Технология интерактивной визуализации тематических моделей

Федоряка Дмитрий

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель
профессор РАН, д.ф.-м.н.
К. В. Воронцов

22 июня 2017

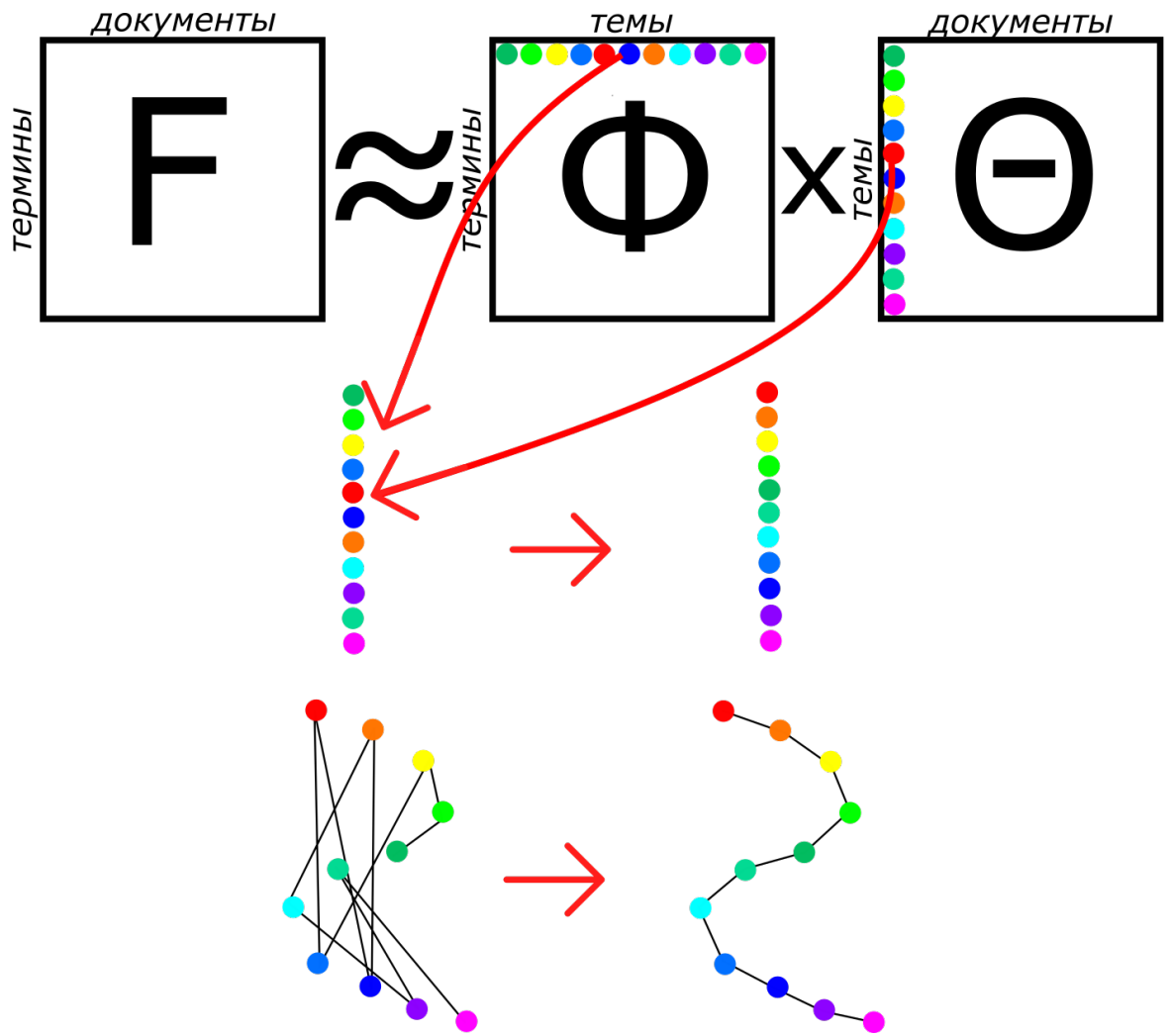
Тематическое моделирование

D — коллекция документов,
W — словарь терминов,
T — множество тем.

$F_{wd} = p(w|d)$ — частоты;
$\varphi_{wt} = p(w|t)$;
$\theta_{td} = p(t|d)$

$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) \Leftrightarrow F = \Phi\Theta$

Тематический спектр



Задача: упорядочить темы так, чтобы близкие по смыслу темы оказались близкими в списке.

Постановка задачи

Введём *функцию расстояния* между темами

$$\rho : T \times T \to [0, +\infty)$$

Матрица расстояний:

$$R[i, j] = \rho(t_i, t_j)$$

Тематический спектр — такая перестановка тем, для которой минимальна сумма расстояний между соседними темами:

$$\pi^* = \arg \min_{\pi \in S_{|T|}} \sum_{i=1}^{|T|-1} \rho(t_{\pi_i}, t_{\pi_{i+1}})$$

Функции расстояния

Евклидово расстояние

$$\rho_E(t,s) = \sqrt{\sum_{w \in W} \left(\varphi_{wt} - \varphi_{ws}\right)^2}$$

Манхэттенское расстояние

$$\rho_M(t,s) = \sum_{w \in W} \left|\varphi_{wt} - \varphi_{ws}\right|$$

Косинусное расстояние

$$\rho_C(t,s) = 1 - \frac{1}{\|t\|\|s\|} \sum_{w \in W} \varphi_{wt}\varphi_{ws}; \quad \|t\| = \sqrt{\sum_{w \in W} \varphi_{wt}^2}$$

Функции расстояния

Расстояние Хеллингера

$$\rho_H(t,s) = \sqrt{\frac{1}{2}\sum_{w\in W}\left(\sqrt{\varphi_{wt}} - \sqrt{\varphi_{ws}}\right)^2}$$

Расстояние Йенсена-Шеннона

$$\rho_{JS}(t,s) = H\left(\frac{\Phi_t + \Phi_s}{2}\right) - \frac{1}{2}\left(H(\Phi_t) + H(\Phi_s)\right); H(u) = -\sum_i u_i \ln u_i$$

Расстояние Жаккара

$$\rho_J(t,s) = 1 - \frac{\left|\left\{w \in W | \varphi_{wt} > \frac{1}{|W|} \wedge \varphi_{ws} > \frac{1}{|W|}\right\}\right|}{\left|\left\{w \in W | \varphi_{wt} > \frac{1}{|W|} \vee \varphi_{ws} > \frac{1}{|W|}\right\}\right|}$$

Решение задачи спектра

Алгоритмы:

- Агломеративная кластеризация;
- Многомерное шкалирование (MDS, t-SNE);
- Симуляция отжига;
- Алгоритм LKH для задачи коммивояжёра [1].

[1] Helsgaun, K. An effective implementation of the Lin-Kernighan traveling salesman heuristic. // European Journal of Operational Research. 2000

Пример спектра (postnauka)

1. остров, земля, период, там, территория, океан, где, более, вид, найти, вулкан, находиться, южный
2. растение, япония, раса, при, более, чем, например, исследование, вид, страна, население
3. вид, эволюция, самец, мозг, самка, животное, отбор, ген, более, птица, наш, между, чтобы, чем, друг
4. мозг, нейрон, при, заболевание, наш, пациент, состояние, система, болезнь, сон, исследование
5. клетка, музей, стволовой, ткань, организм, чтобы, опухоль, система, использовать, технология
6. клетка, ген, днк, организм, молекула, геном, белок, белка, бактерия, система, процесс, жизнь
7. система, материал, задача, структура, метод, компьютер, дать, при, химический, область, химия
8. квантовый, свет, волна, атом, информация, фотон, сигнал, использовать, два, при, частота, состояние
9. частица, энергия, кварк, взаимодействие, магнитный, электрон, масса, физика, бозон, протон, модель
10. звезда, галактика, земля, планета, вселенная, дыра, чёрный, объект, солнце, масса, наш, система
11. теория, пространство, вселенная, закон, физика, математический, уравнение, число, два, мир, система
12. наш, сеть, информация, дать, объект, культура, задача, например, образ, память, слово, разный
13. язык, слово, русский, например, говорить, словарь, речь, разный, языковой, текст, два, лингвист
14. наука, учёный, научный, потому, чтобы, лекция, хороший, университет, сейчас, наш, заниматься
15. экономический, экономика, страна, чтобы, более, рынок, компания, цена, решение, деньга, работа, чем
16. страна, война, государство, политический, россия, советский, власть, политика, германия, стать
17. ребёнок, женщина, мужчина, жизнь, культура, общество, себя, семья, социальный, советский, женский
18. город, пространство, социальный, городской, общество, место, культурный, жизнь, более, современный
19. исследование, социальный, поведение, группа, решение, and, the, теория, проблема, наука
20. социальный, социология, мир, теория, объект, социологический, действие, событие, социолог, наука
21. политический, философия, идея, наука, свобода, понятие, революция, история, философ, век, себя
22. право, власть, закон, король, век, римский, бог, себя, церковь, правовой, политический, суд, два
23. век, история, русский, исторический, имя, традиция, христианский, культура, историк, текст, уже
24. себя, искусство, литература, говорить, потому, мир, сам, миф, жизнь, слово, текст, роман, век
25. книга, фильм, автор, кино, pcourse, num, читатель, посвятить, тема, история, исследование, работа

Пример спектра (lenta)

1. спортсмен, допинг, олимпиада, рио, де, россия, проба, жанейро, wada, олимпийский_игра, соревнование
2. команда, матч, счёт, клуб, победа, чемпионат, турнир, минута, футболист, встреча, летний, футбол
3. евро, евровидение, страна, россия, конкурс, франция, болельщик, англия, украина, футбол, певец
4. пройти, мероприятие, россия, акция, фестиваль, москва, фильм, участник, картина, театр, музей
5. фильм, сериал, продукт, актёр, компания, продукция, процент, россия, книга, товар, картина, сезон
6. россия, москва, турист, процент, россиянин, страна, отель, рейс, путешественник, город, тысяча
7. процент, доллар, рубль, нефть, цена, россия, баррель, страна, уровень, вырасти, рынок, рост
8. компания, миллиард_рубль, процент, миллиард_доллар, россия, сумма, миллион_доллар, банк, банка
9. закон, законопроект, документ, реклама, использование, деятельность, поправка, внести, организация
10. россия, страна, керченский_пролив, российский, боинг, работа, чайка, ряд, гражданин, аэропорт
11. партия, кандидат, журналист, праймериза, выбор, единый_россия, госдума, выборы
12. россия, украина, крым, решение, киев, депутат, вопрос, отношение, страна, мнение, право, москва
13. россия, страна, турция, сша, ес, евросоюз, москва, санкция, отношение, украина, вопрос, государство
14. россия, сирия, исламский_государство, сша, нато, иго, запретить, террорист, страна, боевик
15. ракета, путин, россия, запуск, глава_государство, союз, спутник, президент
16. учёный, клетка, исследование, исследователь, ген, университет, оказаться, процент, помощь, организм
17. земля, животное, учёный, животный, тысяча, звезда, планета, обнаружить, кошка, территория, жизнь
18. самолёт, километр, машина, борт, пассажир, вертолёт, погибнуть, лайнер, пилот, час, район, яхта
19. полицейский, полиция, мужчина, задержать, автомобиль, улица, москва, пострадать, life
20. статья, убийство, задержать, суд, отношение, ук_рф, подозревать, следствие, обвинять, трамп, часть
21. ребёнок, женщина, мужчина, летний, дом, сын, семья, мальчик, жена, полиция, дочь, школа, врач
22. видео, youtube, ролик, фото, фотография, канал, снимка, auto, instagram, девушка, страница, группа
23. facebook, пользователь, интернет, страница, twitter, пост, написать, соцсеть, вконтакте, аккаунт
24. устройство, смартфон, компания, мотоциклист, игра, байкер, видео, миллион_доллар, робот, молодая
25. бренд, модель, компания, обувь, основать, одежда, релиз, коллекция, редакция, часы, поступить

Оценивание качества спектра

Целевой функционал (сумма расстояний между соседями)

$$NDS(\pi) = \sum_{i=1}^{N-1} R[\pi_i, \pi_{i+1}]; \quad N = |T|$$

Средний ранг соседа

$$\text{rank}(v|u) = \left|\left\{w \in \overline{1,N} \middle| R[w,u] < R[v,u]\right\}\right|$$

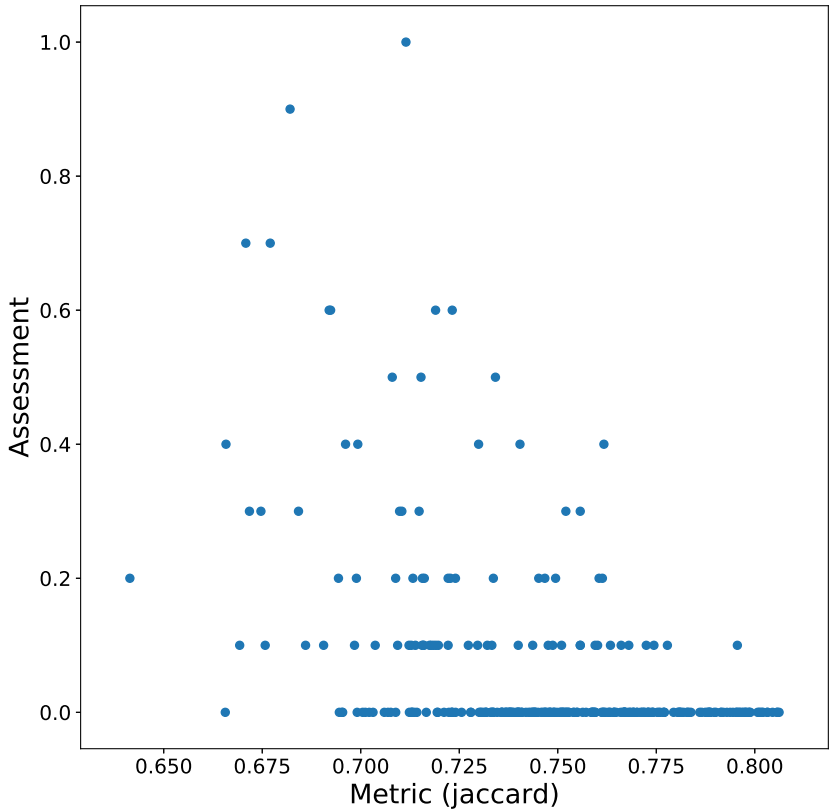
$$MNR(\pi) = \frac{1}{2N-2} \sum_{i=1}^{N-1} \left(\text{rank}(\pi_{i-1}|\pi_i) + \text{rank}(\pi_i|\pi_{i-1})\right)$$

Кривая расстояний

$$\text{DDC}(d) = \frac{1}{N-d} \sum_{i=1}^{N-d} R[i, i+d]$$

Оценивание близости тем с помощью асессоров

- Показать тему (каждую K раз);
- Попросить выбрать из остальных тем несколько близких по смыслу;
- Матрица оценок: $C_{ij} = \frac{\nu_{ij}+\nu_{ji}}{2K}$, где ν_{ij} — сколько раз тема i была указана, как близкая к j.



Асессорские меры качества

Корреляция

$$\mathrm{AMC}(\pi) = \frac{\sum_{i<j}(R_{ij} - \overline{R})(C_{ij} - \overline{C})}{\sqrt{\sum_{i<j}(R_{ij} - \overline{R})^2}\sqrt{\sum_{i<j}(C_{ij} - \overline{C})^2}}$$

Штраф за отдаление

$$\mathrm{ADP}(\pi) = \sum_{i<j} C_{ij}(|\pi_i^{-1} - \pi_j^{-1}| - 1)$$

Средняя несхожесть соседей

$$\mathrm{AMND} = 1 - \frac{1}{N-1}\sum_{i=1}^{N-1} C[\pi_i, \pi_{i+1}]$$

Асессорские меры качества

Доля несхожих соседей

$$\mathrm{ADNP} = \frac{1}{N-1} \sum_{i=1}^{N-1} \Big[C[\pi_i, \pi_{i+1}] = 0 \Big]$$

Кривая оценка-расстояние

$$\mathrm{ADC}(d) = \frac{1}{N-d} \sum_{i=1}^{N-d} C[i, i+d]$$

Эксперименты

- Коллекции
  - postnauka — `postnauka.ru`, 2012-2016, $|D| = 3446$ $|W| = 35531$;
  - lenta — `lenta.ru`, апрель-июнь 2016, $|D| = 8639$, $|W| = 51634$.
- Тематические модели: $|T| = 25$.
- Асессорские оценки: $K = 5$.
- Сравнивались все алгоритмы по всем метрикам.

Сравнение алгоритмов

postnauka, расстояние Жаккара:

Алгоритм	NDS	MNR	ADP	AMND	ADNP
No arranging	17.9758	12.8125	154.40	0.91	0.62
LKH	**16.7725**	**2.5208**	**53.90**	**0.72**	**0.21**
Annealing	16.8223	3.0208	64.40	0.74	0.29
t-SNE	17.9245	12.7917	140.70	0.90	0.71
MDS	18.0651	14.0833	129.80	0.97	0.79
Agl. Clust.	16.8427	3.3125	55.60	0.75	0.33

lenta, расстояние Хеллингера:

Алгоритм	NDS	MNR	ADP	AMND	ADNP
No arranging	20.4540	13.1667	174.90	0.97	0.83
LKH	**19.0180**	**3.0000**	**82.50**	**0.62**	**0.21**
Annealing	19.0661	3.4375	126.50	0.62	0.29
t-SNE	20.6573	14.9375	192.90	0.98	0.79
MDS	20.7519	15.8542	184.40	0.97	0.88
Agl. Clust.	19.0804	3.7917	94.70	0.62	0.25

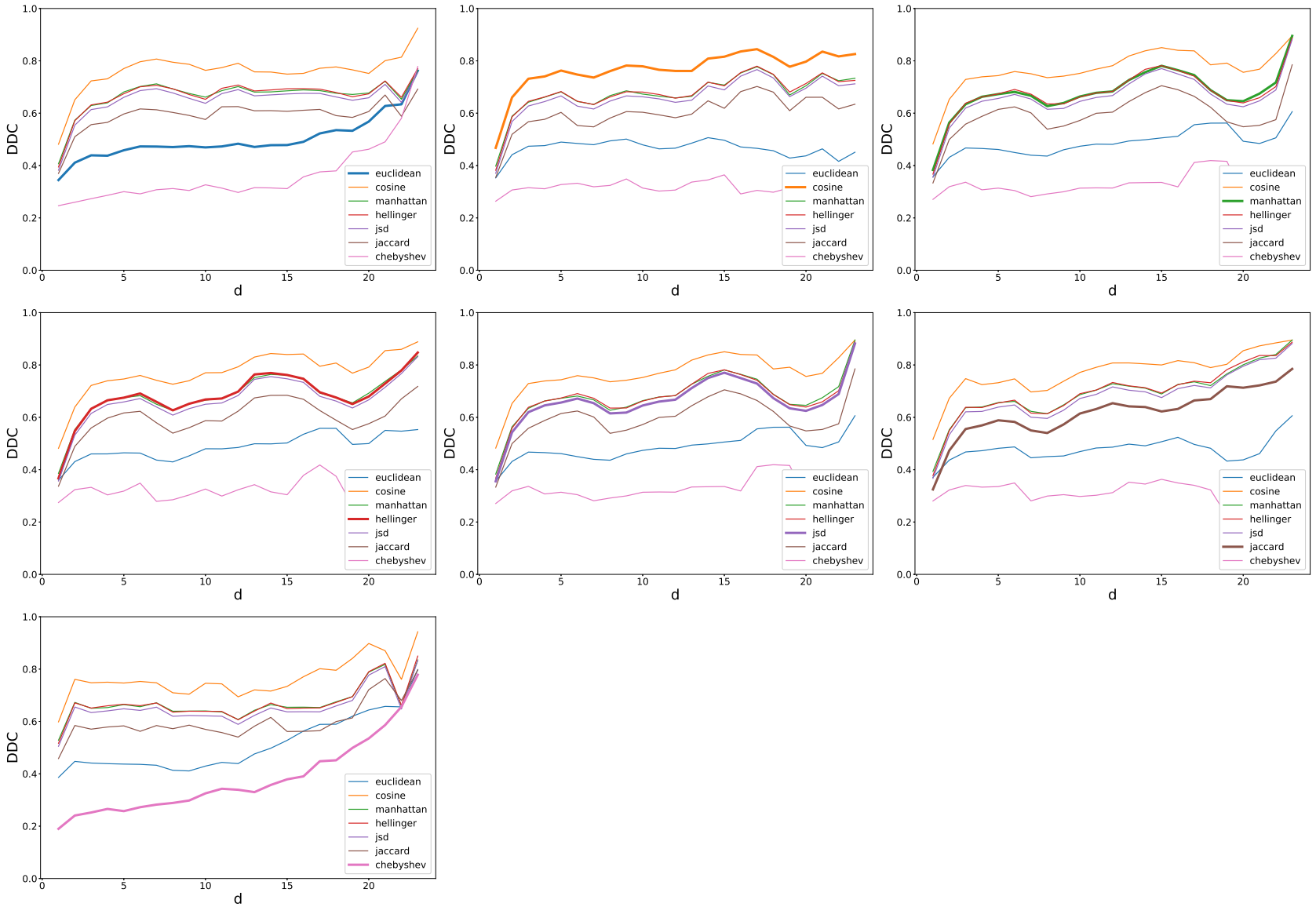
Алгоритм LKH лучше по всем мерам качества.

Сравнение функций расстояния (postnauka)

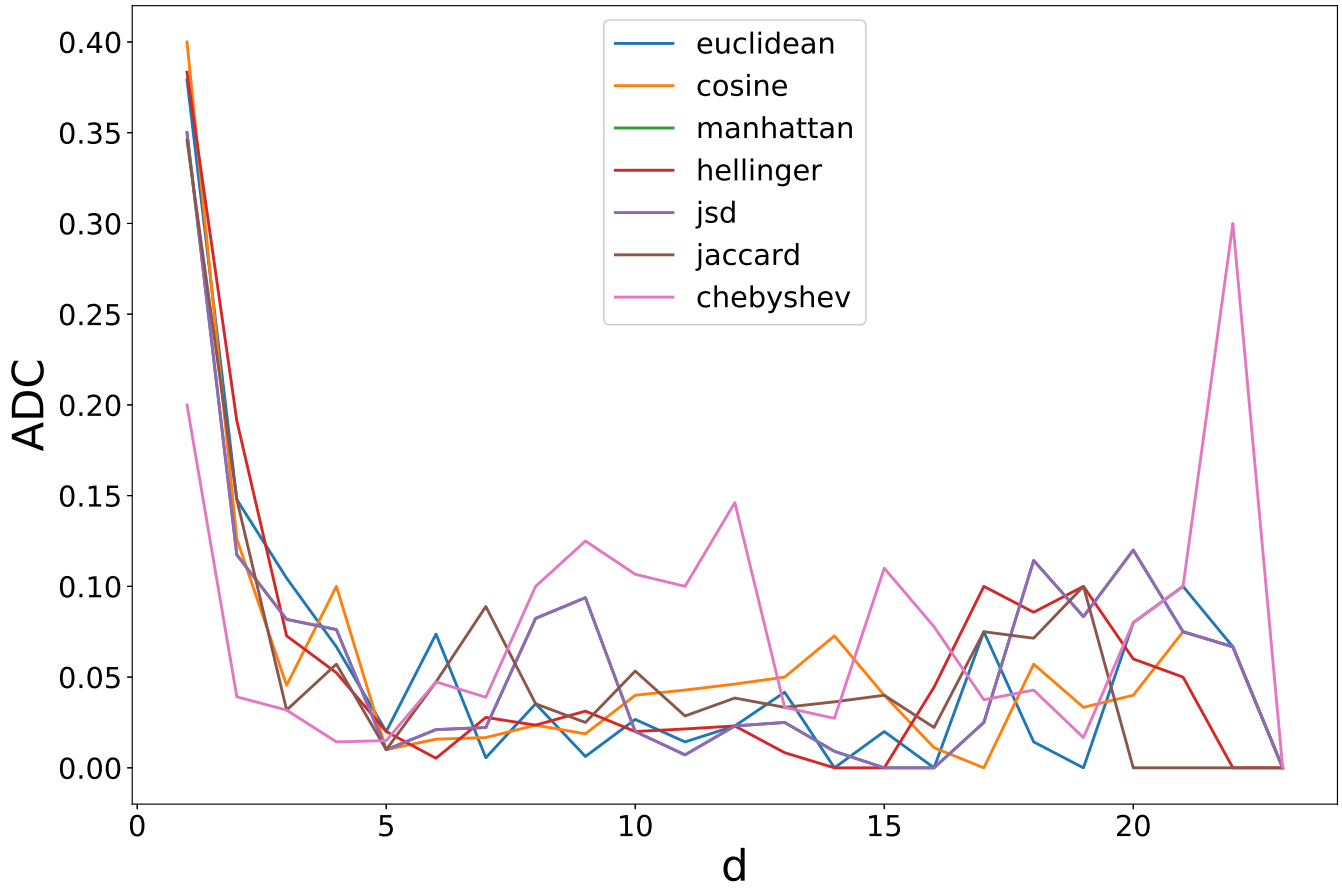
Метрика	MNR	ADP	AMND	ADNP	AMC
euclidean	7.2917	64.00	0.75	0.2917	-0.13
cosine	4.1875	**46.10**	0.70	0.2500	-0.36
manhattan	**2.2083**	54.20	0.72	**0.1667**	-0.49
hellinger	2.2292	66.00	**0.68**	0.2500	**-0.51**
jsd	2.2708	58.70	0.70	0.2083	-0.50
jaccard	2.5208	53.90	0.72	0.2083	-0.46
chebyshev	7.5625	127.80	0.85	0.4167	-0.06
Random permutations	12.2792	136.72	0.94	0.74	

Предположение: функции euclidean, manhattan, cosine, hellinger, jsd, jaccard примерно одинаково хороши.

Сравнение функций расстояния: DDC

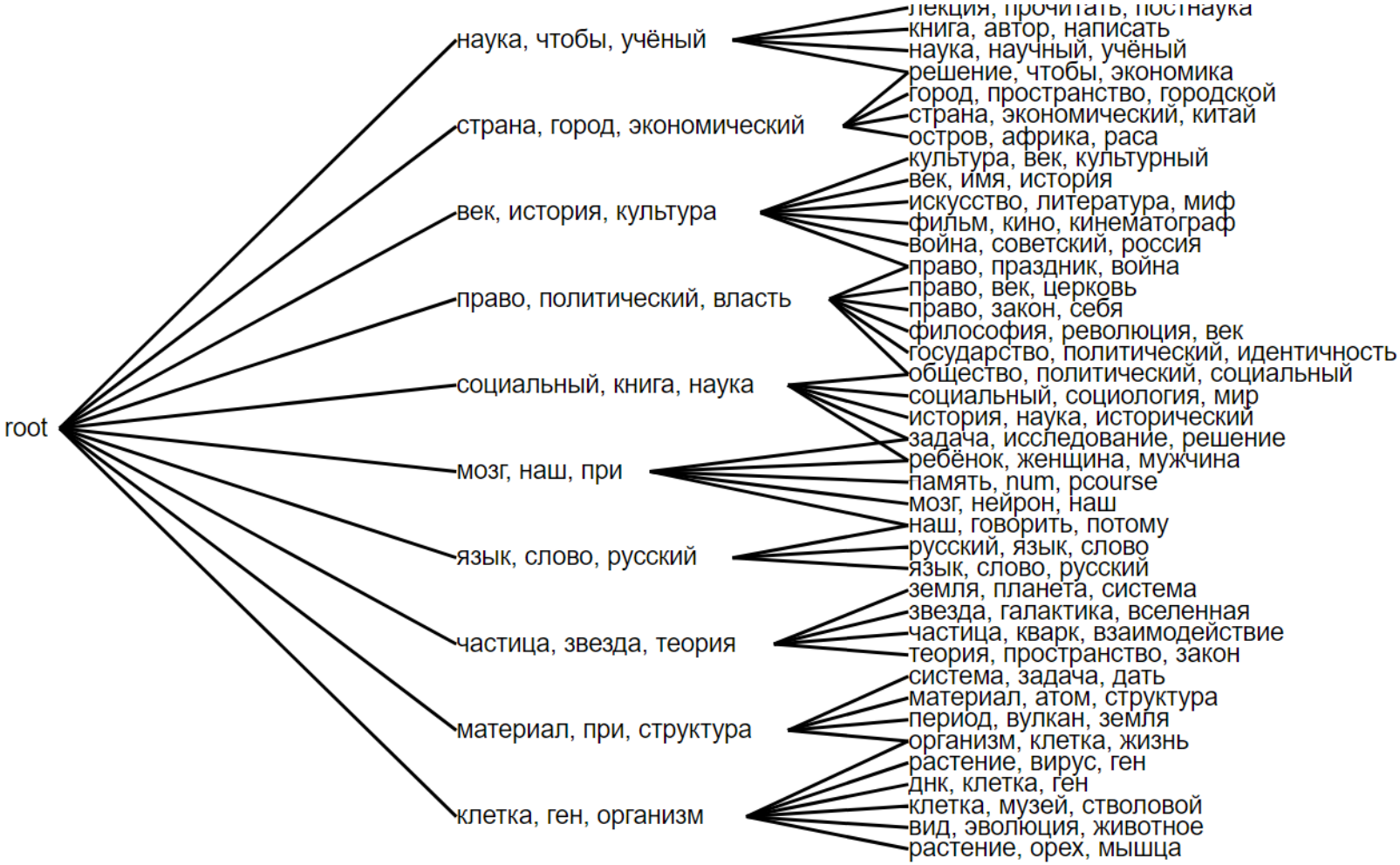


Сравнение функций расстояния: ADC



Расстояния: евклидово, манхэттенское, косинусное, Хеллингера, Йенсена-Шеннона и Жаккара примерно одинаково хороши для оценки семантической близости тем.

Спектр иерархической тематической модели



Построение двухуровневого спектра

- Модифицировать матрицу расстояний на нижнем уровне: умножить на $\beta < 1$ расстояния между всеми темами, имеющими общего родителя;
- Найти оптимальную перестановку на нижнем уровне и зафиксировать её;
- Переставляя темы на верхнем уровне, минимизировать число пересечений рёбер
  - Эвристики: медиан, барицентров, быстрой сортировки;
  - Точное решение: задача к сводится к задаче целочисленного линейного программирования $\big(O(|T_1|^2)$ переменных, $O(|T_1|^3)$ ограничений $\big)$, которую можно решать методом ветвей и границ.

VisARTM

- Web-приложение для работы с тематическими моделями;
- Доступно в Интернете: `http://visartm.vdi.mipt.ru`;
- Автоматическое построение тематических моделей с помощью BigARTM;
- Текстовые интерактивные визуализации документов, тем, терминов, модальностей;
- Визуализация иерархических моделей (вложенными прямоугольниками, многоугольниками или кругами);
- Визуализация тематических моделей во времени;
- Тематические спектры;
- Сбор асессорских оценок.

VisARTM: Визуализация документа

Химические коммуникации планктона

Эколог Егор Задереев 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 Задереев.

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.

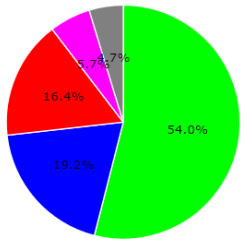
[post id="33793"]

Химические сигналы от хищников заставляют зоопланктон мигрировать. Это одно из самых масштабных на планете перемещений биомассы, которые ежесуточно 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 цепь, сеть взаимодействий до сих пор слабо исследована.

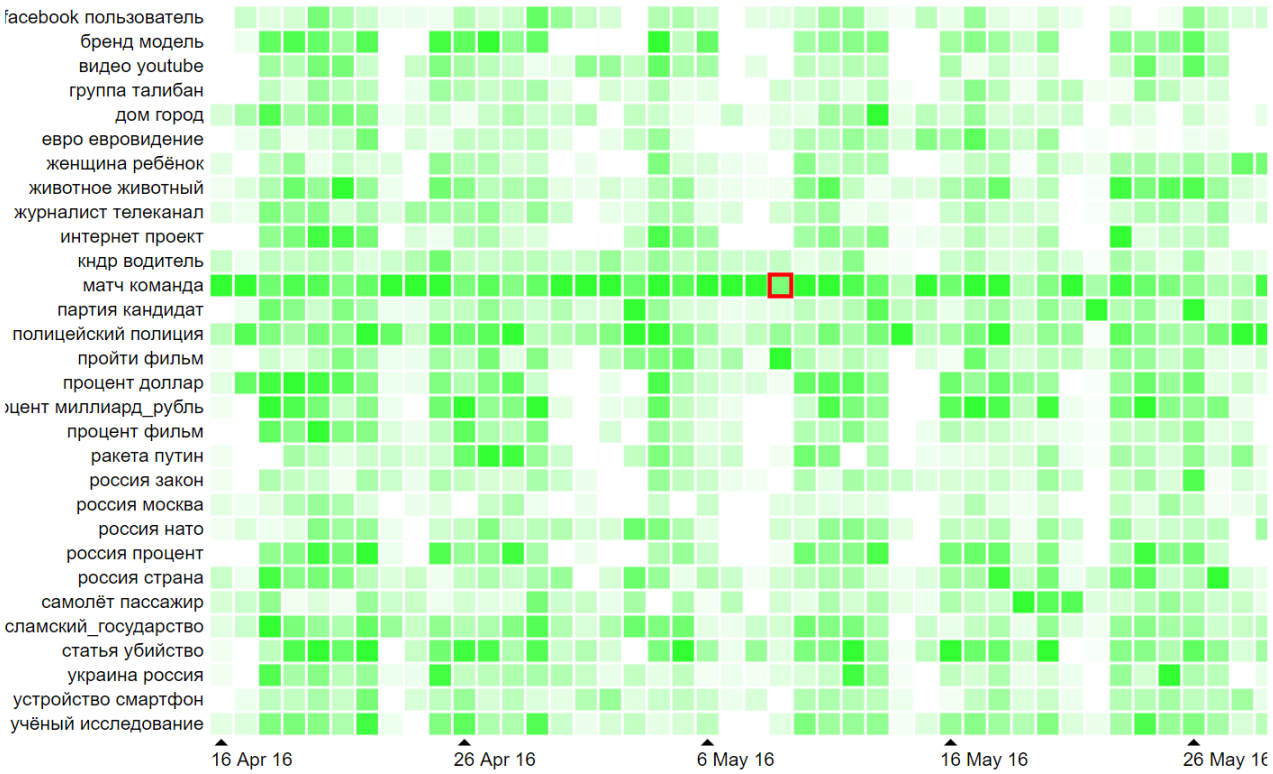
Dataset: postnauka
Time: Dec. 14, 2014, 3 p.m.
View original
index_id: 1866
text_id: 36719.txt
Terms count: 0
Unique terms count: 0
Model: flat-20
Highlighting: Words

Topic distribution



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

VisARTM: Визуализация темпоральной модели



Group by : day

Normalize by: column

Labels placement: left

Spectrum

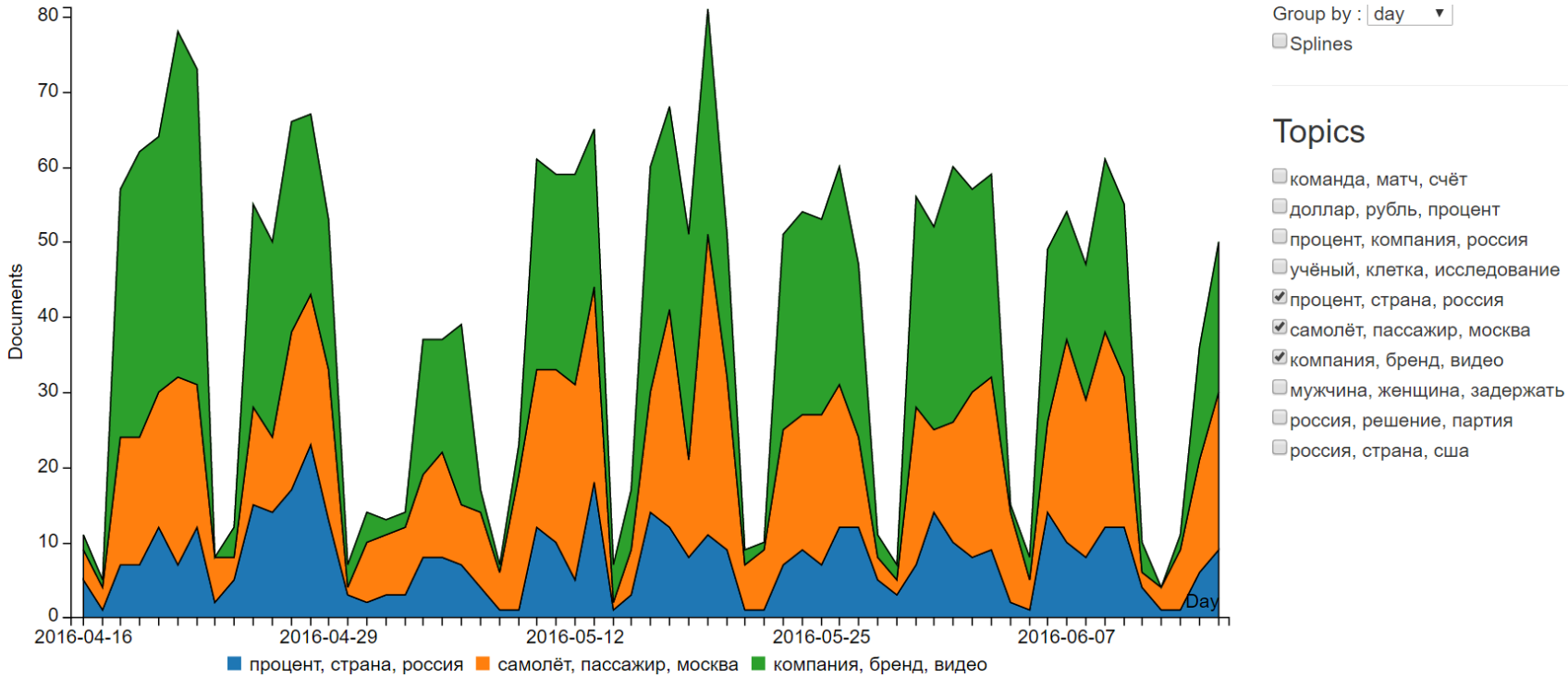
матч команда

9 May 16

Documents: 9

Форвард «Нью-Йорк Айлендерс» получил вызов в сборную России по хоккею
Выступающий в Испании украинский футболист Коноплянка сменил фамилию
Тренер сборной Латвии назвал задачу команды на матч с Россией
Полузащитник сборной России Черышев пропустит Евро-2016 из-за травмы
Сборная России обыграла Латвию на ЧМ по хоккею
Белорусы пропустили восемь шайб от канадцев на ЧМ
Российские арбитры назвали слова Гинера в адрес Иванова недопустимыми
Хоккеист сборной России Апальков получил перелом и покинул расположение команды
Американцы уступили финнам на ЧМ по хоккею

VisARTM: Визуализация темпоральной модели



VisARTM: Визуализация иерархической модели

VisARTM: Визуализация иерархической модели

Результаты, выносимые на защиту

1. Разработаны алгоритмы построения тематического спектра.
2. Предложены методы оценивания качества тематического спектра.
3. Создана информационная система для визуализации тематических моделей.