Классификация научных текстов по отраслям знаний *

Сухарева А. В., Царьков С. В.
suhareva_anzhela@mail.ru
Московский физико-технический институт

В этой статье рассматривается задача классификации публикаций, относящихся к различным предметным областям. Классификация выполняется на основе предметного словаря. Разработанный алгоритм применяется для оценивания и сравнения различных методов автоматического выделения терминов. Особенность исследования в том, что была сравнена униграммная классификация на всех словах с n-граммной классификацией. Униграммная модель имеет ряд недостатков, которые устраняются в модели n-грамм. Проведенные нами эксперименты показали, что использование n-грамм позволяет повысить качество классификации научных текстов.
Задача была решена с помощью линейного многоклассового классификатора (на основе наивного байесовского классификатора) с отбором признаков, имеющим линейное по числу объектов и числу признаков время обучения. В экспериментах на задаче классификации научных текстов его качество сравнивается с SVM по униграммным и n-граммным признакам.

Ключевые слова: *наивный байесовский классификатор, AUC, униграммная модель, n-граммная модель, отбор признаков.*

Classification of scientific texts by industry knowledge*

Sukhareva A. V.,[1]***, Tsarkov S. V.***[2]
Moscow Institute of Physics and Technology

In this article we consider the problem of classification of publications that relate to different subject areas. The classification is performed using the dictionary. The algorithm will be used for evaluation and comparison of different methods for automatic selection of terms. The study compared unigram classification on all words with n-gram classification. It is shown that the use of n-grams can improve the quality of classification of scientific texts.
The problem was solved with a linear multi-class classifier (based on Naive Bayes classifier) with feature selection, having linearly on the number of objects and the number of signs of the training. In experiments on the problem of classifying the quality of scientific texts compared with SVM on unigram and n-gram features.

Keywords: *Naive Bayes classifier, AUC, unigram model, n-gram model, feature selection.*

Введение

В настоящее время наблюдается всплеск научных работ, посвященных рубрикации текстов на основе методов машинного обучения [1, 2, 3]. Классификация документов используется, например, в электронных библиотеках научных публикаций для автоматического заполнения метаописаний при поступлении новых документов в библиотеку. Одна из таких коллекций публикаций рассматривается в качестве выборки. Применение методов машинного обучения для классификации текстов очень эффективно при наличии каче-

Работа выполнена при финансовой поддержке РФФИ, проект РФФИ: 14-07-31176. Научный руководитель: Воронцов К. В. Консультант: Царьков С. В.

ственно размеченной обучающей коллекции. В докладах конференции [4] было отмечено, что для больших рубрикаторов (более 500 рубрик) из-за трудности формирования непротиворечивой обучающей выборки единственный работающий подход — трудоемкое ручное описание смысла каждой рубрики. Таким образом, задача разработки эффективных алгоритмов классификации является актуальной.

Рассмотрим один из самых популярных и старейших подхов к классификации — байесовский подход. Байесовский классификатор [5] позволяет определить вероятность принадлежности объекта к одному из классов. При этом выдвигается предположение о независимости влияния на эту вероятность различных атрибутов объектов — так называемое предположение об условной независимости классов, которое существенно упрощает сопутствующие вычисления. Байесовский классификатор относит объект к определенному классу тогда и только тогда, когда апостериорная вероятность принадлежности объекта к этому классу больше апостериорной вероятности принадлежности объекта к любому другому классу.

Байесовский подход к классификации основан на теореме [6, 7], утверждающей, что если плотности распределения каждого из классов известны, то искомый алгоритм можно выписать в явном аналитическом виде. Более того, этот алгоритм оптимален, то есть обладает минимальной вероятностью ошибок.

На практике плотности распределения классов, как правило, не известны. Их приходится оценивать (восстанавливать) по обучающей выборке. В результате байесовский алгоритм перестаёт быть оптимальным, так как восстановить плотность по выборке можно только с некоторой погрешностью. Чем короче выборка, тем выше шансы подогнать распределение под конкретные данные и столкнуться с эффектом переобучения.

Байесовский подход к классификации лежит в основе достаточно удачных алгоритмов классификации. Одним из них является наивный байесовский классификатор.

Наивный байесовский классификатор [6, 7, 8] — простой вероятностный классификатор, основанный на применении Теоремы Байеса со строгими (наивными) предположениями о независимости. В зависимости от точной природы вероятностной модели, наивные байесовские классификаторы могут обучаться очень эффективно. Во многих практических приложениях, для оценки параметров для наивных байесовых моделей используют метод максимального правдоподобия; другими словами, можно работать с наивной байесовской моделью, не веря в байесовскую вероятность и не используя байесовские методы.

Преимуществами наивного байесовского классификатора являются: малое количество данных для обучения, необходимых для оценки параметров, простота реализации и низкие вычислительные затраты при обучении и классификации. В тех редких случаях, когда признаки действительно независимы (или почти независимы), наивный байесовский классификатор (почти) оптимален. Основной его недостаток — относительно низкое качество классификации в большинстве реальных задач.

Результат классификации зависит не только от выбора алгоритма, но и от того, какой набор характеристик используется для составления вектора признаков. Наиболее распространенный способ представления документа в задачах компьюторной лингвистики и поиска — это униграммы и n-граммы. Униграммная модель, так называемый «мешок слов» («bag of words»), наиболее популярная модель представления текстовых документов, которая рассматривает каждый термин в качестве независимой случайной величины вне контекста и связи с другими словами текста. n-граммы получены с помощью алгоритма автоматического выделения ключевых фраз [9] по коллекции текстовых документов. Они образуют лексикон вероятностной тематической модели. В данной работе исследует-

ся качество классификации в зависимости от применения униграмм и n-грамм в качестве признакового описания документов.

Описание данных

Выборка — коллекция из 6707 авторефератов диссертаций, которые представляют собой тексты на русском языке. Часть публикаций была отнесена к обучающей выборке (2469 автореферата), другая часть — к тестовой (4238 автореферата). В задаче рассматривается 22 класса документов. Распределение документов по классам представлено в табл.1, из которой видно, что данная задача является задачей с несбалансированными классами (unbalanced classes), поскольку некоторые классы встречаются гораздо чаще, чем остальные.

Таблица 1. Статистика по данным.

Отрасль наук	**Обучающая выборка**		**Тестовая выборка**	
	кол-во авторефератов	доля в выборке, %	кол-во авторефератов	доля в выборке, %
архитектура	4	0,162	4	0,094
биологические науки	172	6,966	305	7,197
ветеринарные науки	13	0,0,527	30	0,708
географические науки	15	0,608	31	0,731
геолого-минералогические науки	26	1,053	48	1,133
искусствоведение	18	0,729	33	0,779
исторические науки	116	4,698	195	4,601
культурология	9	0,365	19	0,448
медицинские науки	523	21,183	885	20,882
педагогические науки	127	5,144	242	5,710
политические науки	27	1,094	59	1,392
психологические науки	30	1,215	54	1,274
сельскохозяйственные науки	67	2,714	124	2,926
социологические науки	42	1,701	61	1,439
технические науки	353	14,297	613	14,464
фармацевтические науки	7	0,284	16	0,378
физико-математические науки	198	8,019	321	7,574
филологические науки	145	5,873	226	5,333
философские науки	82	3,321	148	3,492
химические науки	66	2,973	111	2,619
экономические науки	350	14,176	577	13,615
юридические науки	79	3,200	136	3,209

В данной работе сравниваются два вида признаков: униграммы и n-граммы. На вход алгоритма подаются словари коллекции текстов, один из них состоит из 139229 униграмм, другой содержит 667566 n-грамм. Униграммы были получены после следующей обработки:

— слова текстов лемматизированы;

— удалены стоп-слова и служебные части речи;
— удалены слова, встречающиеся менее двух раз в документе.

Для автоматического выделения n-грамм в тематических моделях коллекций текстовых документов использовался двухэтапный алгоритм [9]. На первом этапе формируется избыточно большой лексикон из n-грамм, отобранных по морфологическим признакам и статистическим критериям релевантности и устойчивости n-грамм. На втором этапе отбираются n-граммы, наиболее полезные для тематической модели, что позволяет существенно сократить лексикон без ухудшения качества тематической модели.

Постановка задачи

Разработать линейный многоклассовый классификатор с отбором признаков, имеющий линейное по числу объектов и числу признаков время обучения. В качестве признаков используются частоты униграмм и n-грамм. В экспериментах на задаче классификации научных текстов его качество сравнивается с SVM по униграммным и n-граммным признакам.

Введем следующие обозначения:

X – коллекция текстовых документов, состоящая из документов x;

W – словарь коллекции текстов, состоящий из n-грамм w;

Y – конечное множество классов, состоящее из классов y.

Предполагается, что документы $x \in X$ описываются бинарными признаками $(x^1, \ldots, x^n)$:

$$b_w(x) = [f_w(x) \geqslant th], \tag{1}$$

где $b_w(x) \in \{0, 1\}$ — бинарный признак, $f_w(x)$ — частота встречаемости n-граммы w в документе x, th — порог встречаемости n-граммы w.

Задача многоклассовая. Для выбора других классов будем использовать стратегии каждый-против-каждого и каждый-против-всех. Заметим, что для каждого класса важна только небольшая часть признаков, которые позволяют отнести данный документ x к некоторому классу y. Отбор информативных терминов-признаков для каждого класса из множества Y заключается в том, чтобы отсортировать n-граммы по убыванию абсолютного значения весов n-грамм и взять первые $K = \{k_1, \ldots, k_m\}$, где m — количество классификаторов. K — настраиваемый параметр модели. Для определения веса термина зафиксируем один класс y. Вычисляем для каждой n-граммы w $tf(w, y)$ и $tf(w, \overline{y})$ в соответствии со стратегией каждый-против-каждого:

$$tf(w, y) = \frac{\sum\limits_{x \in Y \bigcap w \in x} x}{\sum\limits_{x \in Y} x}, \tag{2}$$

где $tf(w, y)$ — доля документов класса y с признаком w.

$$tf(w, \overline{y}) = \frac{\sum\limits_{x \in \overline{Y} \bigcap w \in x} x}{\sum\limits_{x \in \overline{Y}} x}, \tag{3}$$

где $tf(w, \overline{y})$ — доля документов фиксированного класса $\overline{y}$ с признаком w.

Определяем вес термина w в классе y $wt(w, y)$:

$$wt(w, y) = \sqrt{tf(w, y)} - \sqrt{tf(w, \overline{y})} \tag{4}$$

Аналогично вес термина определяется в стратегии каждый-против-всех, только в качестве объектов класса $\overline{y}$ выступают документы всех остальных классов, кроме документов класса y.

Требуется построить эмпирическую оценку плотности распределения, приближающую неизвестную плотность вероятностного распределения, сгенерировавшего обучающую выборку X^l.

Гипотеза 1. Признаки $x^1, \ldots, x^n$ являются независимыми случайными величинами. Следовательно, функции правдоподобия классов представимы в ввиде

$$p_y(x) = p_{y^1}(x^1) \cdots p_{y^n}(x^n), \tag{5}$$

где $p_{y^j}(x^j)$ — плотность распределения значений j-го признака для класса y.

Алгоритм восстанвления плотности распределения на основе гипотезы 1. называется наивным байесовским классификатором [6, 7]. В работе наивный байесовский классификатор используется для построения линейного многоклассового классификатора:

$$a(x) = \arg\max_y \sum_{w=1}^{K} wt(w, y) x^w, x \in X, y \in Y, \tag{6}$$

где $K = \{k_1, \ldots, k_m\}$ — число информативных признаков, m — число бинарных классификаторов, $wt(w, y)$ — вес признака w в классе y, x^w — признак объекта x с истинным номером w, который существовал у признака w до сортировки признаков.

При стратегии каждый-против-каждого $m = \frac{c(c+1)}{2}$, где c — количество классов, а при стратегии каждый-против-всех $m = c = 22$.

В качестве функционала качества, по которому будет вестись сравнение моделей, используется AUC [6, 7].

Описание алгоритма

Ставится задача восстановления зависимости $y = f(x)$ по точкам обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$.

Дано пространство объектов $X = \mathbb{R}^n$, множество классов $Y = \{1, \ldots, c\}$.

Необходимо найти по обучающей выборке X^l параметр $K = \{k_1, \ldots, k_m\}$ — число информтивных признаков алгоритма классификации (6).

В качестве критерия качества используется AUC. На выходе алгоритма получаем: AUC на контоле в зависимости от параметра K.

Эксперименты показали, что оптимальным порогом встречаемости термина в документе является три, то есть термин должен встретиться в документе не менее трех раз, чтобы мы посчитали, что он там есть. Исходные документы хранятся в виде «мешка слов»(униграммы или n-граммы). Операцией объединения слов публикаций получаем словарь классиификации. С помощью словаря по формуле (1) документы преобразуются в бинарные признаки. В результате векторизации документов получим матрицу «объекты-признаки»: строки — объекты (документы публикаций), столбцы — признаки.

По обучающей выборке X^l вычисляем матрицу весов признаков: вычислить веса $wt(x^j, y)$ по формуле (4) для всех признаков $j = 1 \ldots n$ для каждого класса $y \in Y$. Сортируем признаки по убыванию $|wt(x^j, y)|$.

Алгоритм 1 Построение ROC-кривой и вычисление AUC за $O(l)$.

Вход: выборка X^l, функция $SCORE(x, wt(w, y)) = \sum_{w=1}^{K} wt(w, y)x^w, x \in X, y \in Y$;

Выход: $\{(FPR_i, TPR_i)\}_{i=0}^{l}$, AUC — площадь под ROC-кривой;

```
l_0 := Σ_{i=1}^{l} [y_i = 0]; l_1 := Σ_{i=1}^{l} [y_i = 1];
упорядочить выборку по убыванию SCORE(x, wt(w, y));
(FPR_i, TPR_i) := (0, 0); AUC := 0;
для i := 1, ..., l
    если y_i = 0 то
        сместиться на один шаг вправо:
        FTR_i := FPR_{i-1} + 1/l_0; TPR_i := TPR_{i-1};
        AUC := AUC + 1/l_0 TPR_i;
    иначе
        сместиться на один шаг вверх:
        TPR_i := TPR_{i-1} + 1/l_1; FPR_i := FPR_{i-1};
```

Решение многоклассовой задачи сводится к решению задачи бинарной классификации для каждой пары классов, причем порядок классов в паре не важен в силу симметрии задачи относительно критерия качества AUC. Всего получается m моделей бинарной классификации. На контольной выборке, используя обученные веса и ответы, вычисляем критерий качества классификации AUC и параметр k_i для каждой модели $i \in \{1, \ldots, m\}$, при котором AUC достигает максимума. Вначале положим $k_i = 1$. Вычисляем для каждой модели дискриминантную функцию и критерий качества классификации AUC [10] по алгоритму 6.

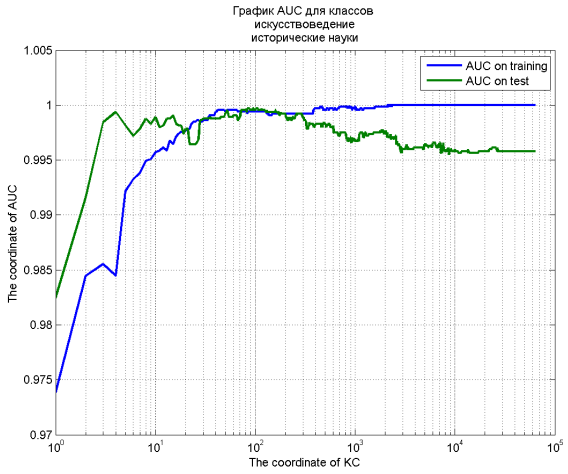
Вычислительный эксперимент

В ходе экспериментов были рассмотрены униграммная и n-граммная модели признакового описания объектов. В качестве стратегии разбиения на классы использовались стратегии каждый-против-всех и каждый-против-каждого. Вычислялся AUC на контрольной выборке в зависимлсти от числа признаков K.

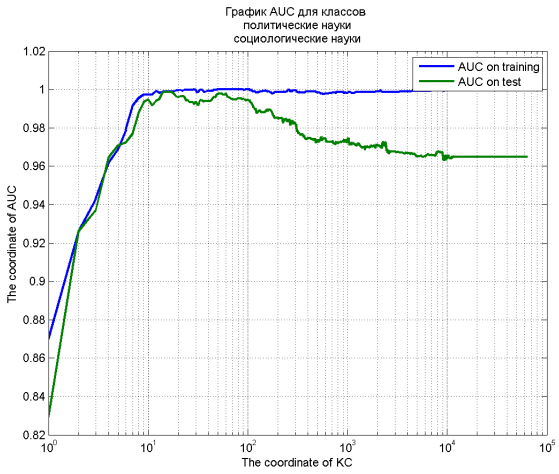
На многих графиках AUC видно, что если брать первые 10 тысяч признаков и более, качество ухудшается (см. Рис.1). Это связано с тем, что слишком много признаков на обучении являются информативными, и в имеющемся объёме данных просто не хватает надёжной статистики, чтобы совсем хорошо отсортировать признаки. В результате мы наблюдаем интересный эффект — как переобучается отбор признаков. Оказывается, что использовать первые 500 признаков — гораздо надёжнее, чем первые 10 тысяч. То есть при таком соотношении числа объектов и числа признаков наивный байесовский классификатор переобучается. При числе признаков меньшем 10 может наблюдаться сильное влияние шума (см. Рис.1(а)), из-за которого критерий качества на конторле лучше, чем на обучении.

Эксперименты показали, что использование n-грамм в качестве признаков наивного байесовского классификатора приводит к улучшению качества классификации (см. Рис.2 — Рис.6). Сравнение построенного классификатора NB с использованием метода top-K с SVM приводится в таблице (см. Рис.7). При представлении многоклассового классификатора как совокупность бинарных классификаторов возникает необходимость вычислить агрегированный показатель качества классификации, объединяющий показатели отдель-

ных классификаторов. Для этого существует два метода. При макроусреднении (macro average) вычисляется взвешенное среднее значение по классам. При микроусреднении (micro average) объединяются решения на уровне документов по всем классам.

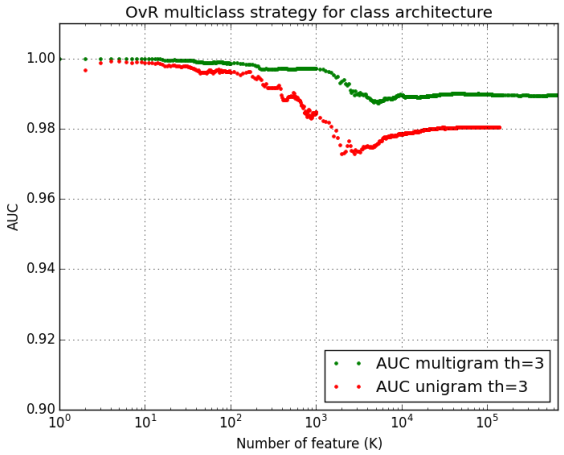


(a) Сильное влияние шума при небольшом числе признаков.



(b) Шум не влияет на критерий качества.

Рис. 1. Переобучение отбора признаков в униграммной модели.



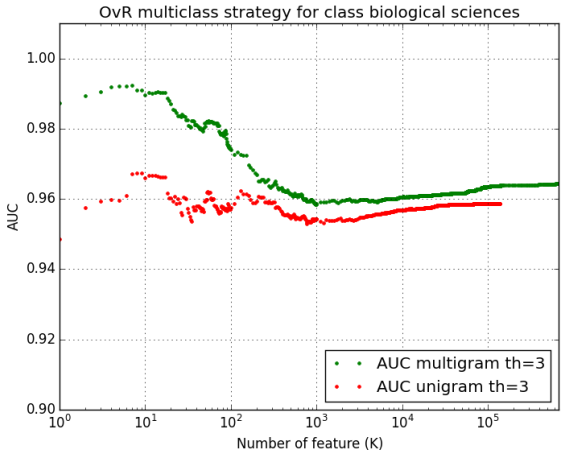


Рис. 2. Сравнение униграммной и n-граммной моделей.

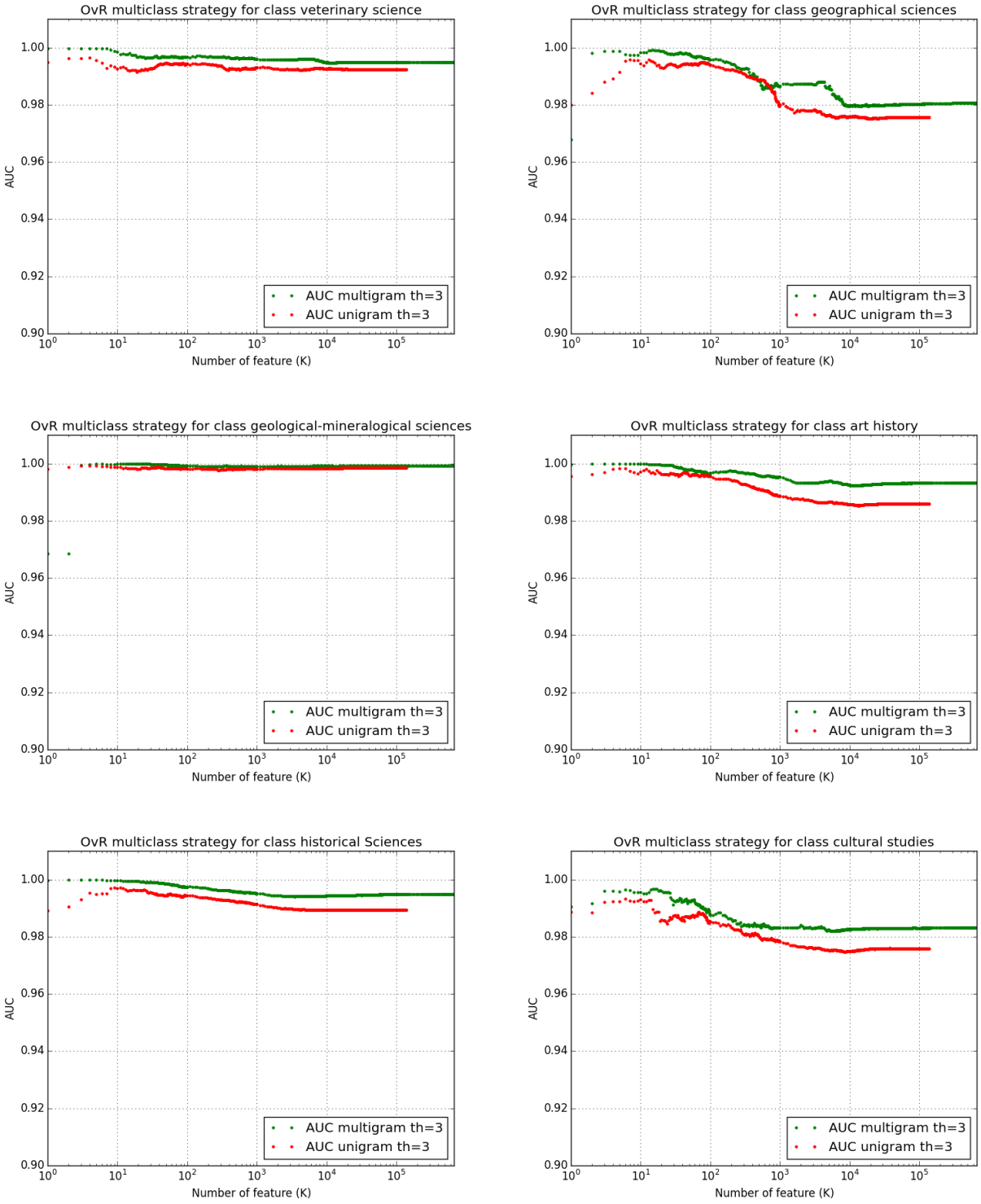


Рис. 3. Сравнение униграммной и n-граммной моделей.

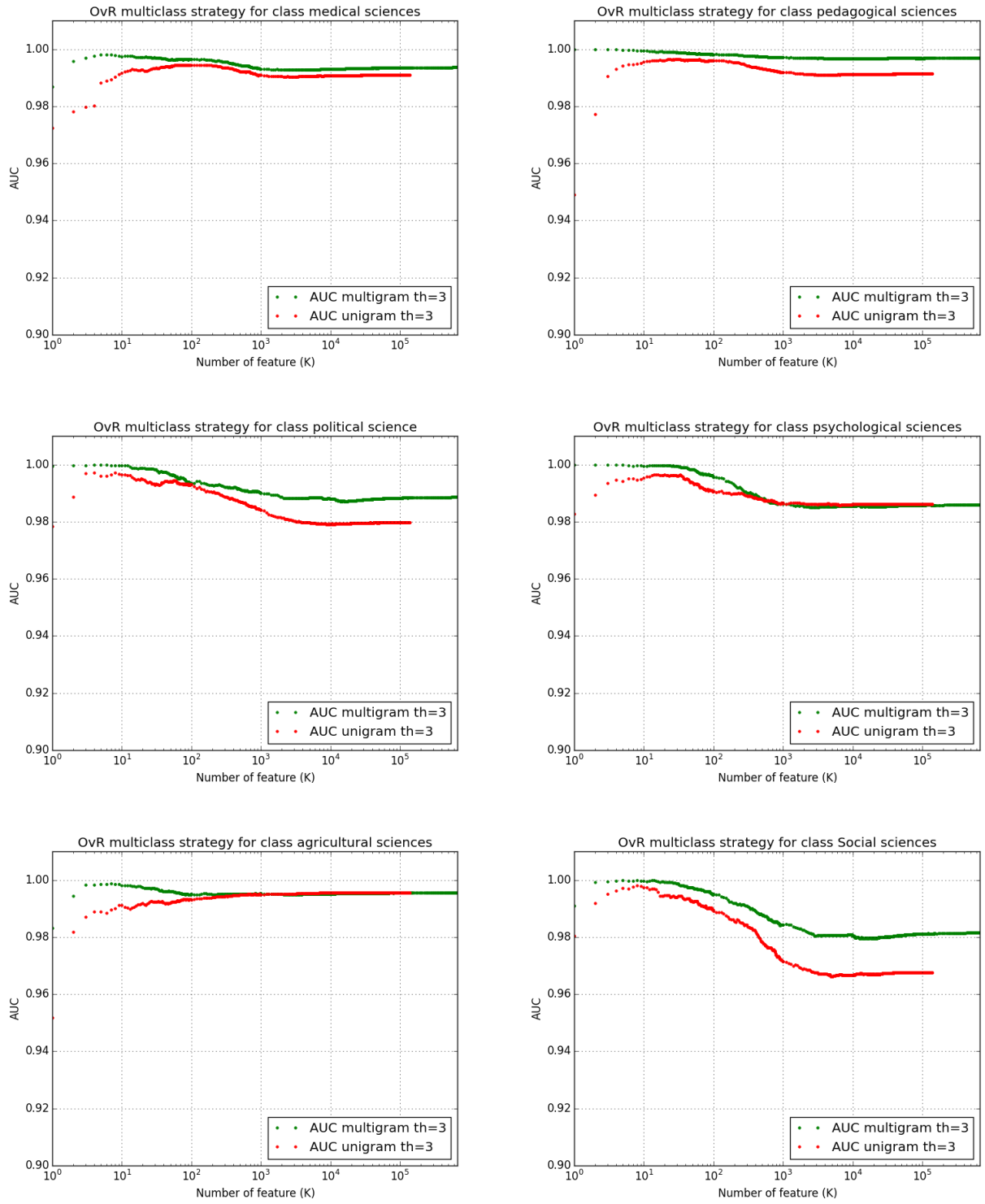


Рис. 4. Сравнение униграммной и n-граммной моделей.

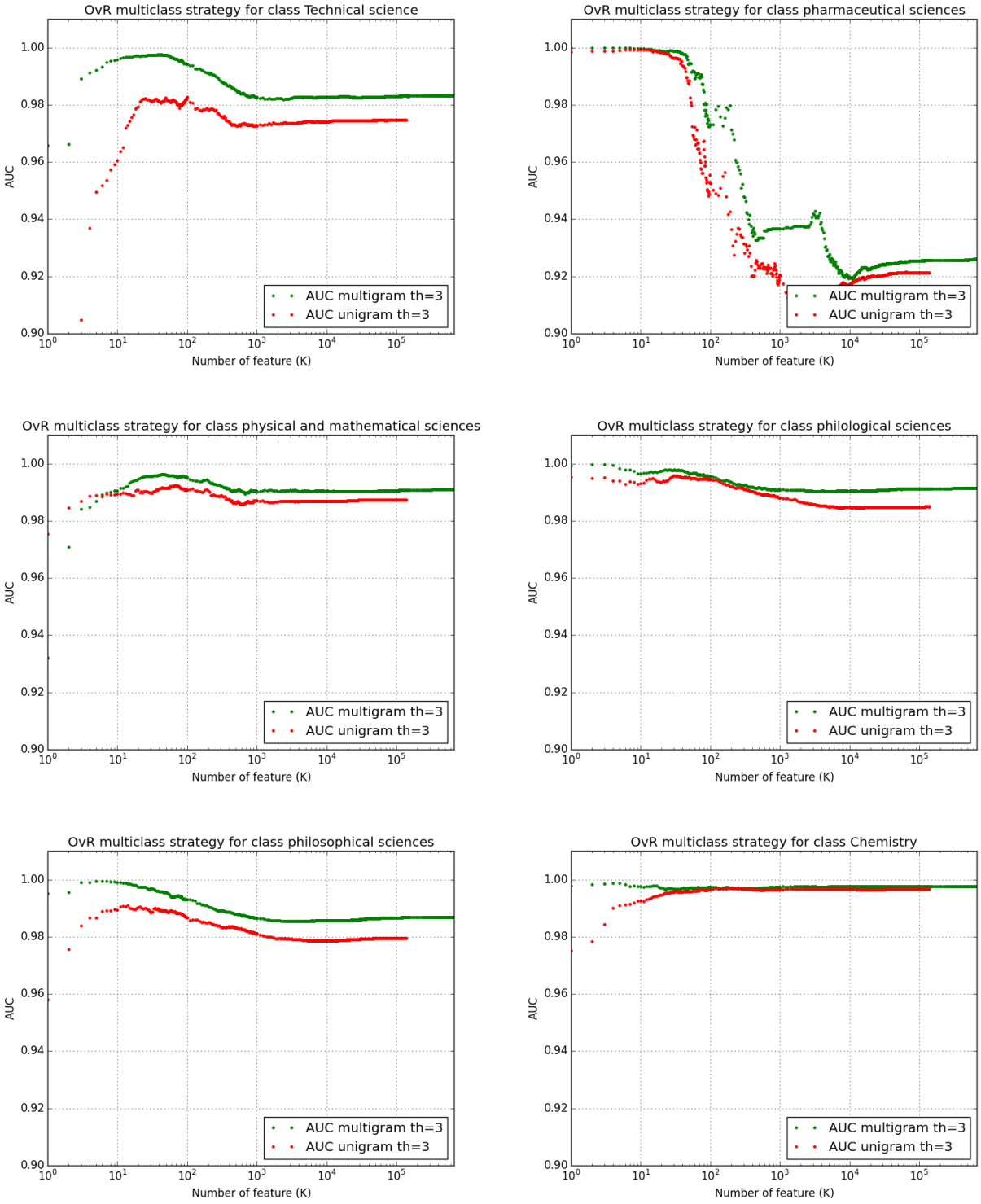


Рис. 5. Сравнение униграммной и n-граммной моделей.

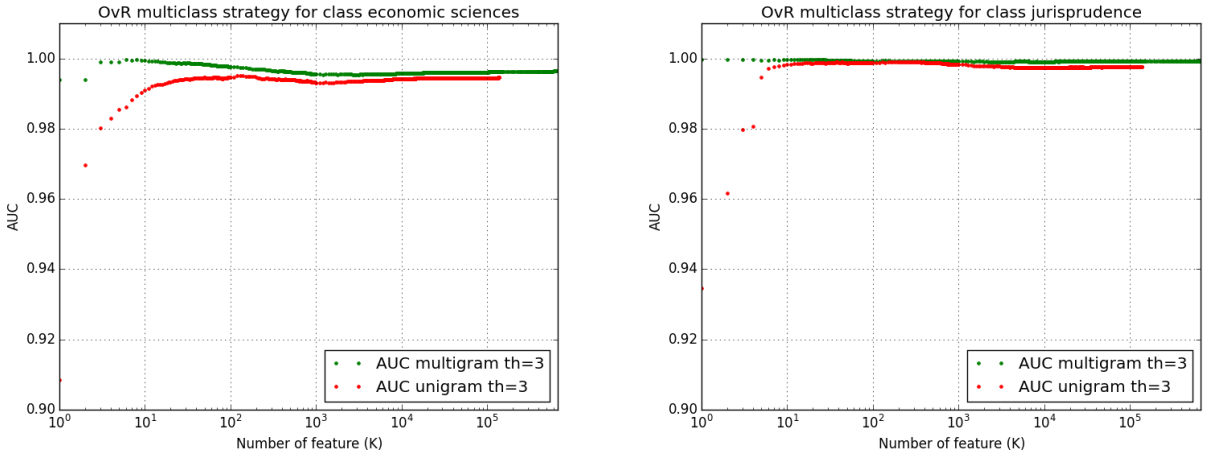


Рис. 6. Сравнение униграммной и n-граммной моделей.

Классы	n-gram		unigram	
	NB_topK	SVM	NB_topK	SVM
архитектура	1,0000	0,9991	0,9993	0,9984
биологические науки	0,9923	0,9907	0,9584	0,9791
ветеринарные науки	0,9998	0,9984	0,9972	0,9980
географические науки	0,9993	0,9958	0,9914	0,9946
геолого-минералогические науки	0,9999	0,9996	0,9988	0,9991
искусствоведение	1,0000	0,9998	1,0000	0,9998
исторические науки	0,9999	0,9998	0,9953	0,9995
культурология	0,9967	0,9907	0,9848	0,9769
медицинские науки	0,9980	0,9982	0,9938	0,9962
педагогические науки	0,9999	0,9995	0,9947	0,9990
политические науки	0,9999	0,9976	0,9939	0,9953
психологические науки	1,0000	0,9994	0,9955	0,9992
сельскохозяйственные науки	0,9987	0,9979	0,9926	0,9960
социологические науки	0,9999	0,9993	0,9973	0,9993
технические науки	0,9975	0,9988	0,9844	0,9961
фармацевтические науки	1,0000	0,9987	0,9708	0,9976
физико-математические науки	0,9963	0,9988	0,9920	0,9985
филологические науки	0,9994	0,9997	0,9969	0,9997
философские науки	0,9994	0,9987	0,9898	0,9957
химические науки	0,9988	0,9988	0,9943	0,9977
экономические науки	0,9996	0,9997	0,9951	0,9988
юридические науки	0,9998	0,9999	0,9984	0,9998
macro average	0,9989	0,9981	0,9916	0,9961
micro average	0,9937	0,9985	0,9668	0,9972

Рис. 7. Сравнение NB с SVM по униграммным и n-граммным признакам.

При стратегии каждый-против-всех всего методом top-K было отобрано 223 информативных n-граммы, из них уникальных — 200. Значения параметра $K = \{k_1, ..., k_m\}$

находятся в диапазоне от 5 до 48 признаков для различных классов. Приведем список отобранных признаков для некоторых классов:

- класс биологические науки:
  1. биологический наука 0.89
  2. биологический 0.62
  3. диссертационный исследование -0.54
  4. идея -0.52
  5. биология 0.51
  6. сущность -0.49
  7. власть -0.47
  8. общественный -0.47
- класс исторические науки:
  1. доктор исторический наука 1.00
  2. исторический наука 0.93
  3. историография 0.82
  4. хронологический 0.76
  5. историк 0.71
- класс филологические науки:
  1. наука филологический 0.96
  2. филологический 0.90
  3. филология 0.69
  4. слово 0.67
  5. языковой 0.65
- класс юридические науки:
  1. наука юридический 0.95
  2. правовой регулирование 0.78
  3. акт правовой 0.78
  4. юридический 0.76
  5. законодатель 0.76
  6. акт нормативный 0.74
  7. государство право 0.73
  8. судебный 0.73
  9. норма право 0.73
  10. законодательство 0.73
  11. конституционный 0.72
  12. правоотношение 0.72
  13. правоприменительный 0.71

Заключение

В работе представлены эксперименты по сравнению униграммной и n-граммной моделей классификации. Эксперименты, проведённые на русскоязычных авторефератах диссертаций, показывают, что n-граммы дают лучшие результаты, чем, униграммы по критерию качества AUC. Модель «мешка слов», в которой каждый документ рассматривается как набор встречающихся в нем слов, имеет ряд недостатков. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов в документах друг от друга. На данный момент проведено множество исследований, посвященных изучению вопроса добавления словосочетаний, n-грамм и многословных терминов в тематические

модели. Однако часто это приводит к ухудшению качества модели в связи с увеличением размера словаря или к значительному усложнению модели [11, 12, 13]. В данном случае быстро и эффективно работает наивный байесовский классификатор.

В статье рассматривается линейный многоклассовый классификатор (на основе наивного байесовского классификатора) с отбором признаков, имеющим линейное по числу объектов и числу признаков время обучения. Проведены эксперименты по сравнению построенного классификатора с SVM по униграммным и n-граммным признакам на задаче классификации научных текстов.

Литература

[1] Гринева М., Гринев М., and Лизоркин Д. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов. In *Труды Института системного программирования РАН*, 2009.

[2] Aizawa A. An information-theoretic perspective of tf–idf measures. *Information Processing and Management: an International Journal*, page 45 – 65, 2003.

[3] Браславский П. and Соколов Е. Сравнение пяти методов извлечения терминов произвольной длины. In *Компьютерная лингвистика и интеллектуальные технологии: ежегодная Международная конференция «Диалог» (Бекасово, 4–8 июня 2008 г.)*, Вып. 7 (14), pages 67–74, М.: РГГУ, 2008.

[4] Harding S.M. The INQUERY Retrieval System Callan J.P., Croft W.B. Proceedings of dexa-92, 3rd international conference on database and expert systems applications. pages 78–83, 1992.

[5] Pat Langley George H. John. Estimating continuous distributions in bayesian classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

[6] К.В. Воронцов. Курс лекций К.В. Воронцова. Машинное обучение., 2011.

[7] Hastie T., Tibshirani R., and Friedman J. *The Elements of Statistical Learning*. Springer, 2001.

[8] Енюков И. С. Мешалкин Л. Д. Айвазян С. А., Бухштабер В. М. *Прикладная статистика: классификация и снижение размерности*. Финансы и статистика, М., 1989.

[9] Царьков С.В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов. *Естественные и технические науки №6(62).*, pages С. 456–464., 2012.

[10] К.В. Воронцов. Задача диагностики заболеваний по электрокардиограмме, 2014.

[11] H. Wallach. Topic modeling: beyond bagofwords. In *In the Proceedings of the 23rd International Conference on Machine Learning*, page pp. 977–984, 2006.

[12] A. McCallum X. Wang and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *In the Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, page pp. 697–702, 2007.

[13] . H. Wallach, pp. 977–984. Topic modeling: beyond bagofwords. *In the Proceedings of the 23rd International Conference on Machine Learning*, 2006.