• Вероятностные языковые модели •
Лекция 2.
Модели сочетаемости слов

Константин Вячеславович Воронцов
k.vorontsov@iai.msu.ru

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

ВМК МГУ • 10 марта 2026

Содержание

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.

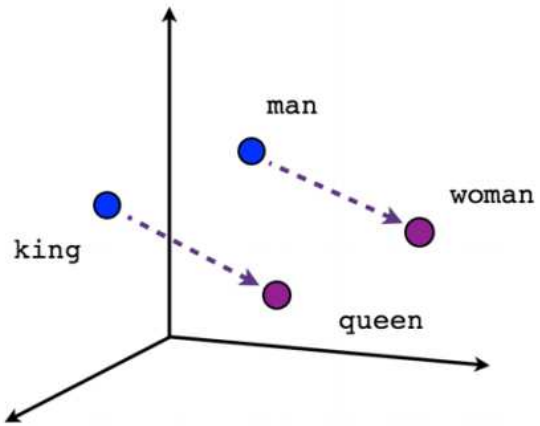
здание–дом, кран–смеситель, функция–отображение

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.

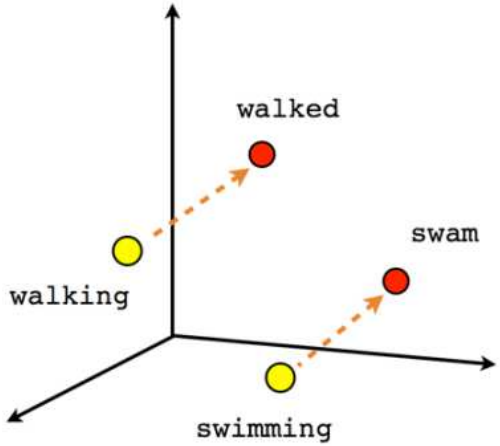
Задача семантического векторного представления слов

Задача: по наблюдаемой синтагматической близости слов построить *векторные представления слов* (word embedding, WE) $x_w \in \mathbb{R}^d$, $w \in W$, отражающие их парадигматическую близость, т.е. близкие по смыслу слова должны иметь близкие векторы.

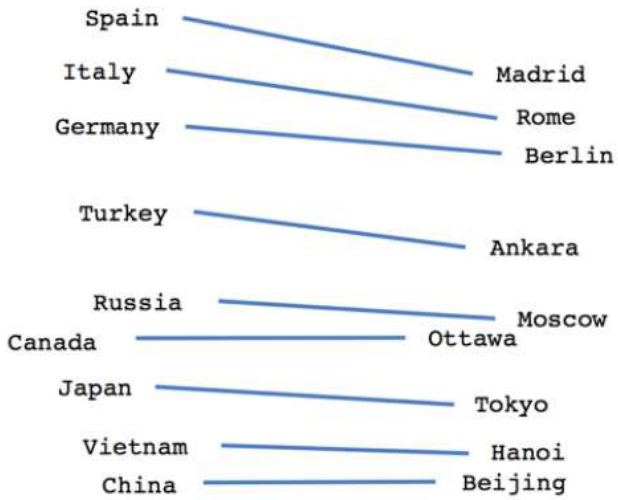
Способ проверки — задача семантической аналогии слов: по трём словам угадать четвёртое.



Male-Female



Verb tense



Country-Capital

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

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

Критерий: максимум правдоподобия для предсказания слов

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

или для классификации пар слов на 2 класса 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): $n_d < \dim p(t|d) = |T|$

- 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

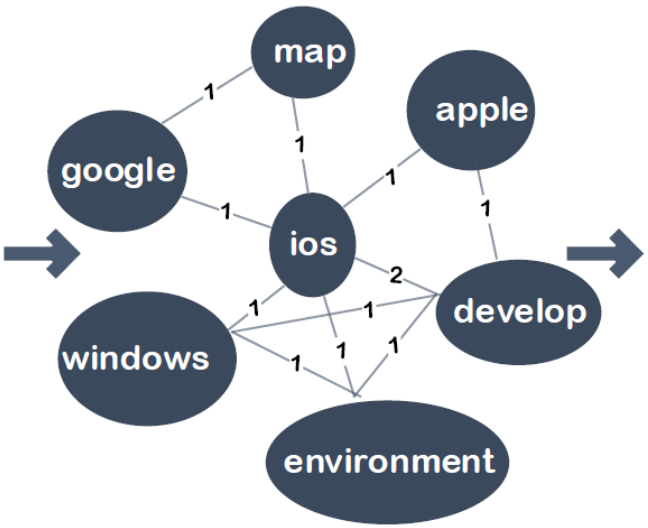
Идея: моделировать не документы, а связи между словами.

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.

Когерентность — мера интерпретируемости тем

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),

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,

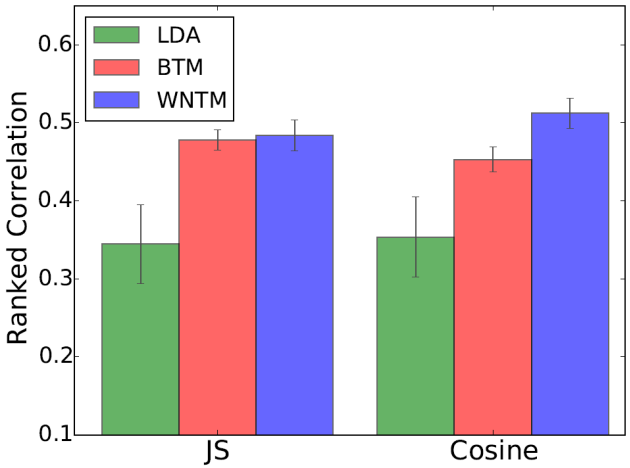
P_{uv}, P_u можно вычислять по другой коллекции (Википедии).

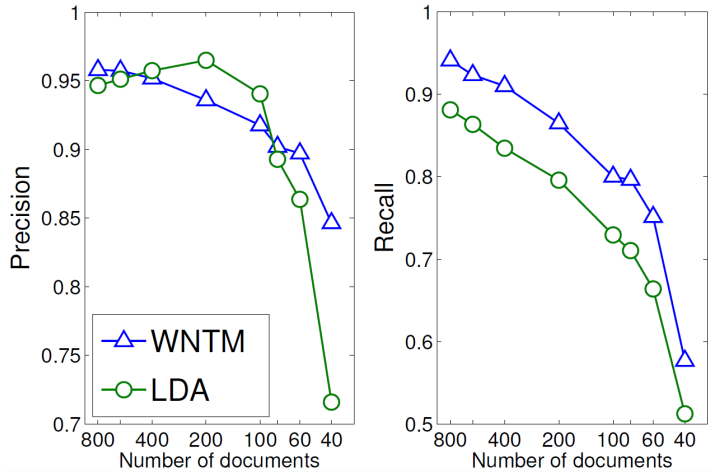
Экспериментально установлено: когерентность темы сильно коррелирует с экспертной оценкой интерпретируемости темы.

Newman D. et al. Automatic evaluation of topic coherence. HLT 2010.

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:

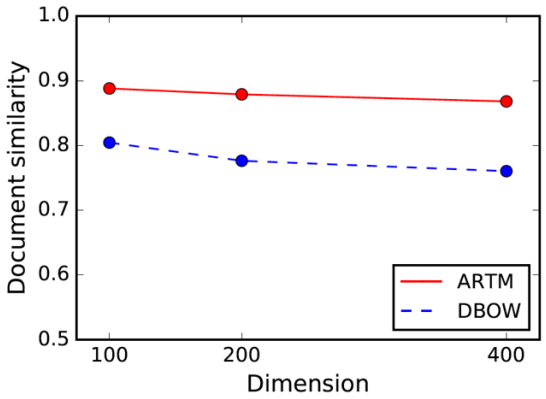
- scottish scotland edinburgh glasgow mps oxford educated cambridge college aberdeen dundee royal uk scots fellows fife corpus kingdom thistle eton angus
- game games video gameplay multiplayer puzzle mario nintendo player gaming pok playable mortal super kombat adventure rpg ds puzzles online smash 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.

Несколько терминов из лингвистики

Единица языка, в зависимости от уровня членения текста — фонема, морфема, слово, словосочетание, фраза, предложение

Сочетаемость (co-occurrence) — свойство языковых единиц сочетаться в речи, образуя единицы более высокого уровня

Виды сочетаемости:

- *контактная* — языковые единицы идут подряд
- *дистантная* — не обязательно подряд (например, битермы)

Виды контактной сочетаемости:

- *n-грамма* — последовательность из n единиц языка
- *коллокация* — *n-грамма* слов, встречающаяся в корпусе гораздо чаще, чем при их чисто случайном соединении
- *словосочетание* — *n-грамма* слов, связанных по смыслу и грамматически, обозначающая единое понятие

Пример 1. n-граммы улучшают интерпретируемость тем

Коллекция 20Conf заголовков научных статей DBLP, тема «Information Retrieval»

Terms	*Phrases*
search	information retrieval
web	social networks
retrieval	web search
information	search engine
based	support vector machine
model	information extraction
document	web page
query	question answering
text	text classification
social	collaborative filtering
user	topic model

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora // VLDB, 2015.

Пример 2. Биграммы улучшают интерпретируемость тем

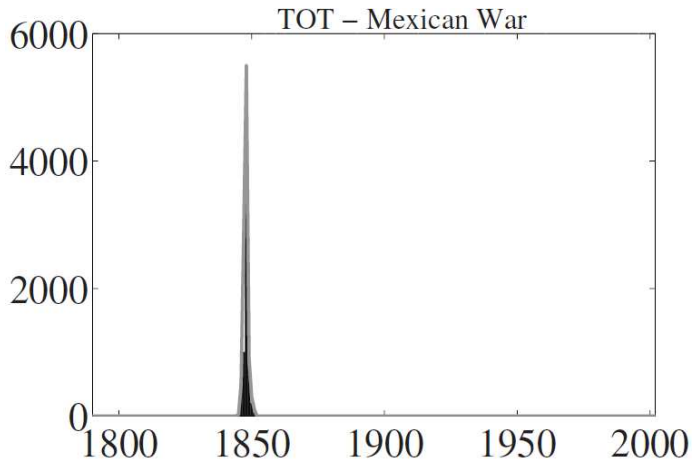
Коллекция 1000 статей конференций ММРО, ИОИ на русском

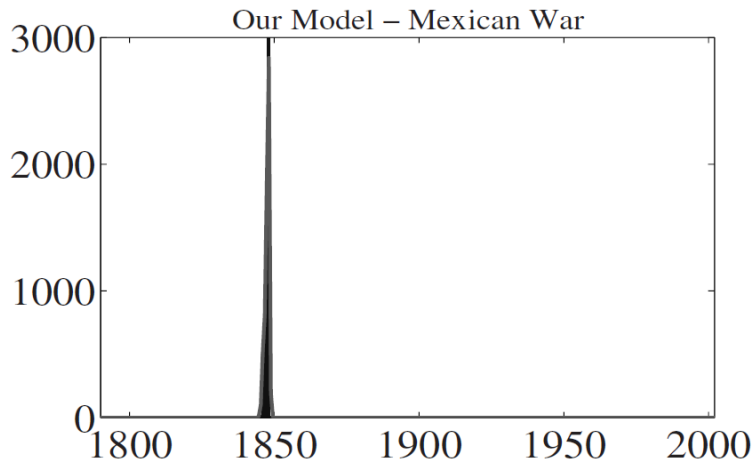
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

Пример 3. Совмещение темпоральной и n-граммной модели

По коллекции выступлений президентов США





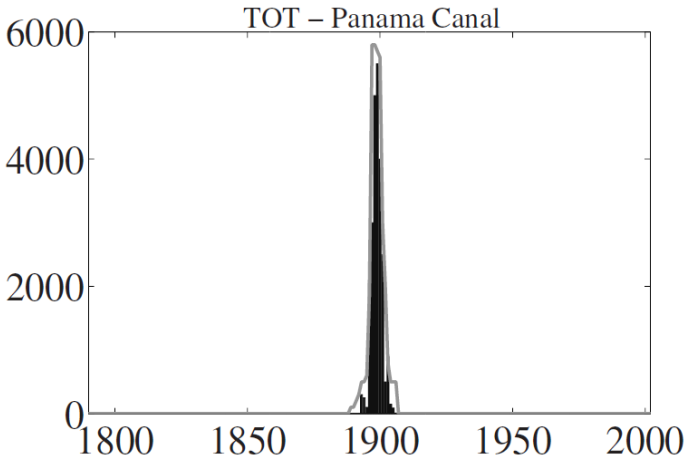
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

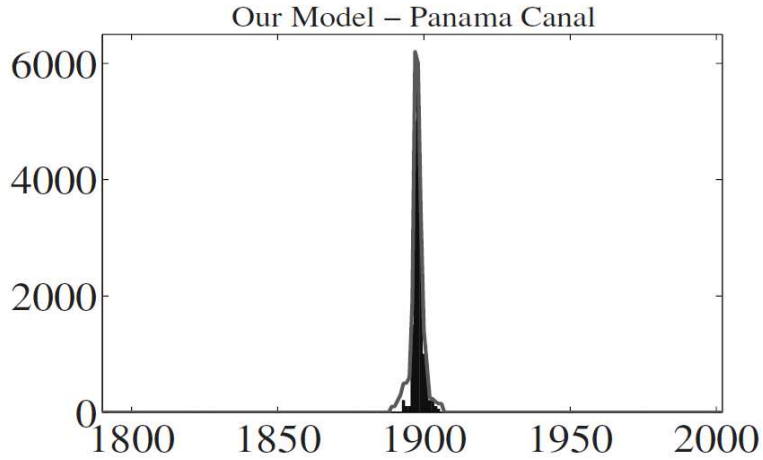
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Пример 3. Совмещение темпоральной и *n*-граммной модели

По коллекции выступлений президентов США





1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Задача автоматического выделения терминов

Термин — фраза (n-грамма) со следующим набором свойств:

1. *высокая частотность* (frequency):
   много раз встречается в коллекции;
2. *контактная сочетаемость слов* (collocation):
   состоит из слов, неслучайно часто встречающихся вместе;
3. *полнота* (completeness):
   является максимальной по включению цепочкой слов;
4. *синтаксическая связность* (syntactic connectedness):
   является грамматически корректным словосочетанием;
5. *тематичность* (topicality):
   часто встречается в узком подмножестве тем.

Сумма технологий для ATE (Authomatic Term Extraction):
TopMine 1 2 3 + UDPipe 4 + BigARTM 5

Алгоритм TopMine: определения и основные идеи

- $C(a_1, \ldots, a_k)$ — хэш-таблица частот k-грамм, $a_i \in W$, $C(w) = n_w$ для всех униграмм $w \in W$: $n_w \geqslant \varepsilon_1$
- ε_k — пороговое значение частоты частых k-грамм
- $A_{d,k}$ — множество позиций i в документе d, с которых начинаются все частые k-граммы:

$$C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k$$

- Свойство антимонотонности:

$$C(a_1, \ldots, a_k) \geqslant C(a_1, \ldots, a_k, a_{k+1})$$

- Основной шаг алгоритма: для всех $i = 1, \ldots, n_d$

  если $(i \in A_{d,k})$ **и** $(i+1 \in A_{d,k})$ **то** $++C(w_{d,i}, \ldots, w_{d,i+k})$

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora. VLDB, 2015.

Алгоритм TopMine: быстрый поиск всех частых k-грамм

Вход: коллекция D, пороги ε_k;
Выход: хэш-таблица частот $C(a_1, \ldots, a_k)$, $k = 1, \ldots, k_{\max}$;

$C(w) := n_w$ для всех $w \in W$;
$A_{d,0} := \{1, \ldots, n_d\}$;
для $k := 1, \ldots, k_{\max}$
  для всех $d \in D$
    $A_{d,k} := \left\{ i \in A_{d,k-1} \mid C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k \right\}$;
    для всех $i \in A_{d,k}$
      если $i+1 \in A_{d,k}$ **то** $++C(w_{d,i}, \ldots, w_{d,i+k})$;
  оставить только частые k-граммы: $C(a_1, \ldots, a_k) \geqslant \varepsilon_k$;

Преимущество алгоритма: линейная память и скорость.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora. VLDB, 2015.

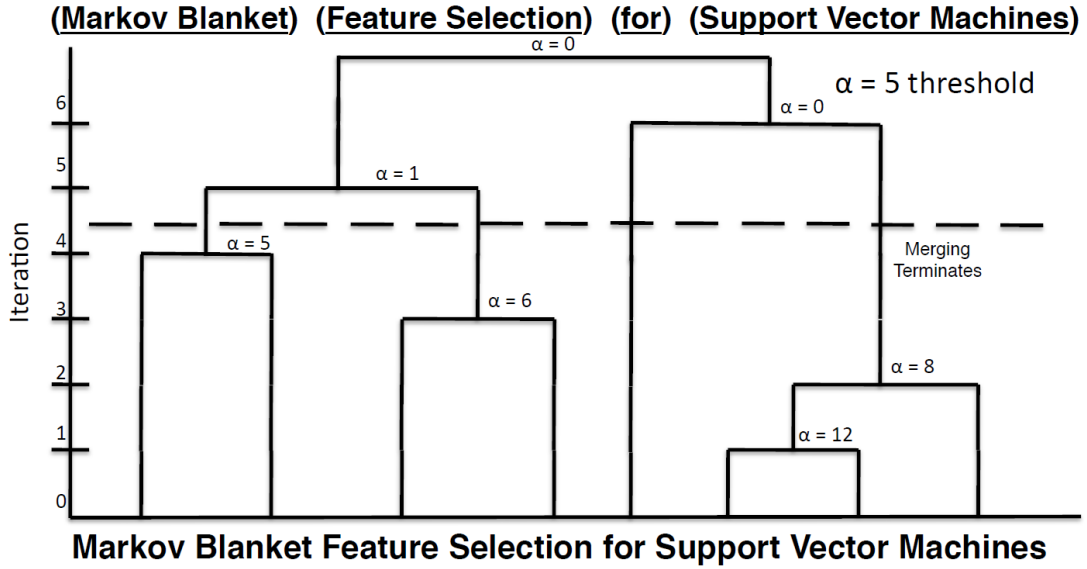
Алгоритм TopMine: отбор фраз по частоте и полноте

Итеративное слияние фраз с понижением значимости α.

p_u — оценка вероятности встретить фразу u

p_{uv} — оценка вероятности встретить фразу uv

Критерии: $\text{SignificanceScore} = \dfrac{p_{uv} - p_u p_v}{\sqrt{p_{uv}}}$ или $\text{PMI} = \log \dfrac{p_{uv}}{p_u p_v}$



Синтаксические анализаторы (UDPipe, SyntaxNet)

Вход: список предложений

Выход, для каждого слова в каждом предложении:

- id (порядковый номер слова в предложении)
- id родительского слова (0 для корня)
- исходное слово
- нормальная форма
- часть речи: `NOUN`, `VERB`, `ADJ`, `ADV`, ...
- член предложения: `nsubj`, `dobj`, `conj`, `cc`, `nmod`, ...

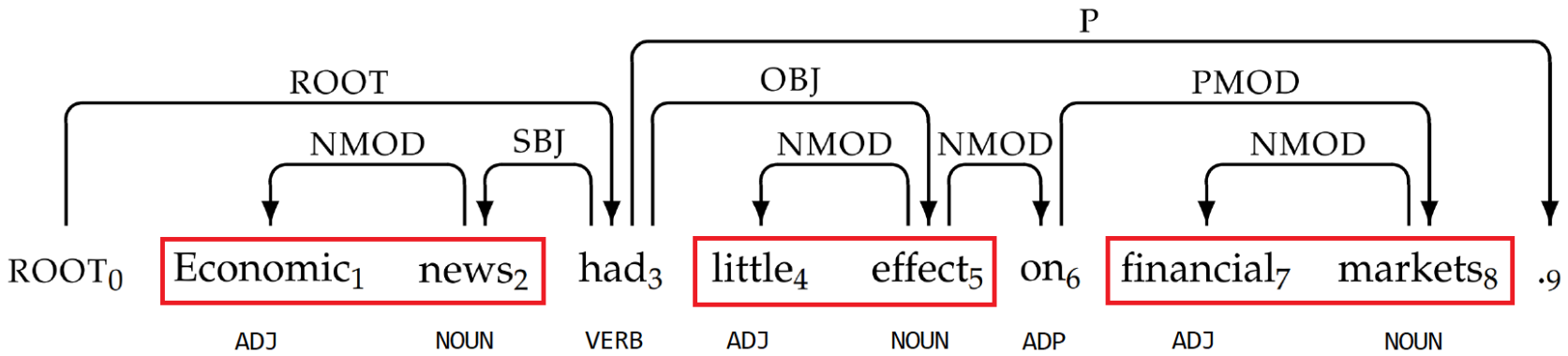
UDPipe (Universal Dependencies), 60 языков, включая русский

Google SyntaxNet — предобученная нейросеть поверх TensorFlow, поддерживает 40 языков, включая русский.

D.Andor, C.Alberti, D.dWeiss, A.Severyn, A.Presta, K.Ganchev, S.Petrov, M.Collins. Globally Normalized Transition-Based Neural Networks. 2016.

Использование дерева зависимостей для отбора терминов

Пример дерева зависимостей:



Варианты стратегий отбора терминов-кандидатов:

- брать все поддеревья
- брать все именные группы (корень — NOUN)
- не брать CONJ, SCONJ, DET, AUX, INTJ, PART, PUNCT, SYM

Announcing SyntaxNet: the world's most accurate parser goes open source.
https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

Денис Кирьянов. Изучаем синтаксические парсеры для русского языка. 2018.
https://habr.com/ru/company/sberbank/blog/418701

Критерии тематичности фраз

Насколько далеко $p(t|w) = \phi_{wt}\frac{n_t}{n_w}$ от равномерного $p_0(t) = \frac{1}{|T|}$.

Дивергенция Кульбака-Лейблера:

$$\mathrm{KL}(w) = \mathrm{KL}(p_0\|p) = \sum_{t\in T}\frac{1}{|T|}\ln\frac{\frac{1}{|T|}}{p(t|w)} \to \max$$

Дивергенция Йенсена-Шеннона (метрика, не имеет проблем с нулевыми вероятностями), где $\bar{p}(t|w) = \frac{1}{2}(p(t|w) + \frac{1}{|T|})$:

$$\mathrm{JS}(w) = \frac{1}{2}\mathrm{KL}(p_0\|\bar{p}) + \frac{1}{2}\mathrm{KL}(p\|\bar{p}) \to \max$$

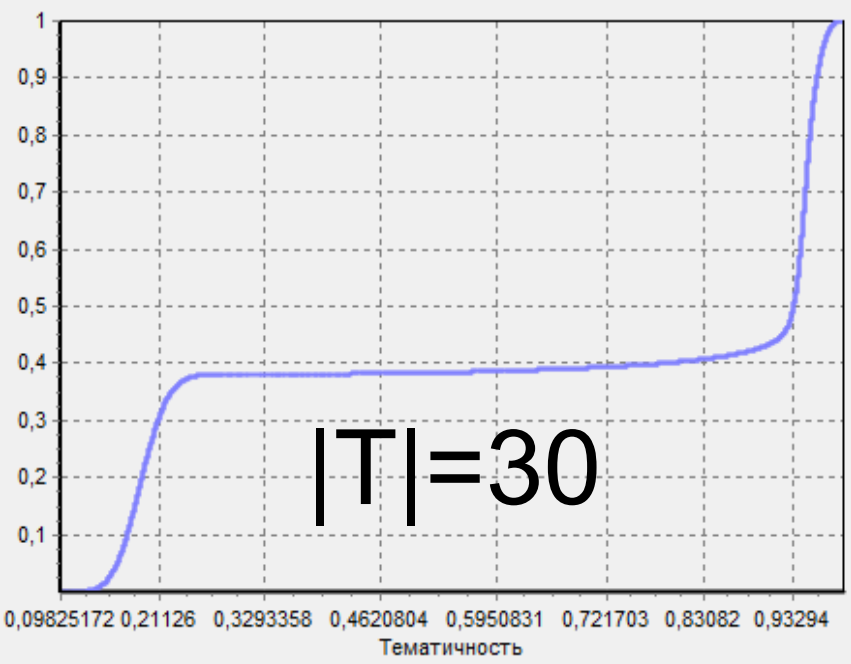
Нормированная сумма степенных функций, $\gamma > 1$:

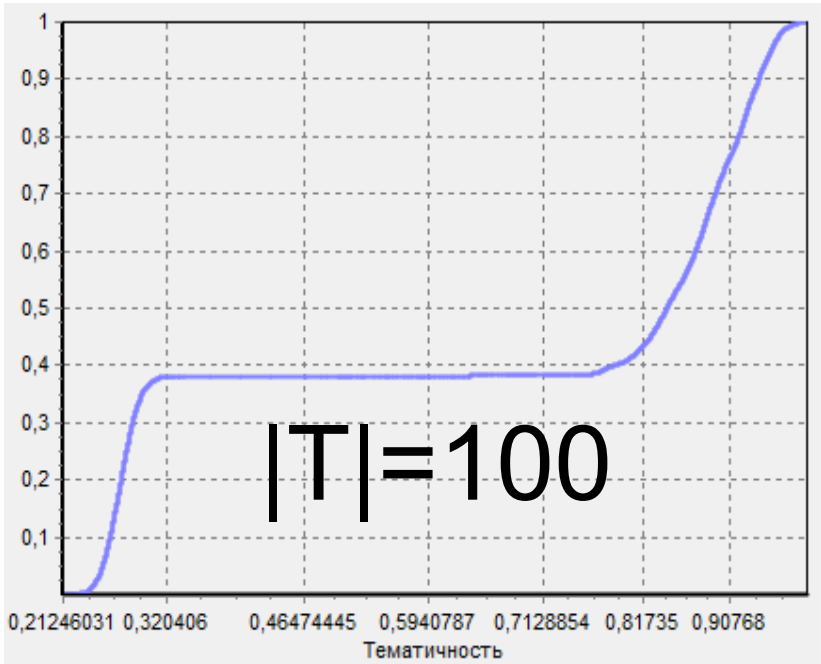
$$\text{Тематичность}(w) = |T|^{\gamma-1}\sum_{t\in T}p(t|w)^{\gamma} \to \max$$

Фразы чётко разделяются на тематичные и нетематичные

Коллекция SyntagRus $|D| = 600$, словарь $|W| = 46\,000$ фраз, тематические модели LDA с числом тем $T = 30,\ 100$.

Эмпирическое распределение фраз по тематичности:





Пограничный слой между тематичными и нетематичными фразами очень узкий $\approx \frac{200}{46\,000}$ и слабо зависит от числа тем.

Основной эксперимент ATE: SyntaxNet + TopMine + BigARTM

- Коллекция $|D| = 3200$ аннотаций статей NIPS (Neural Information Processing Systems), $n = 500\,000$ слов
- Ручная разметка небольшого *случайного* подмножества (2000 n-грамм) на термины / не-термины
- Train : Test = 1000 : 1000
- 7 статистических признаков из TopMine
- 2 синтаксических признака из SyntaxNet
- 3 тематических признака из BigARTM, 30 тем
- две модели классификации:
  логистическая регрессия, градиентный бустинг

Владимир Полушин. Тематические модели для ранжирования рекомендаций текстового контента. Бакалаврская диссертация, ВМК МГУ, 2017.

Сравнение методов автоматического отбора терминов

Найти *как можно больше терминов* — полнота важнее точности

Группа признаков			Линейная модель			Градиентный бустинг		
Синт	Стат	Тем	AUC	Точность	Полнота	AUC	Точность	Полнота
+			0.83	0.20	0.91	0.83	0.20	0.91
	+		0.71	0.09	0.94	0.73	0.11	0.90
		+	0.92	0.32	1.00	0.95	0.32	1.00
+	+		0.88	0.22	0.91	0.88	0.24	0.91
+		+	0.91	0.36	0.91	0.95	0.34	0.99
	+	+	0.93	0.29	0.94	0.98	0.34	1.00
+	+	+	**0.95**	**0.38**	**0.91**	**0.97**	**0.41**	**0.99**

Стат < Син < Син+Стат < Тем < Стат+Тем, Син+Тем < Стат+Син+Тем

- Тематические признаки существенно повышают качество
- Синтаксические признаки можно не использовать

Конкурс RuTermEval конференции «Диалог»

Классификации научных терминов по 3 классам:

- **specific term:** термины, специфичные доменно (для конкретной предметной области) и лексически
- **common term:** термины, специфичные только доменно (могут быть известны и употребляться неспециалистами)
- **nomen:** номенклатурные наименования доменно специфичных объектов, материальных объектов данного домена (программы, базы данных, наборы данных, языки, корпуса, словари и т.д.)

Данные: 850 аннотаций по домену компьютерной лингвистики: 65K токенов, 18K разметок терминов (уникальных 6534), неразмеченные тексты того же домена и двух других доменов

---

https://dialogue-conf.org/evaluation/rutermeval-2024/

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. Эксперимент с выделением терминов надо повторить
   на репрезентативных и качественно размеченных данных

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
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: $\min\big(5, \lfloor \text{score}/20 \rfloor\big)$ по 5-балльной шкале.

Задания к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Биграммная модель коллекции: $p(w|v) = \xi_{wv}$,
где v — слово, идущее в тексте перед w.
Найти параметры модели ξ_{wv}.

2. Биграммная модель документов: $p(w|v,d) = \xi_{dwv}$.
Найти параметры модели ξ_{dwv}.

Подсказка: применить условия ККТ или основную лемму.

3. Творческое задание (возможны разные решения)
Предложите модель, разделяющую 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.

4. Пользуясь основной леммой, докажите, что регуляризатор битермов эквивалентен добавлению псевдодокументов d_u в исходную коллекцию (см. слайд 13)

Прикладная исследовательская задача:
автоматическое выделение научных терминов (ATE)

- Дано:
  коллекция размеченных текстов конкурса ruTermEval;
  неразмеченная коллекция текстов той же тематики
- Найти:
  метод ATE на основе комбинирования ARTM и TopMine;
  обоснование, что синтаксические и др. методы не нужны;
  зависимость качества ATE от объёма неразмеченных данных
- Критерий:
  качество ATE (Prec, Rec, F1) на размеченных данных

Примеры датасетов для практических заданий по курсу

1. Открытые датасеты (английский): 20NG, NIPS, KOS
2. Ранжированные результаты поиска научных статей (по данным eLibrary, arXiv, PubMed)
3. Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
4. Техноблоги: Хабр (русский), TechCrunch (английский)
5. Данные социальных сетей: VK, Twitter, Telegram,...
6. Статьи по Complexity Sciences (хронокарта науки)

- 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,
  — реализация данного сценария как модуля в среде Orange

- «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: помочь в понимании предметной области

Открытые проблемы тематического моделирования

1. Тематические модели внимания последовательного текста
2. Проблема несбалансированности тем в коллекции
3. Измерение интерпретируемости тем (когерентность)
4. Обеспечение 100%-й интерпретируемости тем
5. Автоматическое именование и суммаризация тем
6. Калибровка моделей тематической фильтрации
7. Согласование тем с предобученными эбмедингами LLM
8. Статистические оценки состоятельности тем
9. Обнаружение новых тем или трендов в потоке текстов
10. Обеспечение устойчивости и полноты множества тем
11. Автоматический подбор гиперпараметров, AutoML
12. Гиперграфовые тематические модели для RecSys