Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации

Михаил Алексеевич Нокель

Выпускник аспирантуры факультета ВМК МГУ им. М. В. Ломоносова

Научный руководитель: к.ф.-м.н. Н. В. Лукашевич

План доклада

1. Введение в тематическое моделирование

2. Задача интеграции словосочетаний в тематические модели
   - Обзор методов интеграции словосочетаний
   - Предлагаемые алгоритмы: PLSA-SIM и PLSA-ITER
   - Тестирование предложенных алгоритмов
   - Интеграция терминов в вероятностные тематические модели

3. Задача извлечения терминов
   - Обзор методов извлечения терминов
   - Предлагаемые признаки, использующие тематическую информацию
   - Оценка вклада тематических признаков в модель извлечения терминов

Определение

Вероятностная тематическая модель *– модель, определяющая, к каким темам относится каждый документ текстовой коллекции и какие слова образуют каждую тему*

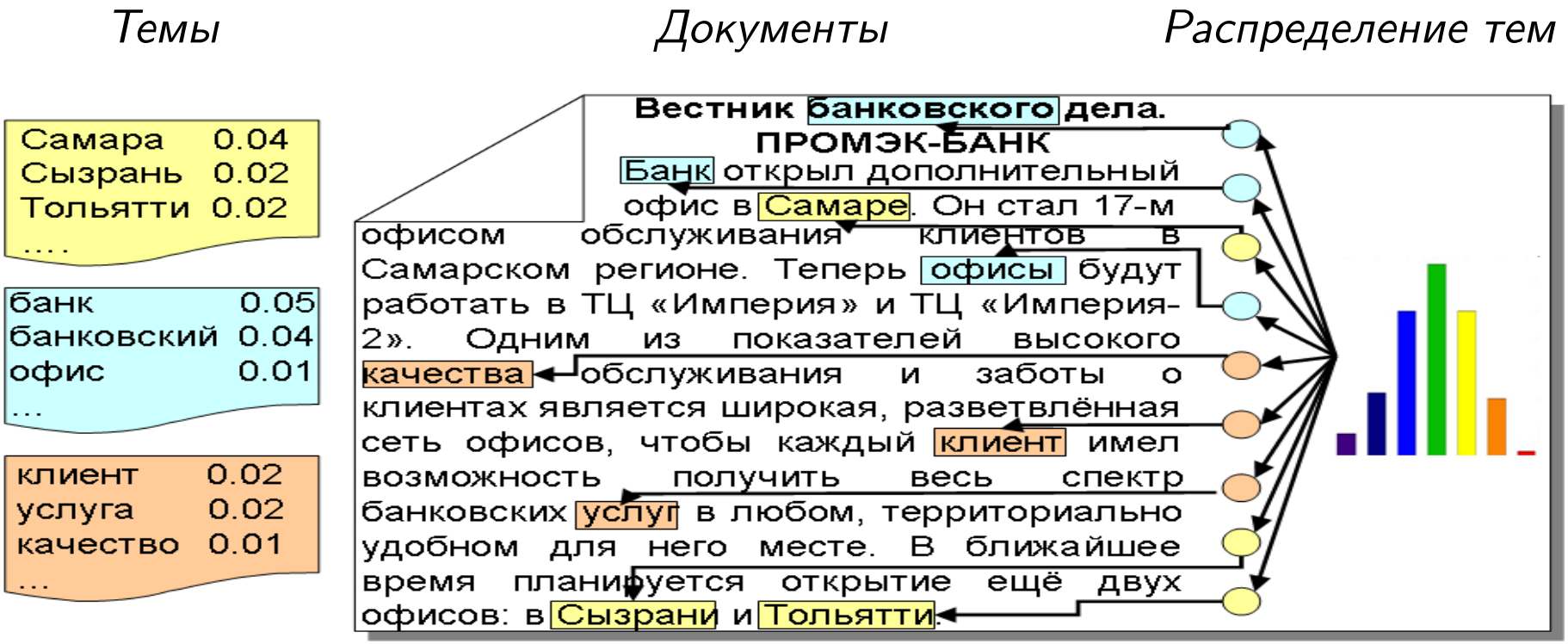
Тема *– дискретное распределение над словами*

Документ *– дискретное распределение над темами*

- **Приложения**:
  - Классификация и категоризация документов
  - Машинный перевод
  - Многодокументное аннотирование
  - ...
- **Примеры тем**:

НЕФТЬ 0.014	СУД 0.034	ГОД 0.05
РОССИЯ 0.012	ДЕЛО 0.014	РОСТ 0.04
ГАЗ 0.01	СУДЕБНЫЙ 0.014	ТЕМП 0.02
НЕФТЯНОЙ 0.009	РФ 0.013	ОБЪЕМ 0.017
СТРАНА 0.009	РЕШЕНИЕ 0.013	ДОЛЯ 0.014

Вероятностная модель порождения документов



Вход: Распределения $P(w|t)$ и $P(t|d)$

Выход: Коллекция документов D

for $d \in D$ **do**

  Задать длину документа n_d

  for $i = 1, \ldots, n_d$ **do**

    Сэмплировать тему t из распределения $P(t|d)$

    Сэмплировать слово w из распределения $P(w|t)$

    Добавить в документ d коллекции D слово w

Общая постановка задачи тематического моделирования

- Вероятностная модель порождения документа d:

$$P(w|d) = \sum_{t \in T} P(w|t)P(t|d)$$

- **Задача** – по наблюдаемым $P(w|d)$ восстановить:
  - $P(w|t)$ – распределение слов w по темам $t \in T$
  - $P(t|d)$ – распределение тем t по документам $d \in D$
- **Решение** – принцип максимума правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \text{ при условиях}$$

$$\phi_{wt} = P(w|t) \geq 0; \quad \sum_{w \in W} \phi_{wt} = 1; \quad \theta_{td} = P(t|d) \geq 0; \quad \sum_{d \in D} \theta_{td} = 1$$

- Наиболее известные алгоритмы:
  - **PLSA** – вероятностный латентный семантический анализ
  - **LDA** – латентное размещение Дирихле

Примеры тем, получаемых стандартными алгоритмами

Тема 1	Тема 2	Тема 3
БЕЗОПАСНОСТЬ 0.029	БАНК 0.048	НЕФТЬ 0.014
ИНТЕРНЕТ 0.016	БАНКА 0.027	РОССИЯ 0.012
ИНФОРМАЦИОННЫЙ 0.015	БЫТЬ 0.022	ГАЗ 0.01
ЗАЩИТА 0.015	НОВЫЙ 0.012	НЕФТЯНОЙ 0.009
ИНФОРМАЦИЯ 0.014	ГОД 0.011	СТРАНА 0.009

Тема 4	Тема 5	Тема 6
ГОД 0.053	ЭКОНОМИКА 0.041	БЫТЬ 0.024
БАНК 0.029	ИНВЕСТИЦИОННЫЙ 0.038	РОССИЯ 0.014
БАНКА 0.017	РАЗВИТИЕ 0.03	ГОД 0.011
РАБОТАТЬ 0.015	ИНВЕСТИЦИЯ 0.025	РФ 0.009
ДОЛЖНОСТЬ 0.014	ЭКОНОМИЧЕСКИЙ 0.023	ЗАКОНОПРОЕКТ 0.008

Проблема

Использование модели «мешка слов», хотя в документах много слов и словосочетаний, связанных между собой по смыслу

План доклада

1. Введение в тематическое моделирование

2. Задача интеграции словосочетаний в тематические модели
   - Обзор методов интеграции словосочетаний
   - Предлагаемые алгоритмы: PLSA-SIM и PLSA-ITER
   - Тестирование предложенных алгоритмов
   - Интеграция терминов в вероятностные тематические модели

3. Задача извлечения терминов
   - Обзор методов извлечения терминов
   - Предлагаемые признаки, использующие тематическую информацию
   - Оценка вклада тематических признаков в модель извлечения терминов

Постановка задачи интеграции словосочетаний в тематические модели

- **Цель**: разработка методов и программных средств построения тематических моделей $M(w, mwe)$, выбирающих наиболее подходящие единицы из слов w и словосочетаний mwe для интеграции и улучшения качества

$$M^*(w, mwe) = \arg \min_{M(w,mwe)} Perplexity(D')$$

$$Perplexity(D') = \exp\left(-\frac{1}{n}\sum_{d\in D'}\sum_{w\in d} n_{dw} \log P(w|d)\right), \text{ где}$$

  - n_{dw} – частотность w в документе d
  - n – общее число слов в контрольной выборке
- **Задача**: исследование и разработка методов построения тематических моделей, учитывающих словосочетания и связи между ними и образующими их словами

Обзор методов интеграции словосочетаний: создание единой унифицированной модели

- Биграммная Тематическая Модель (Wallach et al., 2006)
  - Вероятности слов зависят от вероятностей предыдущих слов
  - $W^2 T$ параметров (ср. WT параметров у LDA и $WT + DT$ параметров у PLSA)
- Модель Словосочетаний LDA (Griffiths et al., 2007)
  - Расширение Биграммной Тематической Модели за счёт дополнительных переменных для одновременного порождения слов и словосочетаний из двух слов
  - $W^2 T + W^2$ параметров
- Тематическая N-граммная Модель (Wang et al., 2009)
  - Усложнение предыдущих моделей для формирования словосочетаний в зависимости от контекста
  - $W^n T$ параметров

Проблема

Существенное усложнение моделей $\Longrightarrow$ интересны в теории, неприменимы на реальных данных

Обзор методов интеграции словосочетаний: предварительное извлечение словосочетаний

Идея

Предварительное извлечение словосочетаний для последующего их добавления в тематические модели

- On collocations and topic models (Lau et al., 2013)
  - Извлечение и ранжирование словосочетаний из двух слов по *t-score*:

$$T\text{-}Score(uv) = \frac{TF(uv) - \frac{TF(u)TF(v)}{|W|}}{\sqrt{TF(uv)}}$$

  - Отбор 1000 лучших словосочетаний по *t-score*
  - Замена в документах слов отобранными словосочетаниями

Проблема

Ухудшение перплексии

Идея

Учесть слова и словосочетания, содержащие общие слова и встречающиеся часто в одних и тех же документах

$$\exists d_i \in D : \forall w \in W : w_1, \ldots, w_k \in S_w, n_{d_i w_1} > 0, \ldots, n_{d_i w_k} > 0 \Longrightarrow r(w_1, \ldots, w_k, t)$$

- S_w – множество похожих слов и *словосочетаний* (именных групп, состоящих из двух слов)
  - \+ Обладающих семантической и тематической близостью
    - *ипотечный – ипотечный кредит – ипотечный рынок*
    - *пенсионная система – пенсионный фонд – пенсионный*
    - *жилищный – жилищный политика – жилищный кредит*
  - – Обладающих семантическими различиями
    - *точка зрения* и *точка*, *зрение*
    - *центральный банк* и *банк данных*
- $n_{d_i w_k}$ – частотность w_k в документе d_i;
- $r(w_1, \ldots, w_k, t)$ – отношение принадлежности $w_1, \ldots, w_k$ темам t

Множества похожих слов и словосочетаний

- Множества похожих слов и словосочетаний $S = \{S_w\}$, где:
  - $S_w = \{w, \bigcup\limits_v wv, \bigcup\limits_v vw\}$ – множество слов и словосочетаний, похожих на слово w
  - w и v – лемматизированные слова
  - vw и wv – лемматизированные словосочетания
- Примеры:

Множество похожих слов и словосочетаний	**Центральное слово**
коллекция, коллекция текстов, коллекция данных	*коллекция*
бюджет, бюджет субъекта, федеральный бюджет	*бюджет*
доллар, доллар США, курс доллара	*доллар*
конечный, конечный счёт, конечный итог	*конечный*
ACL, ACL workshop, proceedings of ACL	*ACL*
article, news article, newspaper article	*article*
baseline, baseline model, baseline system	*baseline*

Предложенный алгоритм PLSA-SIM

Вход: коллекция текстов D, число тем $|T|$, начальные приближения распределений $\Phi = \{P(w|t)\}$ и $\Theta = \{P(t|d)\}$

Выход: распределения Φ и Θ

1 Извлечение и отбор словосочетаний для интеграции

2 Построение множеств похожих слов и словосочетаний $S = \{S_w\}$

3 **while** *не выполнится критерий остановки* **do**

for $d \in D$, $w \in W$, $t \in T$ **do**

4 E-шаг:

$$P(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}$$

5 M-шаг:

$$n'_{dw} = n_{dw} + \sum_{s \in S_w} n_{ds}$$

$$\phi_{wt} = \frac{\sum\limits_{d \in D} n'_{dw} P(t|d,w)}{\sum\limits_{d \in D} \sum\limits_{w \in d} n'_{dw} P(t|d,w)}, \qquad \theta_{td} = \frac{\sum\limits_{w \in d} n'_{dw} P(t|d,w)}{\sum\limits_{w \in W} \sum\limits_{t \in T} n'_{dw} P(t|d,w)}$$

end

end

Теорема 1

Пусть имеется коллекция текстов D со словарём W. Пусть u – самое частотное слово в коллекции D. Тогда при добавлении любых словосочетаний вида uv_j $(j = \overline{1,k})$ в словарь W так, что:

$$\forall d \in D : n_{du} > \sum_{j=1}^{k} n_{duv_j} \text{ при условии } n_{du} > 0,$$

и построении тематической модели алгоритмом PLSA-SIM с одной темой t будут выполнены следующие неравенства:

$$\forall j = \overline{1,k}\ \forall w \in W \setminus \{u, v_1, \ldots, v_k\} :$$

$$P(uv_j|t) \geq P(u|t),$$

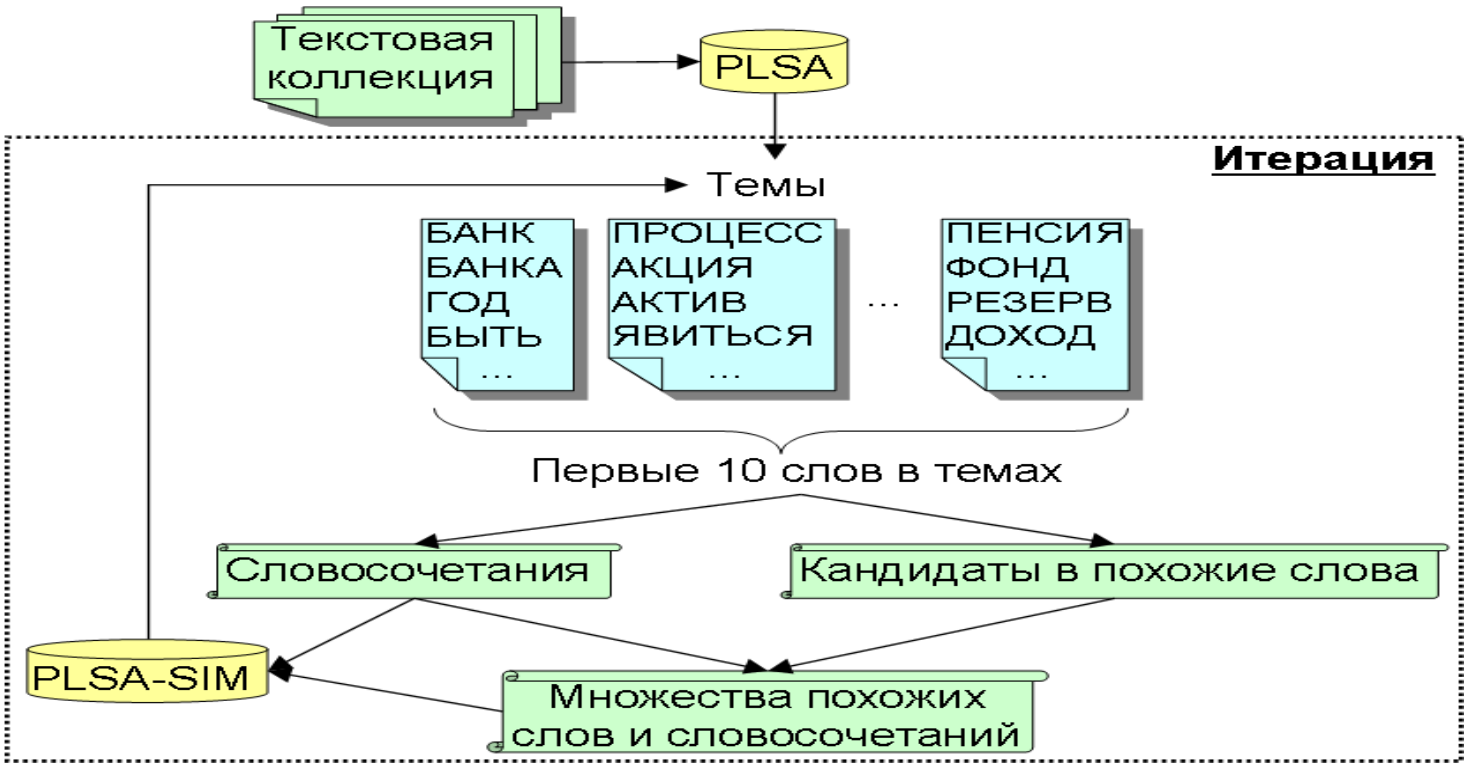
$$P(uv_j|t) > P(w|t) \text{ и } P(u|t) > P(w|t)$$

Итеративная модель учёта словосочетаний в тематических моделях и алгоритм PLSA-ITER

Идея

Автоматический выбор словосочетаний для добавления в тематические модели

- **Решение**: добавить в словарь коллекции все возможные словосочетания из первых слов в каждой теме $W = \{u_i\} \bigcup \{u_{it_k} u_{jt_k}\}$



Предложенный итеративный алгоритм PLSA-ITER

Вход: коллекция текстов D, число тем $|T|$, множество словосочетаний B
Выход: полученные темы T

Запуск алгоритма PLSA для получения тем T
$B_A = \emptyset$
while *не выполнится критерий остановки* **do**
    $S_i = \emptyset$, $B_i = \emptyset$
    for $t \in T$ **do**
        $S_i = S_i \bigcup \{u_1, u_2, \ldots, u_{10}\}$
        for $u_i, u_j \in (u_1, u_2, \ldots, u_{10})$ **do**
            if $u_i u_j \in B$ *и* $u_j u_i \in B$ *и* $TF(u_i u_j) > TF(u_j u_i)$ **then**
                $B_i = B_i \bigcup \{u_i u_j\}$
            end
        end
    end
    Формирование множества похожих слов и словосочетаний S из $S_i \bigcup B_i$
    Запуск PLSA-SIM с множествами S и $B_A = B_A \bigcup B_i$
end

Текстовые коллекции и основная мера качества

Текстовая коллекция	Число текстов	Число слов
Русскоязычные банковские тексты	10422	$\approx$ 18.5 млн
Английская часть корпуса Europarl	9673	$\approx$ 54 млн
Английская часть корпуса JRC-Acquiz	23545	$\approx$ 45 млн
ACL Anthology Reference Corpus	10921	$\approx$ 42 млн

- **Перплексия**, вычисляемая по контрольной выборке

$$Perplexity(D') = \exp\left(-\frac{1}{n}\sum_{d\in D'}\sum_{w\in d} n_{dw}\log P(w|d)\right)$$

  - Чем меньше значение, тем лучше модель предсказывает появление слов w в коллекции D'
  - Для вычисления коллекция разбивается на 2 части
    - D – для обучения модели
    - D' – для вычисления перплексии

Меры интерпретируемости тем

- **Экспертные оценки** – задача классификации на 2 класса в зависимости от того, можно ли дать теме некоторое название

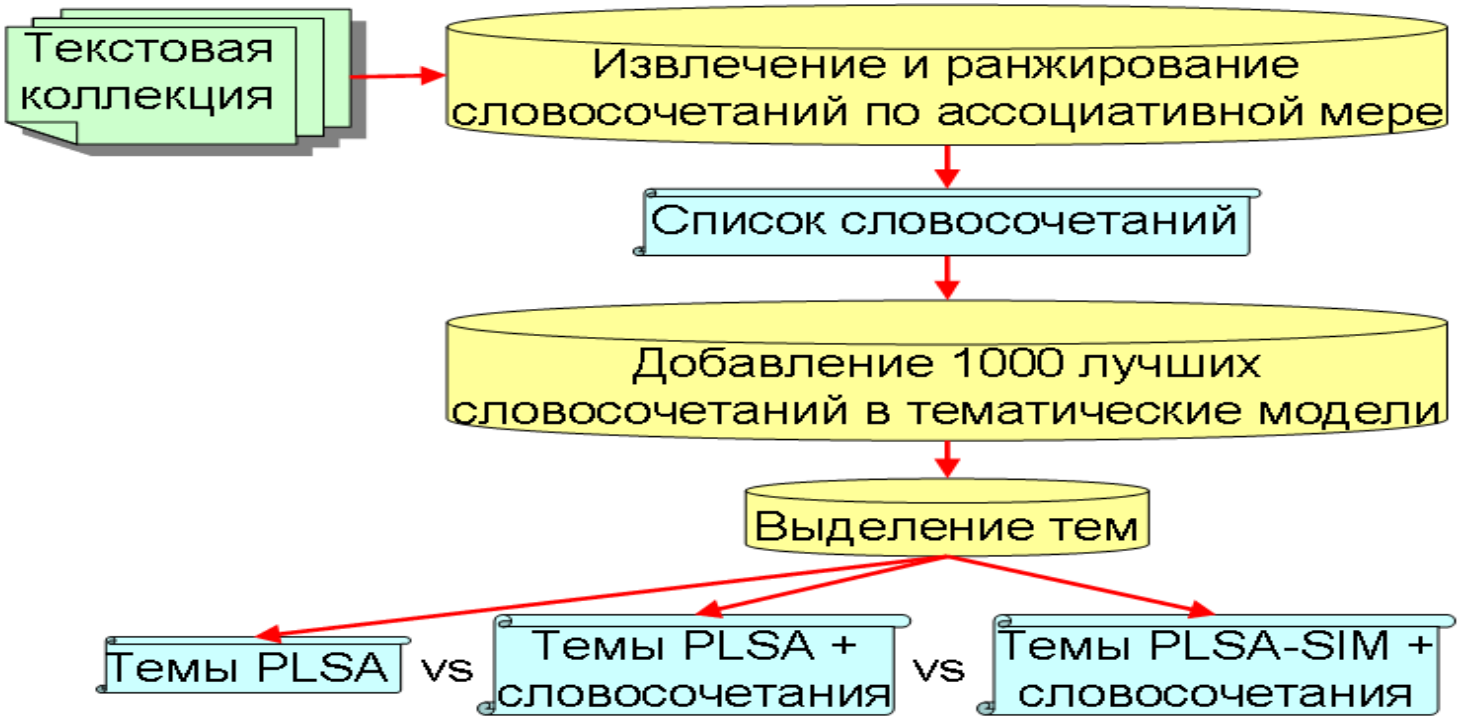
Верхняя часть списка слов из темы	**Название темы**
быть, человек, люди, год, когда, время	–
предприятие, лизинг, имущество, лизинговый	*лизинг*

- Мера согласованности тем **TC-PMI** (Newman et al., 2010)

$$TC\text{-}PMI = \frac{1}{|T|}\sum_{t=1}^{|T|}\sum_{j=2}^{10}\sum_{i=1}^{j-1}\log\frac{P(w_j^t, w_i^t)}{P(w_i^t)P(w_j^t)}$$

  - Показывает высокую корреляцию с экспертными оценками
  - Вероятности вычисляются на внешнем корпусе – Википедии
- Вариант меры согласованности тем **TC-PMI-nSIM**
  - Предлагаемые алгоритмы неявно максимизируют TC-PMI
  - Рассматривает первые 10 слов и словосочетаний, не содержащихся в одних и тех же множествах похожих слов и словосочетаний

Схема тестирования алгоритма PLSA-SIM



- 16 ассоциативных мер
- Базовая мера: частотность
- Выявление 100 тем

Вывод

Все ассоциативные меры распределились по двумя группам в зависимости от того, какие словосочетания они добавляют в тематические модели

Тестирование алгоритма PLSA-SIM: первая группа мер

- Взаимная Информация (MI, Augmented MI, Normalized MI), Хи-квадрат
- Симметричная условная вероятность, отношение логарифма правдоподобия
- Коэффициенты: простого совпадения, Юла, Кульчинского, Сёренсена

Корпус	Модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковский	*PLSA*	1724.2	86.1	86.1
	PLSA + словосочетания	1714.1	84.2	84.2
	PLSA-SIM + словосочетания	1715.4	84.1	84.1
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA + словосочетания	1584.6	55	55
	PLSA-SIM + словосочетания	1591.3	55.2	55.2
JRC-Acquiz	*PLSA*	812.1	67	67
	PLSA + словосочетания	815.4	66.3	66.3
	PLSA-SIM + словосочетания	815.6	66.4	66.4
ACL	*PLSA*	2134.7	74.8	74.8
	PLSA + словосочетания	2138.1	75.5	75.5
	PLSA-SIM + словосочетания	2144.8	75.8	75.8

Тестирование алгоритма PLSA-SIM: вторая группа мер

- Частотность, T-Score, Gravity Count, Cubic MI, True MI
- Модифицированный коэффициент Сёренсена

Корпус	Модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковский	*PLSA*	1724.2	86.1	86.1
	PLSA + словосочетания	2251.8	98.8	98.8
	PLSA-SIM + словосочетания	**1450.6**	**156.5**	**102.6**
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA + словосочетания	1993.5	57.3	57.3
	PLSA-SIM + словосочетания	**1431.6**	**127.7**	**84.7**
JRC-Acquiz	*PLSA*	812.1	67	67
	PLSA + словосочетания	1038.9	72	72
	PLSA-SIM + словосочетания	**743.7**	**108.4**	**74.9**
ACL	*PLSA*	2134.7	74.8	74.8
	PLSA + словосочетания	2619.3	73.7	73.7
	PLSA-SIM + словосочетания	**1806.4**	**152.7**	**87.8**

Тестирование алгоритма PLSA-SIM: экспертная оценка

- Сравнение трёх алгоритмов:
  - Стандартный PLSA
  - PLSA + 1000 самых частотных словосочетаний
  - PLSA-SIM + 1000 самых частотных словосочетаний

Корпус	**Модель**	**Эксперт 1**		**Эксперт 2**	
		+	−	+	−
Банковский	*PLSA*	93	7	92	8
	PLSA + словосочетания	92	8	95	5
	PLSA-SIM + словосочетания	**95**	**5**	**97**	**3**
JRC-Acquiz	*PLSA*	98	2	90	10
	PLSA + словосочетания	96	4	97	3
	PLSA-SIM + словосочетания	**100**	**0**	**100**	**0**
Europarl	*PLSA*	91	9	99	1
	PLSA + словосочетания	94	6	99	1
	PLSA-SIM + словосочетания	**99**	**1**	**100**	**0**

- Подтверждение улучшения качества тем

Примеры тем PLSA и PLSA-SIM

- Темы, полученные стандартным алгоритмом **PLSA**

МОЧЬ 0.017	БЫТЬ 0.011	РЫНОК 0.031
ПРОБЛЕМА 0.013	ЧЕЛОВЕК 0.011	ЦЕНА 0.025
БЫТЬ 0.012	ЖЕНЩИНА 0.01	РОСТ 0.023
СЛУЧАЙ 0.011	МУЖЧИНА 0.006	ГОД 0.015
РЕШЕНИЕ 0.009	ЛЮДИ 0.005	БЫТЬ 0.013
МОЖНО 0.009	ЦВЕТ 0.005	КРИЗИС 0.01
ДАТЬ 0.008	ОДЕЖДА 0.004	МОЧЬ 0.009

- Темы, полученные алгоритмом **PLSA-SIM** с добавлением 1000 самых частотных словосочетаний

ИПОТЕЧНЫЙ КРЕДИТ 0.066	СТРАХОВОЙ КОМПАНИЯ 0.08
ИПОТЕЧНЫЙ БАНК 0.046	СТРАХОВОЙ 0.058
ИПОТЕЧНЫЙ КРЕДИТОВАНИЕ 0.044	СТРАХОВАНИЕ 0.048
КРЕДИТ 0.036	СТРАХОВОЙ РЫНОК 0.047
ИПОТЕЧНЫЙ 0.033	ДОГОВОР СТРАХОВАНИЕ 0.029
ПОТРЕБИТЕЛЬСКИЙ КРЕДИТ 0.03	СТРАХОВАНИЕ ЖИЗНЬ 0.027
ИПОТЕЧНЫЙ РЫНОК 0.026	СТРАХОВОЙ СЛУЧАЙ 0.025

Множества похожих слов и словосочетаний в алгоритме PLSA-ITER

Проблема

Слишком мало множеств похожих слов и словосочетаний $\Longrightarrow$ *необходимо найти больше с помощью стеммеров*

- Модификация множеств похожих слов и словосочетаний S для учёта стеммеров:

$$S_w = \{w,\ \bigcup_u u,\ \bigcup_{u,v} uv : stem(u) = stem(w) \text{ или } stem(v) = stem(w)\}$$

Стеммер	**Множество похожих слов и словосочетаний**
Snowball	*тайна, банковская тайна, тайный*
	право, право собственности, правый, правая сторона
Портер	*fish, fish agreement, fishing, fishing agreement*
	alcohol, use of alcohol, alcoholic, alcoholic product
Ланкастер	*budget, budget year, budgetary, budgetary year*
	culture, european culture, cultural, cultural Europe

Тестирование алгоритма PLSA-ITER

- *Результаты для первой итерации* (число тем $|T| = 100$):

Корпус	Модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковский	*PLSA-SIM*	1450.6	156.5	102.6
	PLSA-ITER + Snowball	**1265.1**	137.6	96.7
Europarl	*PLSA-SIM*	1431.6	127.7	84.7
	PLSA-ITER + Портер	1293.8	99.6	61.2
	PLSA-ITER + Ланкастер	**1077.7**	105	55.2
JRC-Acquiz	*PLSA-SIM*	743.7	108.4	76.9
	PLSA-ITER + Портер	777.7	90.8	68.2
	PLSA-ITER + Ланкастер	**736.5**	94.5	68.6
ACL	*PLSA-SIM*	1806.4	152.7	87.8
	PLSA-ITER + Портер	1853.7	123.6	76.2
	PLSA-ITER + Ланкастер	**1772.1**	121.3	76.5

- Экспертные оценки подтверждают улучшение качества

Тестирование алгоритма PLSA-ITER: первые итерации

Корпус	Итерация	Перплексия	TC-PMI	TC-PMI-nSIM
Банковский	*0 (PLSA)*	1724.2	86.1	86.1
	1	1265.1	137.6	96.7
	2	1257.1	133.5	95
	3	1259.8	134.5	95.7
Europarl	*0 (PLSA)*	1594.3	53.2	53.2
	1	1077.7	105	55.2
	2	1210.8	92.1	55.2
	3	1242.9	80.1	53.2
JRC-Acquiz	*0 (PLSA)*	812.1	67	67
	1	736.5	94.5	68.6
	2	751.9	94.9	67
	3	749.6	99.5	67.7
ACL	*0 (PLSA)*	2134.7	74.8	74.8
	1	1772.1	121.3	76.5
	2	1775.5	139.3	81
	3	1767.6	144.6	83

- Колебание результатов вокруг одних и тех же значений мер качества

Примеры тем PLSA и PLSA-ITER

- Темы, полученные стандартным алгоритмом **PLSA**

МЛРД 0.056	БЫТЬ 0.008	ДОЛЖНЫЙ 0.043
ДОЛЛ 0.028	ОТЕЛЬ 0.006	ДОЛЖЕН 0.043
ДОЛ 0.019	ЧЕЛОВЕК 0.006	БЫТЬ 0.027
ОБЪЕМ 0.018	МОЖНО 0.005	МОЧЬ 0.019
ИНВЕСТИЦИЯ 0.013	ГОД 0.005	ПРИНЦИП 0.015
БАЛАНС 0.013	САМОЛЕТ 0.004	ТРЕБОВАНИЕ 0.013
США 0.011	ВОДА 0.004	НЕОБХОДИМЫЙ 0.012

- Темы, полученные алгоритмом **PLSA-ITER** после первой итерации

АУДИТОРСКИЙ ОРГАНИЗАЦИЯ 0.052	**СТРАХОВОЙ** КОМПАНИЯ 0.127
АУДИТОРСКИЙ ДЕЯТЕЛЬНОСТЬ 0.046	**СТРАХОВОЙ** 0.067
АУДИТОРСКИЙ ПРОВЕРКА 0.04	**СТРАХОВОЙ** РЫНОК 0.053
АУДИТОРСКИЙ 0.039	**СТРАХОВАНИЕ** 0.05
АУДИТ 0.037	КОМПАНИЯ 0.033
ВНУТРЕННИЙ КОНТРОЛЬ 0.033	**СТРАХОВАНИЕ** ЖИЗНЬ 0.03
АУДИТОР 0.027	**СТРАХОВОЙ** СЛУЧАЙ 0.028

Интеграция терминов в вероятностные тематические модели

Определение

Термин – *эталонное слово или словосочетание, выделенное экспертом предметной области*

- *«Золотые стандарты»* – ручные терминологические ресурсы, разработанные экспертами:
  - Банковский тезаурус ($\approx$ 15000 терминов) – банковская коллекция
  - Eurovoc ($\approx$ 15000 терминов) – коллекция Europarl
- Результаты добавления терминов (число тем $|T| = 100$)

Корпус	Модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковский	*PLSA*	1724.2	86.1	86.1
	PLSA-SIM + термины	1465.2	**153.8**	**106.4**
	PLSA-ITER + термины	**1267.6**	134.9	96
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA-SIM + термины	1519.9	**139**	**88.9**
	PLSA-ITER + термины	**1193.5**	97.4	66.6

План доклада

1. Введение в тематическое моделирование

2. Задача интеграции словосочетаний в тематические модели
   - Обзор методов интеграции словосочетаний
   - Предлагаемые алгоритмы: PLSA-SIM и PLSA-ITER
   - Тестирование предложенных алгоритмов
   - Интеграция терминов в вероятностные тематические модели

3. Задача извлечения терминов
   - Обзор методов извлечения терминов
   - Предлагаемые признаки, использующие тематическую информацию
   - Оценка вклада тематических признаков в модель извлечения терминов

Постановка задачи терминологического анализа

Определение

Автоматическое извлечение терминов – *процедура упорядочивания множества слов и словосочетаний S из текстовой коллекции D так, чтобы эталонные термины T_e оказались в начале списка*

- $S = \{\mathbf{s_1} = (f_1(s_1), \ldots, f_n(s_1)), \ldots, \mathbf{s_k} = (f_1(s_k), \ldots, f_n(s_k))\}$ – множество признаковых описаний слов и словосочетаний
- $T_e \subset S$ – множество признаковых описаний эталонных терминов
- $f_i(x) : X \to D_f$ – признак
- **Задача ранжирования**: построить ранжирующую функцию $a : S \to \{0, 1\}$ такую, что

$$i \prec j \Longrightarrow a(\mathbf{s_i}) \prec a(\mathbf{s_j})$$

$$a^* = \arg\max_a |\{\mathbf{t} \in T_e | \forall \mathbf{s} \in S \setminus T_e : a(\mathbf{s}) \prec a(\mathbf{t_e})\}|$$

  - $i \prec j$ – правильный порядок на парах $(i, j) \in \{1, \ldots, l\}$, известный на объектах обучающей выборки $S^m = \{\mathbf{s_1}, \ldots, \mathbf{s_l}\}$

Существующие признаки терминологичности

- **Основанные на частотности** (*8 признаков*)
  - *Идея*: термины встречаются чаще остальных слов
  - *TF, DF, TF-IDF, TF-RIDF, Domain Consensus, Term Variance и др.*

$$TF\text{-}IDF(w) = TF(w) \times \log \frac{|D|}{DF(w)}$$

- **Использующие контрастную коллекцию** (*7 признаков*)
  - *Контрастная коллекция* – коллекция более общей тематики
  - *Идея*: частотности терминов в целевой коллекции значительно больше, чем в контрастной
    - Британский национальный корпус
    - Русский национальный корпус
  - *Weirdness, Relevance, Contrastive Weight, KF-IDF, TF-IDF и др.*

$$Weirdness(w) = \frac{TF(w)}{|W|} \Big/ \frac{TF_r(w)}{|W_r|}$$

Существующие признаки терминологичности

- **Контекстные признаки** (*11 признаков*)
  - Соединяют частотность с данными о контексте употребления
  - *C-Value, NC-Value, Type-LR, Token-FLR, Insideness, Sum3 и др.*

$$C\text{-}Value(w) = TF(w) - \frac{\sum\limits_{p \in P_w} TF(p)}{|P_w|}$$

- **Ассоциативные меры** (*16 признаков*)
  - Оценивают силу связи между составными частями фраз
  - Применимы только для многословных выражений
  - *MI, T-Score, Loglikelihood Ratio, Gravity Count, DC и др.*

$$MI(xy) = \log \frac{|W| \times TF(xy)}{TF(x) \times TF(y)}$$

- **Гибридные признаки** (*3 признака*)
  - Объединение идей признаков нескольких категорий для лучшего ранжирования кандидатов

$$TE(w) = \frac{1}{3}(Weirdness(w) + DomainConsensus(w) + ModifiedDC(w))$$

Прочие признаки терминологичности

- **Modified Gravity Count (MGCount)**
  - Новый контекстный признак
  - Основан на ассоциативной мере Gravity Count

$$MGCount(xy) = \log\left(\frac{TF(xy)l(x)}{TF(x)} + \frac{TF(xy)r(y)}{TF(y)}\right)$$

- **Прочие ad-hoc признаки** (*27 признаков*)
  - Лингвистические признаки:
    - *Многозначность, новизна, специфичность, сущ. и прил.*
  - *TF, DF, TF-IDF (2 варианта), TF-RIDF и Domain Consensus*
    - Для подлежащих (сущ. в имен. падеже)
    - Для слов с большой буквы
    - Для слов с большой буквы, не начинающих предложения
  - *NearTermsFreq* и *NearTermsFreq-IDF*
    - Для кандидатов, находящихся в контекстном окне самых частотных кандидатов в термины
  - *Номер первой позиции* в документах и *Длина термина*

Модель применения тематической информации для извлечения терминов

Проблема

Существующие признаки слабо отражают тот факт, что большинство терминов относятся к той или иной тематике

$$Term = Term_{general} \bigcup \left(\bigcup_{t \in T} Term_t \right)$$

$$\forall t \in T : |Term_{general}| << |Term_t| \text{ и } Term_{general} \bigcap Term_t = \emptyset$$

- Для выявления тематик – **тематические модели**
  - Базовая: каждый документ – отдельная тема
  - Тематические модели на методах кластеризации
    - K-Средних, Сферический K-Средних, метод NMF
    - Иерархическая с одиночным, полным и средним связыванием
  - Вероятностные тематические модели
    - PLSA, LDA

Признаки, использующие тематическую информацию

Идея

Термины относятся к той или иной тематике, обсуждаемой в рамках текстов предметной области

Признак	**Формула**
Частотность (TF)	$\sum_{t \in T} P(w\|t)$
TF-IDF	$TF(w) \times \log \frac{\|T\|}{DF(w)}$
Maximum TF	$\max_{t \in T} P(w\|t)$
Domain Consensus (DC)	$-\sum_{t \in T} (P(w\|t) \times \log P(w\|t))$
Term Score (TS)	$\sum_{t \in T} TS(w\|t)$, $TS(w\|t) = P(w\|t) \times \log \frac{P(w\|t)}{\left(\prod_{t \in T} P(w\|t)\right)^{\frac{1}{\|T\|}}}$
TS-IDF	$TS(w) \times \log \frac{\|T\|}{DF(w)}$
Maximum TS	$\max_{t \in T} TS(w\|t)$

Текстовые коллекции и мера качества

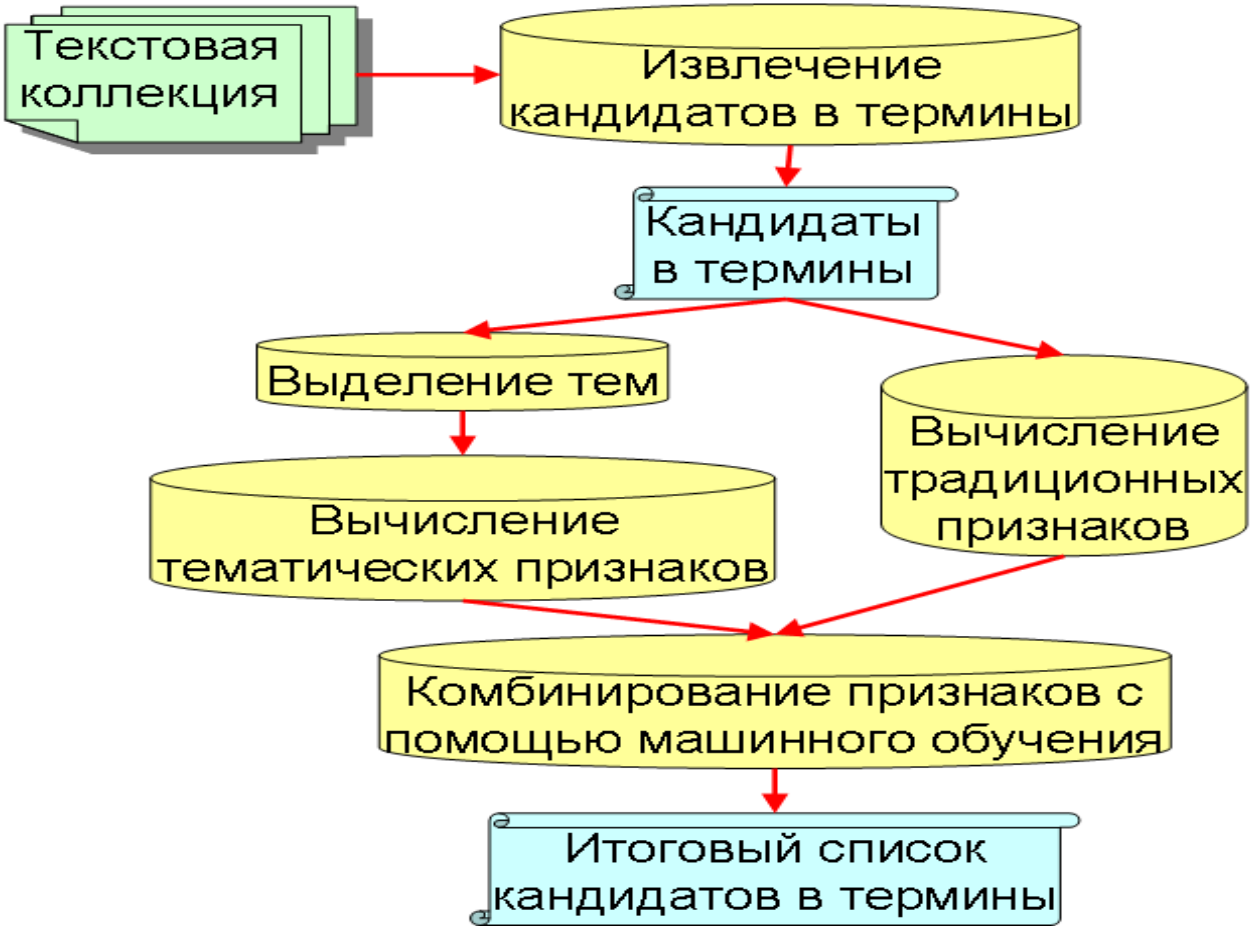
Текстовая коллекция	Число текстов	Число слов
Русскоязычные банковские тексты	10422	$\approx$ 18.5 млн
Английская часть корпуса Europarl	9673	$\approx$ 54 млн

- **Подтверждение терминов**: кандидат считается термином, если он есть в «золотом стандарте»:
  - Банковский тезаурус ($\approx$ 15000 терминов) – банковская коллекция
  - Eurovoc ($\approx$ 15000 терминов) – коллекция Europarl
- **Мера качества** – Средняя Точность (AvP)

$$AvP@n = \frac{\sum_{k=1}^{n} \text{Точность@}k}{\text{число терминов}}$$

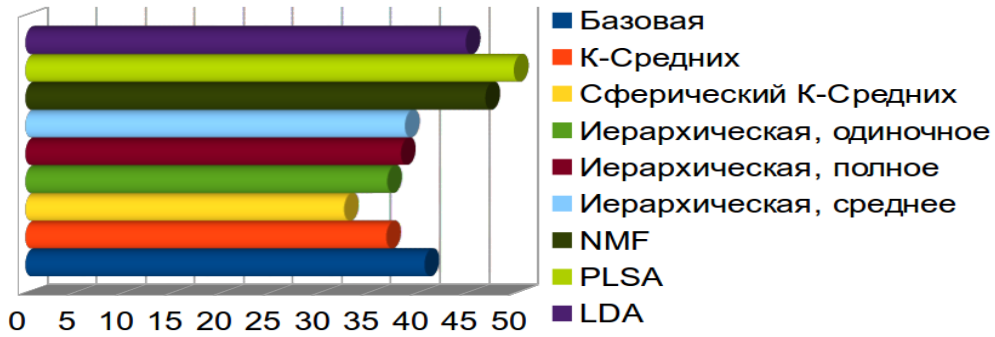
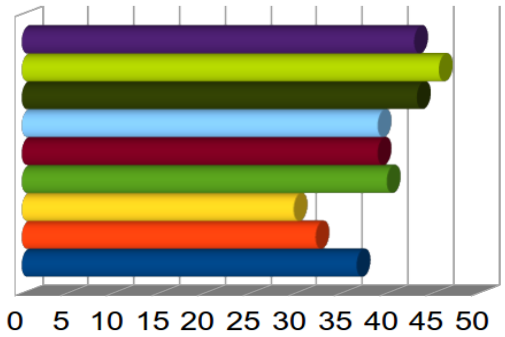
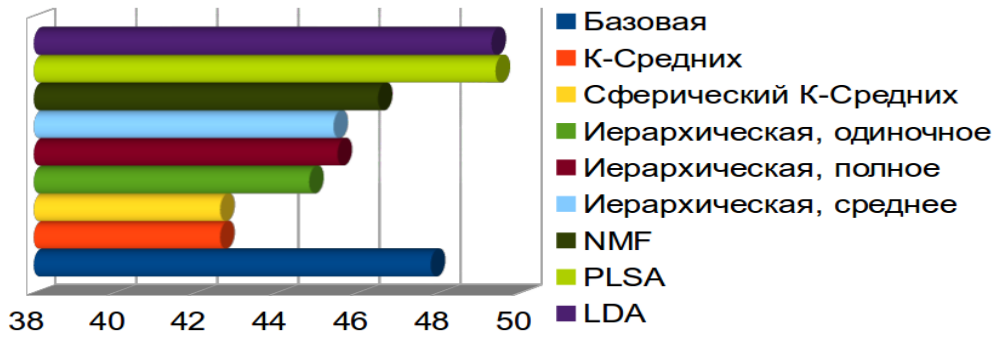
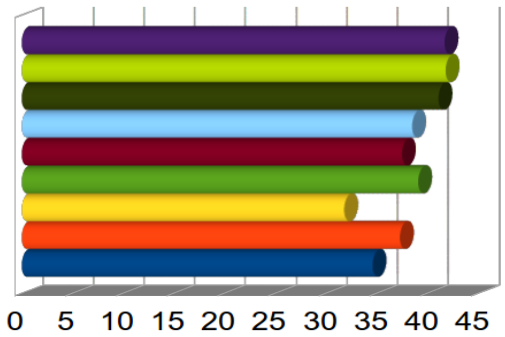
- Число тем $|T| = 100$
- Все признаки рассчитывались для 5000 самых частотных кандидатов в термины

Схема тестирования извлечения терминов



- Комбинирование признаков – метод градиентного бустинга
- Метод скользящего контроля по четырём блокам

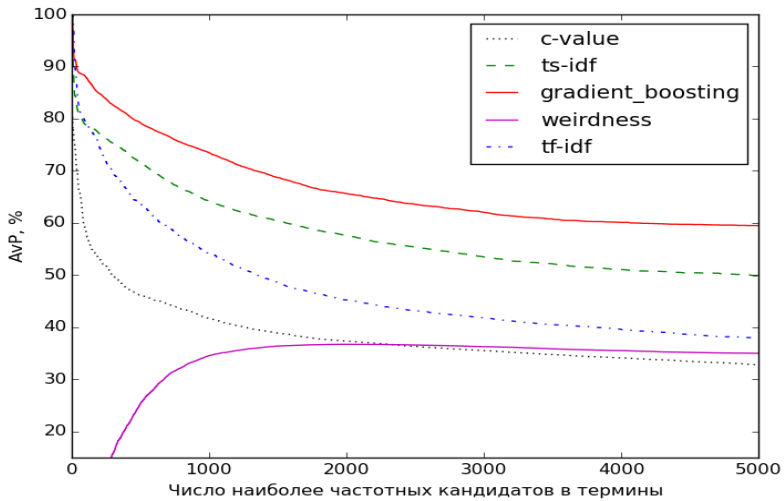
Лучшая тематическая модель и лучший признак

Банковский корпус	Корпус Europarl
Однословные термины 	Однословные термины 
Двусловные термины 	Двусловные термины 

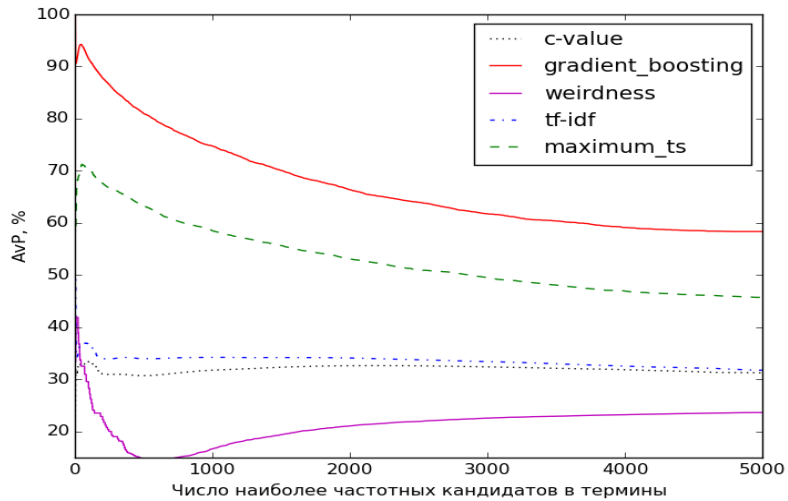
Термины	Корпус	Лучший признак	Прирост качества
Однословные	*Банковский*	**TS-IDF**	**+7.7%**
	Europarl	**Maximum TS**	**+16%**
Двусловные	*Банковский*	**MGCount**	**+0.9%**
	Europarl	**MGCount**	**+1.5%**

Графики средней точности извлечения терминов

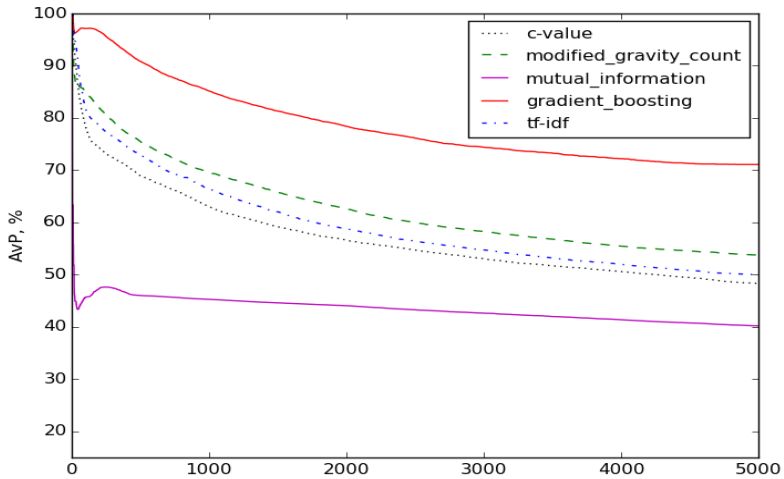
Однословные, банковский корпус



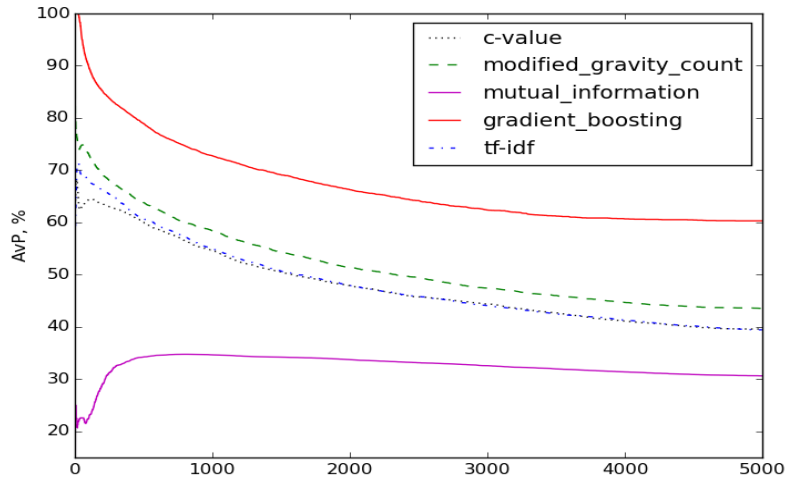
Однословные, корпус Europarl



Двусловные, банковский корпус



Двусловные, корпус Europarl



Вклад тематических признаков в извлечение терминов

- Сравнение моделей
  - Базовая модель без тематических признаков
  - Модель с тематическими признаками
- Комбинирование признаков – метод градиентного бустинга
- Однословные термины

Модель	**Число признаков**	**Банковский корпус**	**Корпус Europarl**
Базовая	54	57.3	58.5
+ тематические	67	**59.0**	**58.7**

- Двусловные термины

Модель	**Число признаков**	**Банковский корпус**	**Корпус Europarl**
Базовая	73	70.8	60.0
+ тематические	86	**71.6**	**60.3**

- Показана статистическая значимость с помощью одностороннего критерия Вилкоксена

Вклад PLSA-SIM в извлечение однословных терминов

- Алгоритм PLSA-SIM + 1000 самых частотных словосочетаний
- Модификация в определении признаков:
  - Замена $P(w|t)$ на $\hat{P}(w|t)$ согласно признаку *C-Value*:

$$\hat{P}(w|t) = P(w|t) - \frac{\sum\limits_{p \in P_w} P(p|t)}{|P_w|}, \text{ где}$$

    - P_w – множество всех добавленных в модель словосочетаний, содержащих слово w
- Сравнение моделей:
  - Базовая модель без признаков по PLSA-SIM
  - Модель с признаками по PLSA-SIM

Модель	**Число признаков**	**Банковский корпус**	**Корпус Europarl**
Базовая	67	59.0	58.7
+ признаки по PLSA-SIM	74	**59.9**	**58.9**

- Показана статистическая значимость результатов с помощью одностороннего критерия Вилкоксена

Примеры извлечённых терминов

- Модель извлечения однословных терминов

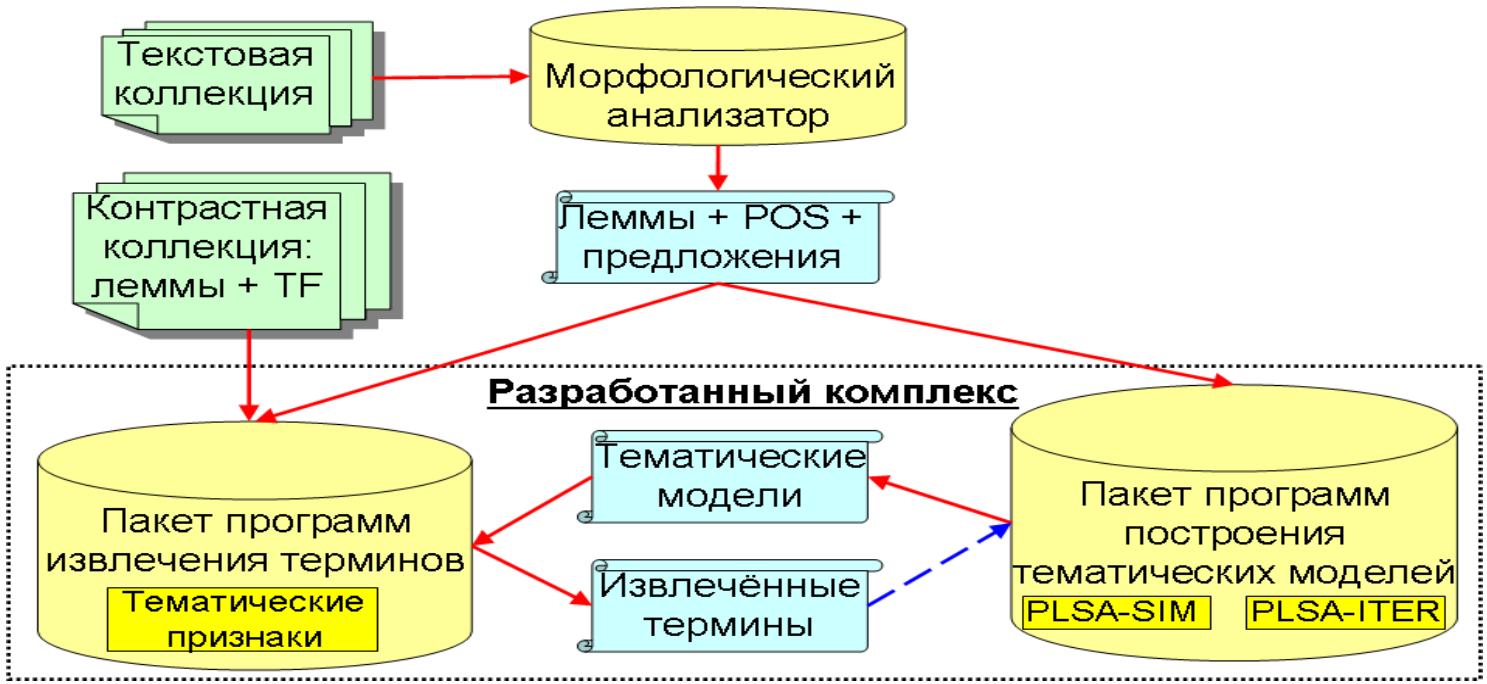
Банковский корпус	Корпус Europarl
Налоговый	**Brazil**
Аудит	**Croatia**
Валютный	**Cuba**
Валюта	**Georgia**
Банковский	**Syria**
Риска	**Iceland**
Страхование	**Taiwan**

- Модель извлечения двусловных терминов

Банковский корпус	Корпус Europarl
Государственный регистрация	**Public service**
Федеральный бюджет	**Health care**
Банковский вклад	**United nation**
Саморегулируемый организация	**Code of conduct**
Финансовый отчётность	**Regional policy**
Доверительный управление	**Rural development**
Иностранный валюта	**Social service**

Система построения тематических моделей на основе лексико-терминологической информации

- В открытом доступе (`https://bitbucket.org/Meister17/dissertation`)



Теорема 2

При условиях $W << N$ и $D << N$ вычислительная сложность алгоритма PLSA-SIM и одной итерации алгоритма PLSA-ITER совпадает с вычислительной сложностью стандартных алгоритмов PLSA и LDA и оказывается равной $O(NTi)$

Основные результаты

- Предложен и реализован новый алгоритм построения тематических моделей, учитывающий словосочетания и улучшающий характеристики качества тематических моделей, включая интерпретацию тем экспертами, что полезно для организации человеко-машинных интерфейсов в информационных системах. Для предложенного метода приведено теоретическое обоснование
- Предложен и реализован новый итеративный алгоритм добавления словосочетаний в тематические модели, улучшающий меру соответствия тематических моделей словам и словосочетаниям текстовых коллекций (перплексию). Для предложенных методов приводятся теоретические оценки вычислительной сложности
- Предложены новые признаки для извлечения терминов, основанные на тематических моделях. Показано, что использование тематической информации улучшает качество извлечения терминов для включения их в базы знаний и терминологические ресурсы
- Разработан и выложен в открытый доступ программный комплекс по построению тематических моделей с использованием лексико-терминологической информации

Публикации по теме доклада (всего 12)

Публикации из списка ВАК:

- Большакова Е.И., Лукашевич Н.В., Нокель М.А. Извлечение однословных терминов из текстовых коллекций на основе методов машинного обучения // Информационные технологии. № 7. 2013. С. 31–37
- Нокель М.А., Лукашевич Н.В. Тематические модели в задаче извлечения однословных терминов // Программная инженерия. № 3. 2014. С. 34–40
- Нокель М.А. Метод учёта структуры биграмм в тематических моделях // Вестник ВГУ, Серия: Системный анализ и информационные технологии. № 4. 2014. С. 89–97
- Нокель М.А., Лукашевич Н.В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами // Вычислительные методы и программирование. Том 16. Выпуск 2. 2015. С. 215–234

Публикации из международной базы цитирования Scopus:

- Bolshakova E., Loukachevitch N., Nokel M. Topic Models Can Improve Domain Term Extraction // ECIR Proceedings. Серия LNCS. Изд. SPRINGER HEIDELBERG. том 7814. 2013. P. 684–687
- Nokel M., Loukachevitch N. Application of topic models to the task of single-word term extraction // CEUR Workshop Proceedings. Vol. 1108. 2013. P. 52–60 (*лучшая работа на конференции RCDL'2013*)
- Nokel M. Topic models: Taking into account similarity between unigrams and bigrams // CEUR Workshop Proceedings. Vol. 1297. 2014. P. 243–252 (*лучшая работа на семинаре молодых учёных RCDL'2014*)

Спасибо за внимание!
Вопросы?