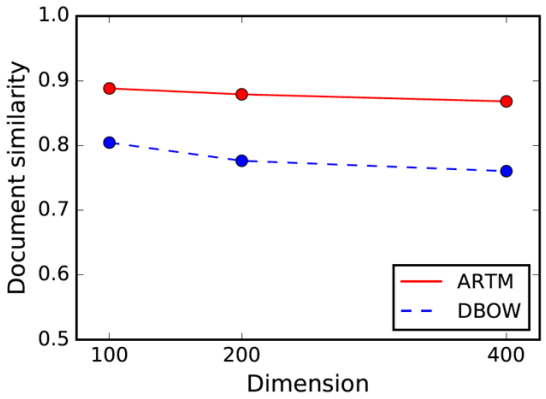
- scottish scotland edinburgh glasgow mps oxford educated cambridge college aberdeen dundee royal uk scots fellows fife corpus kingdom thistle eton angus
- game games video gameplay multiplayer puzzle mario nintendo player gaming pok playable mortal super kombat adventure rpg ds puzzles online smash zelda
- election party elected elections parliament assembly seats members minister legislative electoral liberal council representatives parliamentary democratic

---

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

WN-ARTM на задачах семантической близости документов

ArXiv triplets dataset [Dai et. al, 2015]: 20K троек статей:
⟨ статья A, схожая статья B, непохожая статья C ⟩



- обучение по 1M текстов статей ArXiv
- тестирование на триплетах ArXiv
- Конкурент: DBOW paragraph2vec [Dai et. al, 2015]

WN-ARTM превосходит модель DBOW (distributed bag-of-words) на той же коллекции, на которой DBOW оценивали его авторы

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Регуляризаторы когерентности

Слова, часто встречающиеся рядом, должны иметь общие темы:

$$R(\Phi) = \tau \sum_{t \in T} p(t) \sum_{u \in W} \sum_{v \in W} \frac{n_{uv}}{n_v} \frac{n_{vt}}{n_t} \ln \phi_{ut}$$

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u \in W} \sum_{v \in W} P_{uv} \big[\mathrm{PMI}(u,v) > 0\big] \phi_{ut} \phi_{vt}$$

Тематическая модель битермов (Biterm Topic Model):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} \phi_{ut} \phi_{vt} p(t)$$

Тематическая модель сети слов (Word Network Topic Model):

$$R(\Phi, \Theta) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} \phi_{wt} \theta_{td_u}$$

Mimno D. et al. Optimizing semantic coherence in topic models. EMNLP 2011.
Newman D. et al. Improving topic coherence with regularized topic models. NIPS 2011.

Резюме

1. word2vec [2013] стал революцией в анализе текстов:
   — это векторная модель парной сочетаемости слов
   — вектор кодирует смысл слова по данным о его контекстах
   — в соответствии с гипотезой дистрибутивной семантики
2. В тематическом моделировании эта идея тоже успешна:
   — парная сочетаемость: когерентность: интерпретируемость
   — мешок слов $\rightarrow$ мешок пар слов $\rightarrow$ связный текст
3. Как улучшать интерпретируемость тем:
   — строить тематическую модель по парной сочетаемости
   — использовать регуляризаторы когерентности
   — строить словари n-грамм (терминов, словосочетаний)
   — но это не всё — см. следующие лекции
4. Когерентность (Ньюмана) имеет проблемы
   — нужны новые эксперименты с текстовой когерентностью

Задания по курсу

Задача-минимум: научиться решать задачи анализа текстов с использованием тематического моделирования

Задача-максимум: получить новый научный результат

виды деятельности	оценка
теоретическая задача	X
теоретическая задача*	$2X$
теоретическая задача**	$3X$
решение прикладной задачи	$10X$
обзор по последним PTM/NTM	$10X$
участие в проекте	$20X$
работа над открытой проблемой	$25X$

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

Итоговая оценка: $\min\left(5, \lfloor \text{score}/20 \rfloor\right)$ по 5-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 ККТ 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.

Задания к лекции 2

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайды 27–28)

Прикладная исследовательская задача:
оценивание (внутри)текстовой когерентности

- реализуйте вычисление средневзвешенной когерентности
- предложите способ разметки с помощью LLM
- подберите наилучшее сочетание эвристик rel и coh в калибровочном эксперименте без экспертной разметки
- какие регуляризаторы улучшают текстовую когерентность?
- воспроизводимо ли это улучшение на разных коллекциях?