Суммаризация тем в вероятностных тематических моделях

Евгений Александрович Смирнов

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

Научный руководитель д.ф.-м.н. К. В. Воронцов

Москва,
2016 г.

Дано:

W — словарь терминов

D — коллекция документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} — сколько раз термин w встретился в документе d

n_d — длина документа

Найти:

Параметры модели $\frac{n_{dw}}{n_d} \approx p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача максимизации регуляризованного правдоподобия:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d,w} ln \sum_{t} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

Литература

1. Radev D. R. et al. Centroid-based summarization of multiple documents //Information Processing & Management. – 2004. – Т. 40. – №. 6. – С. 919-938.
2. Clustering by Passing Messages Between Data Points Brendan J. Frey, et al. Science 315 , 972 (2007)
3. Воронцов К.В. Аддитивная регуляризация тематических моделей коллекции текстовых документов. Доклады РАН, 2014. Т. 455., №3. 268-271
4. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Dudarenko Analysis of Images, Social Networks and Texts, 370-381

Машинное обучение
окно
метрический
лес
решающий
Парзен
классификатор
случайный
список
признаки
регрессия
бинаризация
главный
линейный
кластеризация
компоненты
алгомеративный
спектральный
логистический
метод
поиск

Топ-токены униграмной тематической модели

Машинное обучение
окно
метрический
лес
решающий
Парзен
классификатор
случайный
список
признаки
регрессия
бинаризация
главный
линейный
кластеризация
компоненты
алгомеративный
спектральный
логистический
метод
поиск

Топ-токены униграмной тематической модели

Машинное обучение
Метод Парзеновского окна
Метрический классификатор
Случайный лес
Решающий лес
Решающий список
Бинаризация признаков
Случайный поиск
Логистическая регрессия
Метод главных компонент
Линейный классификатор
Агломеративная кластеризация
Спектральная кластеризация

Топ-токены фразовой тематической модели

окно	признаки	Парзен	решающий	логистический
метод	главный	линейный	кластеризация	спектральный
поиск	случайный	список	метрический	классификация
лес	компоненты	регрессия	бинаризация	иерархический

1. Случайный лес представляет собой ансамбль решающих деревьев классификации или регрессии.
2. Спектральная кластеризация состоит из преобразования исходного множества объектов в пространство собственных векторов.
3. Если о структуре сообществ графа очень мало известно, удобно применять алгоритмы иерархической кластеризации.

Цель исследования:

Разработать алгоритм суммаризации тем и методы количественной оценки его качества.

Неформальные принципы включения предложения в суммаризацию темы:

- предложение относится только к одной теме
- предложение является грамматически связным
- предложение является самодостаточным, содержит законченную мысль
- предложения суммаризации вместе максимально полно представляют тему

Алгоритм построения суммаризации тем:

1. Выделить и отфильтровать тематичные предложения
2. Кластеризовать предложения
3. Выбрать наиболее ценные предложения в каждом кластере
4. Отранжировать предложения по возрастанию ценности первых k предложений

Фильтрация тематичных предложений

$p(t|w) = \frac{\phi_{wt}}{\sum\limits_{t \in S} \phi_{wt}}$ — тематический профиль термина w

$p(t|s) = \frac{1}{|s|} \sum\limits_{w \in s} p(t|w)$ — тематический профиль предложения s

$\mathcal{K}(t) = \{w \in W \mid p(t|w) \neq 0\}$ — ядро темы t

Информативность предложения s относительно темы t:

$$\mathrm{I}(s,t) = \sum_{w \in W} [w \in \mathcal{K}(t)].$$

Тематичность предложения s относительно темы t:

$$\mathrm{T}(s,t) = 1 - \frac{p(t|s)}{\|p(t|s)\|_2}.$$

Мера близости предложений:

$$\mathcal{J}(s_i, s_j|t) = \frac{\sum\limits_{w \in \mathcal{K}(t)} [w \in s_i \cap s_j]}{\sum\limits_{w \in \mathcal{K}(t)} [w \in s_i \cup s_j]}$$ — коэффициент Жаккара

$$\mathcal{J}'(s_i, s_j|t) = \frac{\sum\limits_{w \in \mathcal{K}(t)} [w \in s_i \cap s_j] p(t|w)}{\sum\limits_{w \in \mathcal{K}(t)} [w \in s_i \cup s_j] p(t|w)}$$ — модифицированный коэффициент Жаккара

Алгоритм кластеризации Affinity propagation

$s(x_i, x_j)$ — мера близости объектов x_i к x_j
$r(i,k)$ — вес сообщения, отправляемого элементом кластера
$a(i,k)$ — вес сообщения, отправляемого экземпляром кластера

1. $\forall i,k \; a(i,k) = 0$
2. Повторять до сходимости:
   $\forall i,k \; r(i,k) \leftarrow s(i,k) - \max\limits_{k' \neq k}\{a(i,k') + s(i,k')\}$
   $$\forall i \neq k, k \; a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max(0, r(i',k))\right)$$
   $\forall k \; a(k,k) \leftarrow \sum\limits_{i' \neq k} \max(0, r(i',k)$
3. Экземпляры: $c = \{c_1, \ldots, c_N\}$, где
   $c_i = arg\max\limits_k [a(i,k) + r(i,k)]$

Признаки ценности экземпляров кластеров:

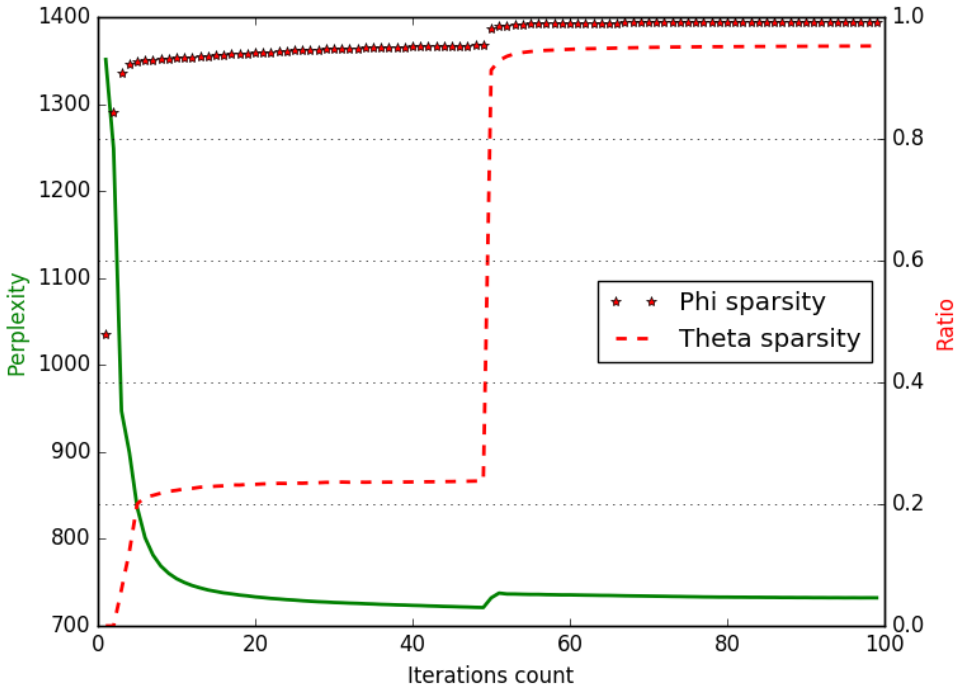
- мощность кластера по числу предложений
- мощность кластера по числу документов
- доля предложений, имеющих не менее K общих слов
- суммарная вероятность тематичных слов $p(t|w)$
- суммарная вероятность тематичных слов $p(w|t)$
- число слов из ядра темы
- длина предложения в словах
- близость к началу документа
- близость к началу раздела

$S(k,t)$ - топ k предложений суммаризации темы t
$p(w|S(k,t))$ - вероятность термина w в суммаризации $S(k,t)$

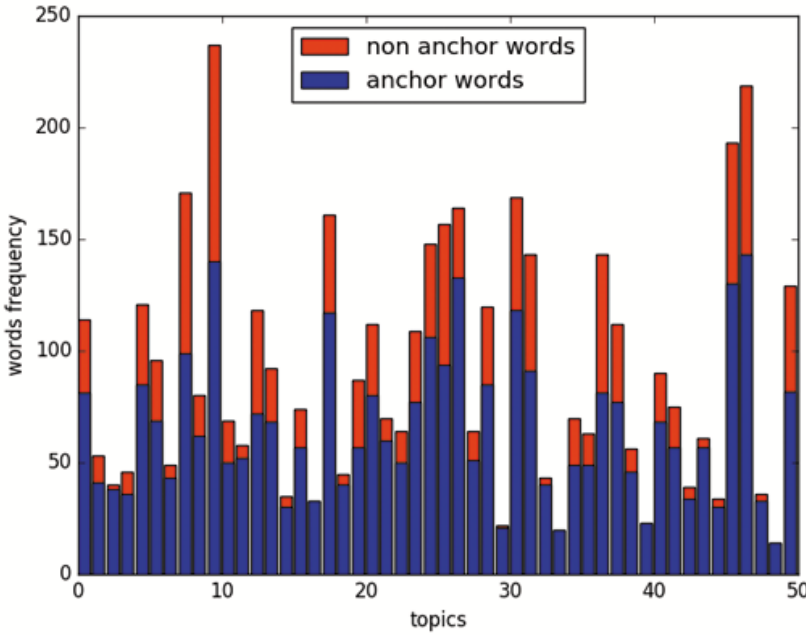
Количественные оценки качества $S(k,t)$

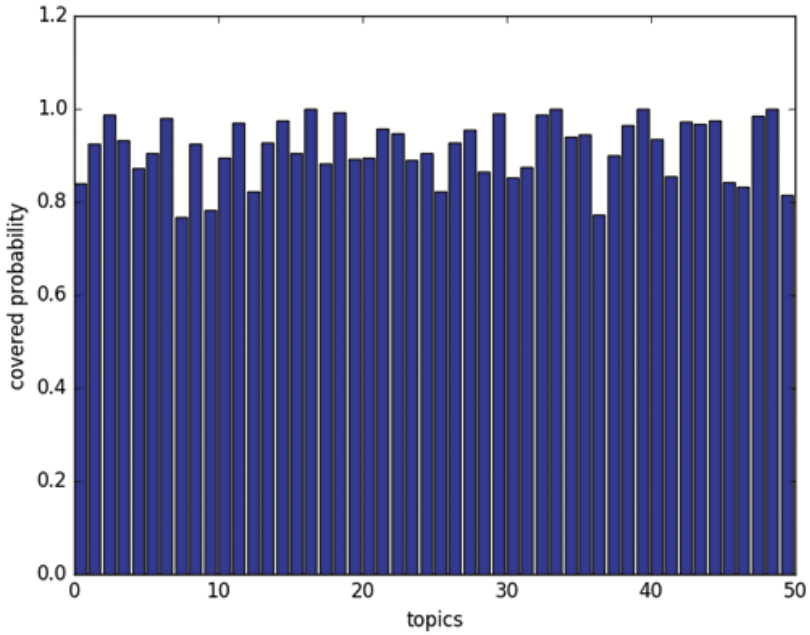
- Точность: $$\frac{\sum_{i=1}^{k}\sum_{w\in s_i}[w\in \mathrm{K}(t)]p(w|t)}{\sum_{i=1}^{k}\sum_{w\in s_i}[w\in \mathrm{K}(t)]p(w|t)+\mathbb{E}_{t'\neq t}p(w|t')}$$
- Полнота: $\sum_{w\in W}[w\in S(k,t)]p(w|t)$
- Представительность: $\cos(p(w|S(k,t)),p(w|t))$

Данные: Коллекция ММРО — ≈ 70000 предложений, $|W| = 7805$, $|D| = 1061$

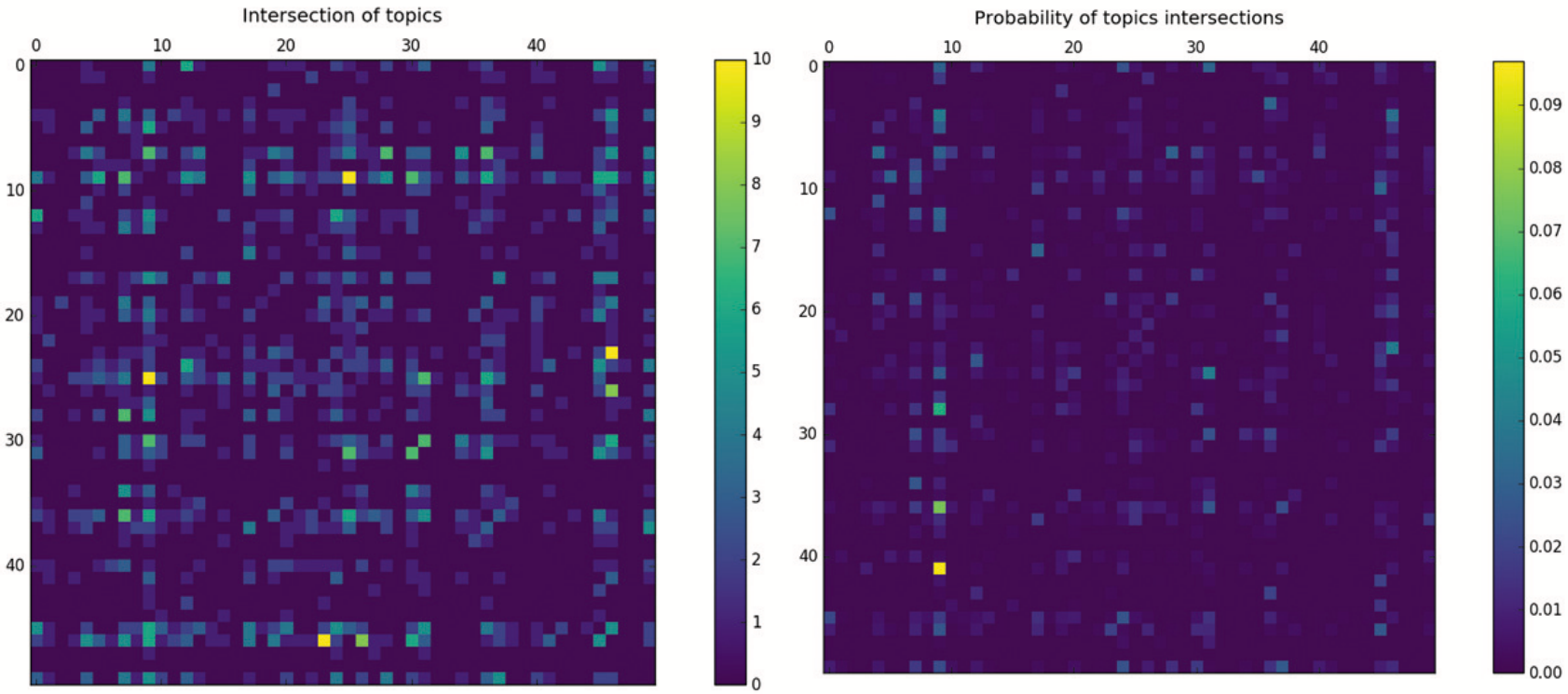


Матрицы Φ и Θ разрежены.

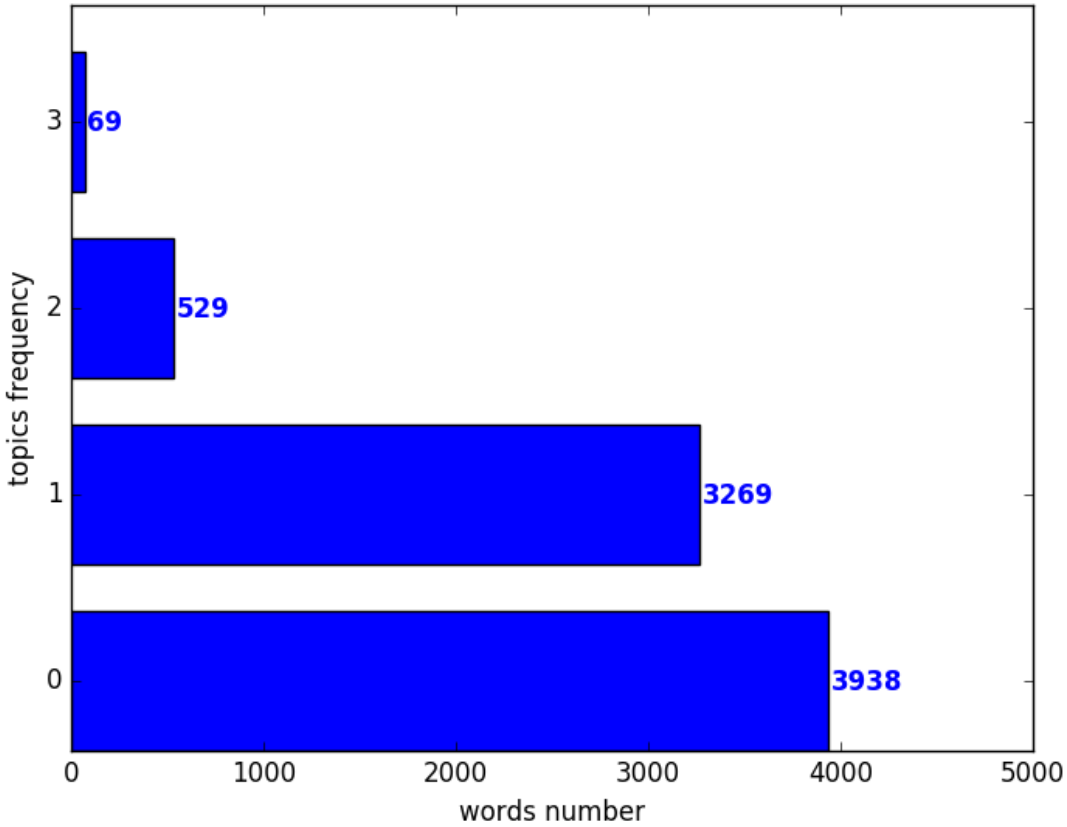




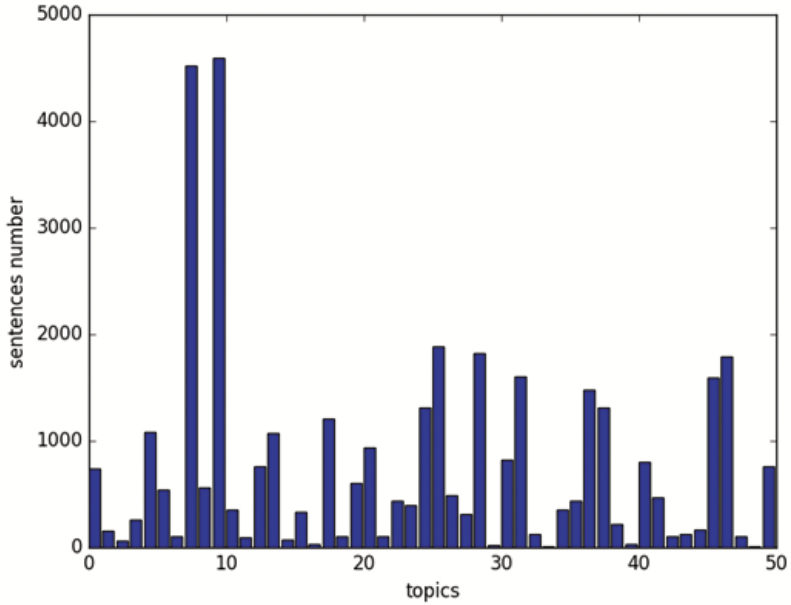
Доля якорных терминов составляет более 65%, покрытие вероятности якорными терминами каждой темы более 75%

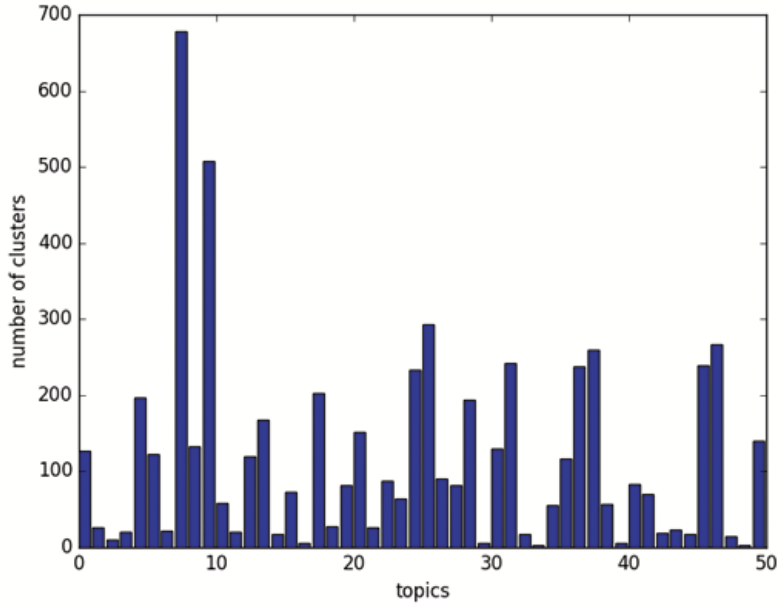


Темы пересекаются не более чем по 10 терминам, вероятностная мера пересечения < 0.1

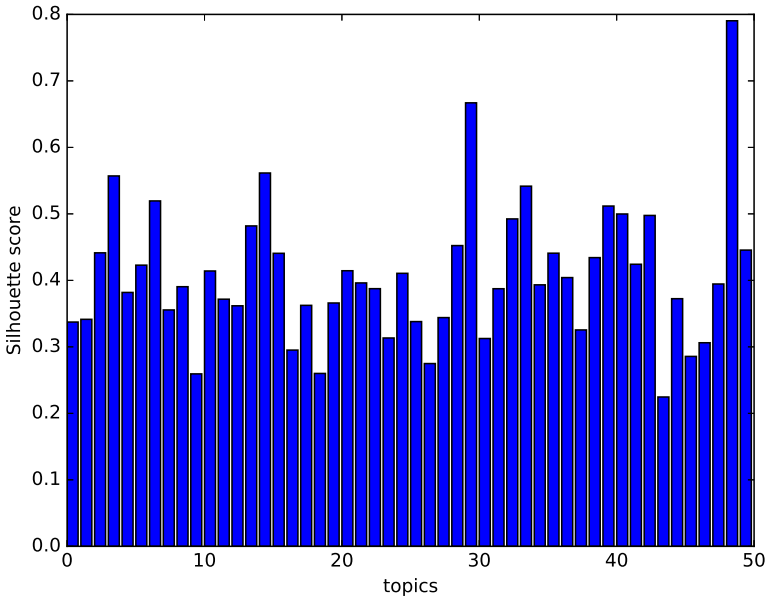


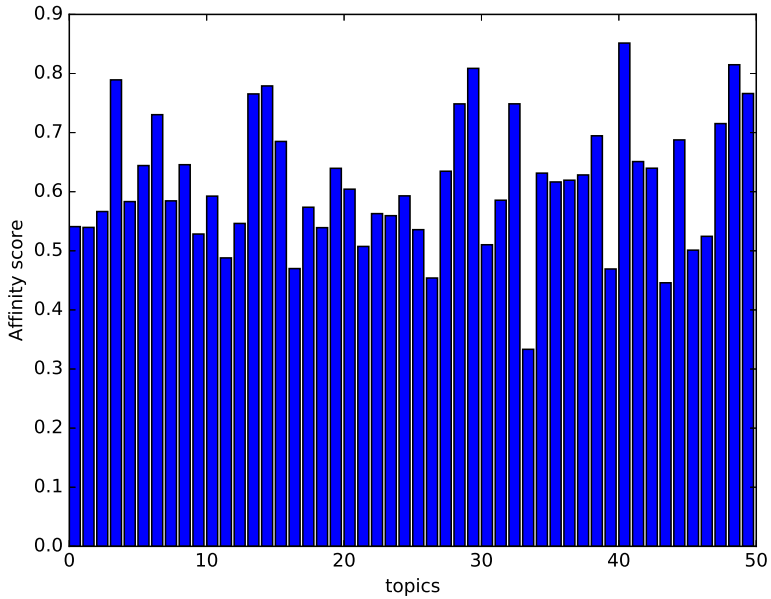
Термины относятся к небольшому числу тем





Число тематичных предложений и кластеров зависит от размера лексического ядра темы

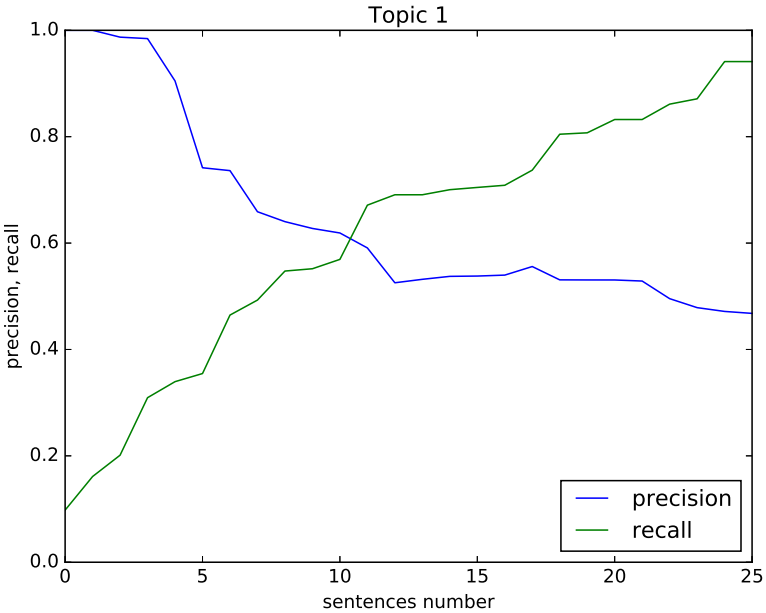


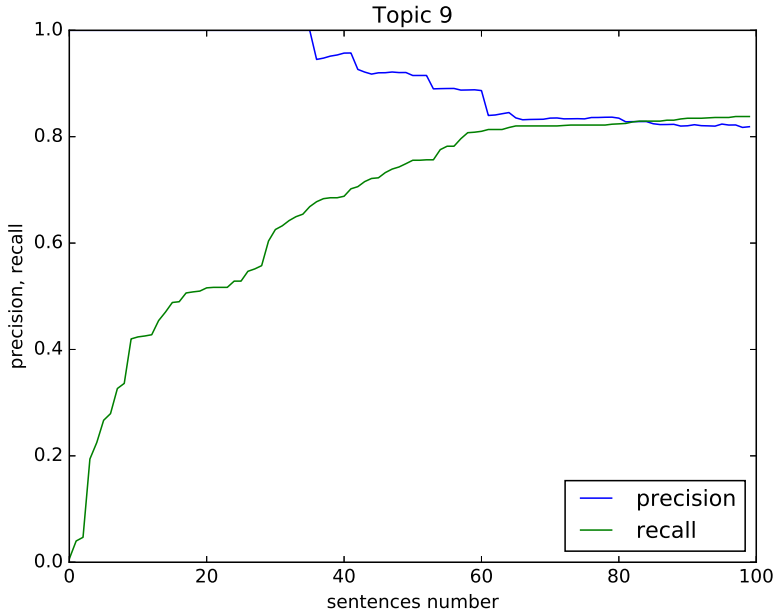


В кластеры объединяются довольно похожие предложения

Суммаризация темы №26:

- *Выявление групп **психологического** риска среди **беременных женщин**.*
- *Данная работа посвящена исследованию **применимости** существующих **артикуляционных** моделей для анализа фонетических и **паралингвистических** особенностей **речи**, выявлению корреляций между **лингвистическими** признаками **эмоций** и их **акустическими** параметрами, построению алгоритмов распознавания **эмоционального** состояния **говорящего**.*
- *Интеллектуализация поддержки принятия диагностических и **профилактических** решений по **депрессии** на основе интеллектуальной системы **ДИАПРОД**.*

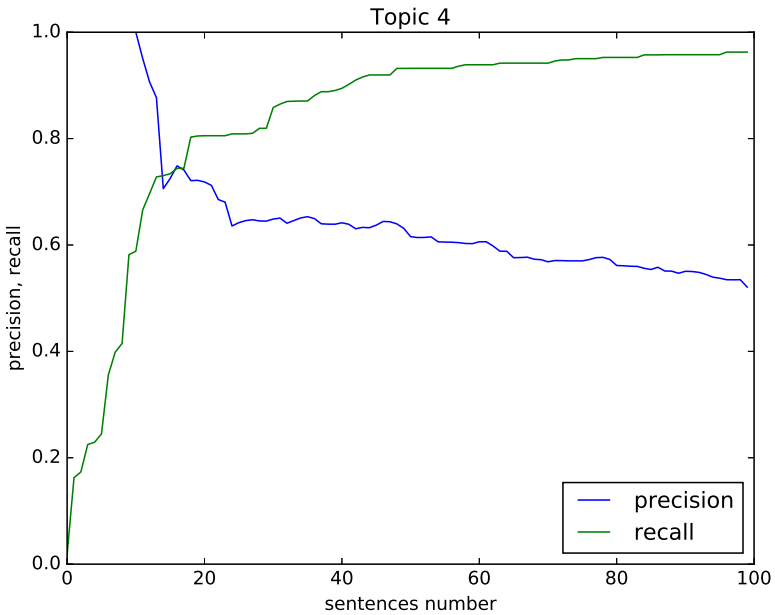


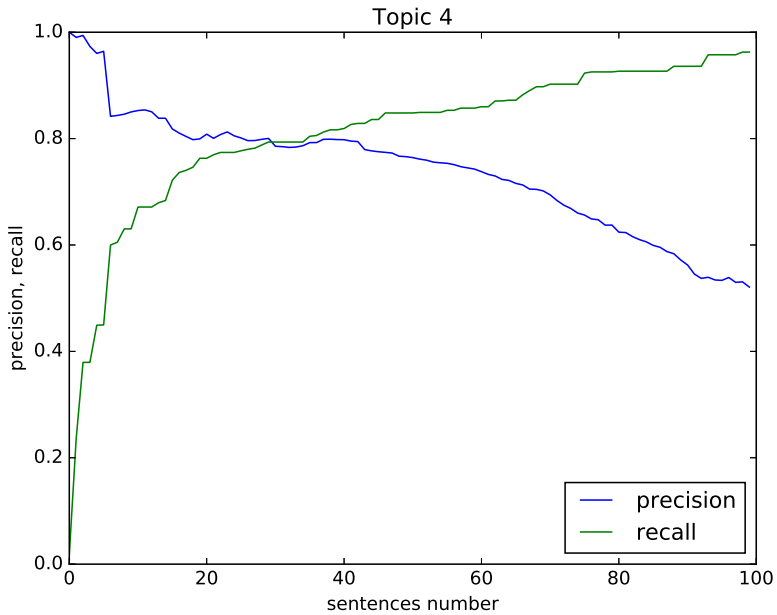


$\dim \mathcal{K}(t) = 53$

$\dim \mathcal{K}(t) = 237$

Качество суммаризации темы зависит от размера её лексического ядра





Количественные оценки качества можно улучшить перестановкой строк суммаризации темы.

Заключение

- Разработан алгоритм суммаризации тем вероятностного-тематического моделирования
- Разработаны методы количественной оценки качества суммаризации тем
- Выполнена суммаризация коллекции ММРО и вычислены количественные оценки её качества