# Оценивание качества выделения терминов в задаче классификации текстовых документов

Сухарева Анжелика
Научный руководитель: д.ф.-м.н. К. В. Воронцов

Московский физико-технический институт
(государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2016 г.

# Постановка задачи

**Задача классификации**:
$X \in R^n$ — коллекция текстовых документов,
$Y = \{1, \ldots, C\}$ – множество классов.
Документы хранятся в виде «мешка слов» и описываются бинарными признаками:

$$b_w(x) = [f_w(x) \geq th],$$

где $f_w(x)$ — частота встречаемости $n$-граммы в документе,
$th$ — порог встречаемости $n$-граммы.
**Найти** зависимости $y = f(x)$ по точкам обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$.
**Критерий качества**: $AUC, MAUC$.

# Задача выделения терминов

**Задача выделения терминов** (*Term Extraction*): по коллекции текстовых документов сформировать лексикон коллекции.
В работе исследуются $n$-граммы, построенные одним из алгоритмов *Term Extraction*:

- на первом этапе формируется избыточный словарь, затем $n$ - граммы отбираются из слов предложений текста **на основе морфологических и синтаксических правил**;
- второй этап (**статистический**) — автоматическое выделения ключевых фраз без привлечения внешней информации. Методы: **TF-IDF**, **Termhood**.

# Цели исследования

- **Цель данного исследования**: разработать способы измерения качества выделения терминов в задачах классификации текстов.
- **Проблемы исследования**:
  - Как качество выделения терминов влияет на качество классификации?
  - Как построить чувствительный критерий качества выделения терминов?
- **Решение**: строить как можно более точные модели классификации и, измеряя их качество, тем самым измерять качество мультиграммных словарей терминов.

# Наивный байесовский классификатор

Оптимальный байесовский классификатор:

$$a(x) = \arg\max_{y \in Y} P(y)p(x|y),$$

где $P(y)$ — вероятности появления объектов каждого из классов, $p(x|y) = p(x;\theta_y)$ — функции правдоподобия классов. NB основан на гипотезе независимости признаков.

## Гипотеза

Если признаки $x^1, \ldots, x^n$ являются независимыми случайными величинами, то

$$p(x|y) = p(x^1, x^2, \ldots, x^n|y) = p(x^1|y) \cdots p(x^n|y),$$

где $p(x^j|y) = p(x^j;\theta_y^j)$ — плотность распределения значений $j$-го признака для класса $y$.

# Экспоненциальное семейство распределений

Распределение $p(x)$ из **экспоненциального семейства распределений**, если его плотность может быть представлена в виде:

$$p(x|\theta,\varphi) = \exp\left(\frac{x\theta - c(\theta)}{a(\varphi)} + h(x,\varphi)\right),$$

где $c(\theta)$, $h(x,\varphi)$, $a(\varphi)$ — функциональные параметры распределения,
$\theta$ и $\varphi$ — числовые параметры,
$\theta$ — параметр *сдвига*,
$\varphi$ — параметр *разброса*.
Обозначим *среднее значение* $j$ признака в классе $y$ как

$$\langle x_i^j \rangle_y = \frac{1}{|X_y|} \sum_{x_i \in X_y} x_i^j.$$

# Наивный линейный байесовский классификатор

Согласно гипотезе независимости признаков **принцип максимума логарифма правдоподобия** принимает вид:

$$\sum_{j=1}^{n} \sum_{y \in Y} \sum_{x_i \in X_y} \ln p(x^j; \theta_y^j) \to \max_{\theta_y^j}.$$

## Теорема К. В. Воронцова

Если одномерные плотности $p(x^j, \theta_j^y)$ принадлежат экспоненциальному семейству распределений и $\Theta = (\theta_j^y)$ является точкой максимума правдоподобия, то

$$\theta_y^j = [c']^{-1}(\langle x_i^j \rangle_y).$$

# Параметры NB с отбором признаков

**Найти**:

1. $w_y^j$ — вес $j$ признака в классе $y$;
2. $K = \{k_1, \ldots, k_m\}$ — число информативных признаков алгоритма классификации.

**Отбор признаков**: метод *Top-K*.
Пусть $Y = \{+1, -1\}$. Веса признаков:

$$w_y^j = \begin{cases} \sqrt{\langle x_i^j \rangle_{+1}} - \sqrt{\langle x_i^j \rangle_{-1}} & , w_y^j > 0 \\ 0 & , \text{в противном случае} \end{cases} \quad (1)$$

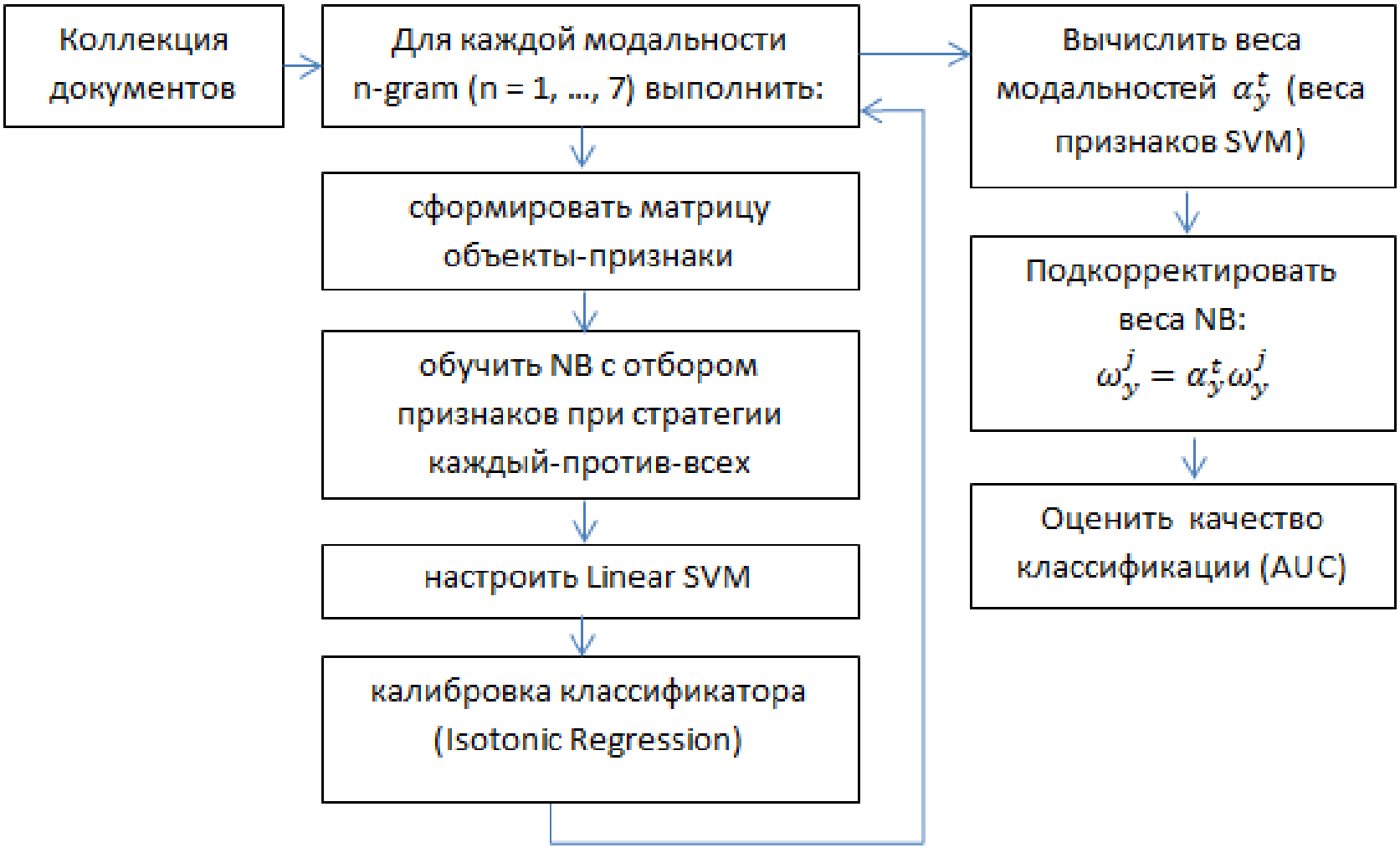
Соответствует распределениям из экспоненциального семейства с $\varphi^j = \varphi = 1$, $a(\varphi) = 1$, $h(x, \varphi) = 0$, $\theta = \sqrt{\mu}$, $c(\theta) = \frac{\theta^3}{3}$.

# Композиция классификаторов

Композиция позволяет точнее настраивать веса NB за счет учета влияния модальности $n$-грамм.
Мультиграммы получены на 1 этапе алгоритма.
Лучшее качество классификации среди композиций у **откалиброванной композиции NB классификаторов**.

| Метод | NB | Композиция | | |
|---|---|---|---|---|
| | | NB+wt SVM | ANN | NB+IR+wt SVM |
| macro average AUC | 0,749 | 0,865 | 0,876 | 0,887 |
| micro average AUC | 0,733 | 0,873 | 0,880 | 0,880 |

# Алгоритм построения откалиброванной композиции



# Тематическая модель классификации

Тематическая модель появления слов в документах:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

Тематическая модель классификации документов:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}$$

где $c$ — класс, $w$ — слово, $t$ — тема, $d$ — документ коллекции.
**Задача максимизации логарифма мультимодального регуляризованного правдоподобия**:

$$\sum_{m,d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \tau \sum_{d,c} m_{dc} \ln \sum_t \psi_{ct}\theta_{td} + R(\Phi, \Theta, \Psi) \to \max_{\Phi,\Theta,\Psi}$$

где $w \in W^m$, $W^m$ — словарь терминов модальности $m$, $m \in M$.

# Аддитивная регуляризация

Аддитивная регуляризация тематической модели (ARTM):

$$R(\Phi, \Theta, \Psi) = \sum_i \tau_i R_i(\Phi, \Theta, \Psi)$$

- разреживание $\Theta$:

$$R(\Theta) = -\tau \sum_{d \in D} \sum_{t \in T} \ln \theta_{td} \to \max$$

- сглаживание $\Phi$:

$$R(\Phi) = \tau \sum_{t \in T} \sum_{w \in W} \ln \phi_{wt} \to \max$$

- декорреляция тем как столбцов $\Phi$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max$$

где $S$ — предметные темы, $S \subset T$.

# Выборка

**Два типа коллекций**:

- коллекции с целыми авторефератами;
- коллекции с фрагментами авторефератов.

Эксперименты с фрагментами авторефератов проводились как на всей выборке, так и на подвыборке.

**Описание подвыборки**:

$X^l$ — обучающая выборка, $|X^l| = 30000$

$X^k$ — контрольная выборка, $|X^k| = 5000$

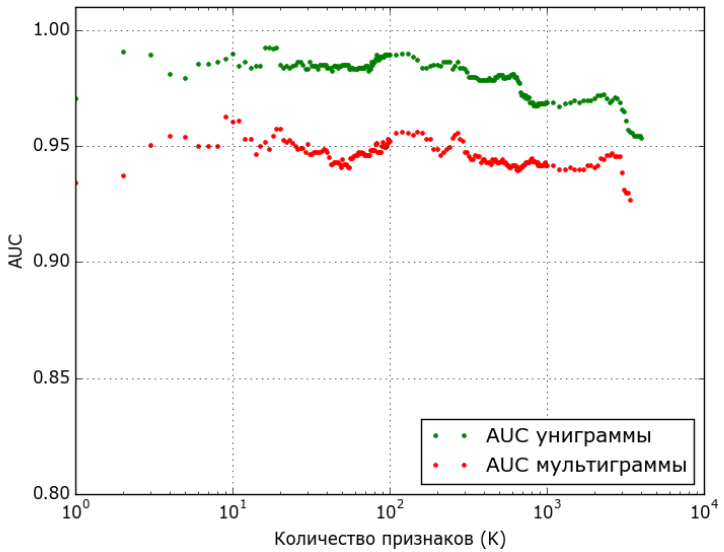
$x = (x^1, \ldots, x^n)$, $n \geq 1$.

# Сравнение униграммной и $n$-граммной моделей

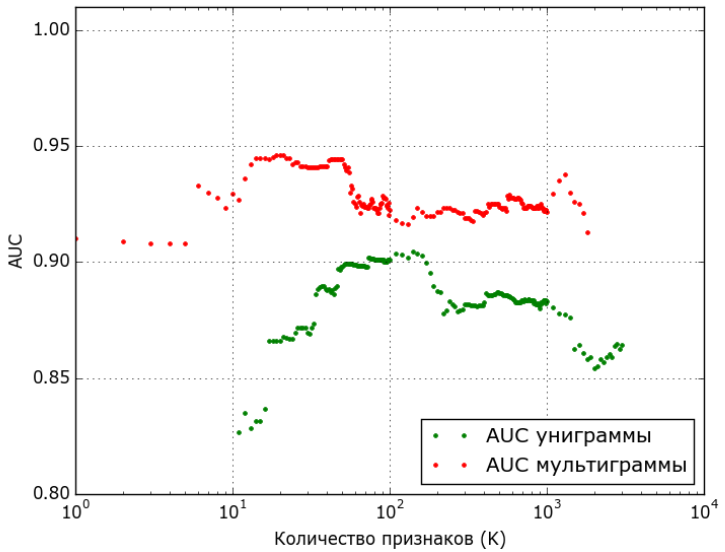
Эксперименты проводились на **целых авторефератах**.

Мультиграммы получены на 1 этапе алгоритма.

**NB стратегия каждый-против-каждого**



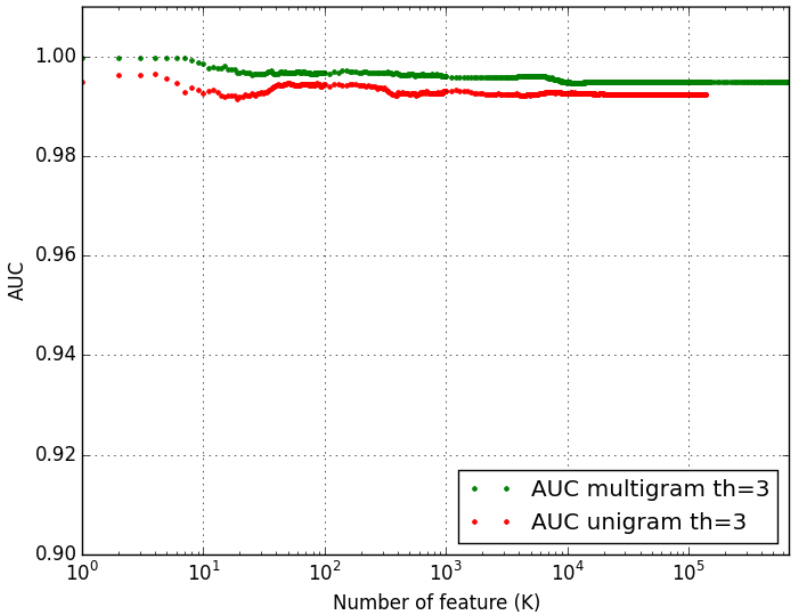

(a) Географические науки против геолого- минералогических.



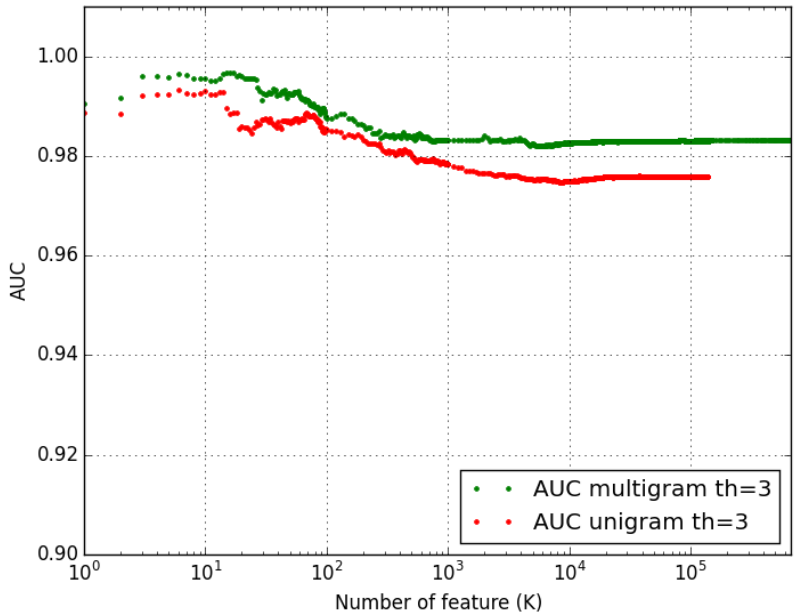

(b) Культурология против философских наук.

# Сравнение униграммной и $n$-граммной моделей

## NB стратегия один-против-всех



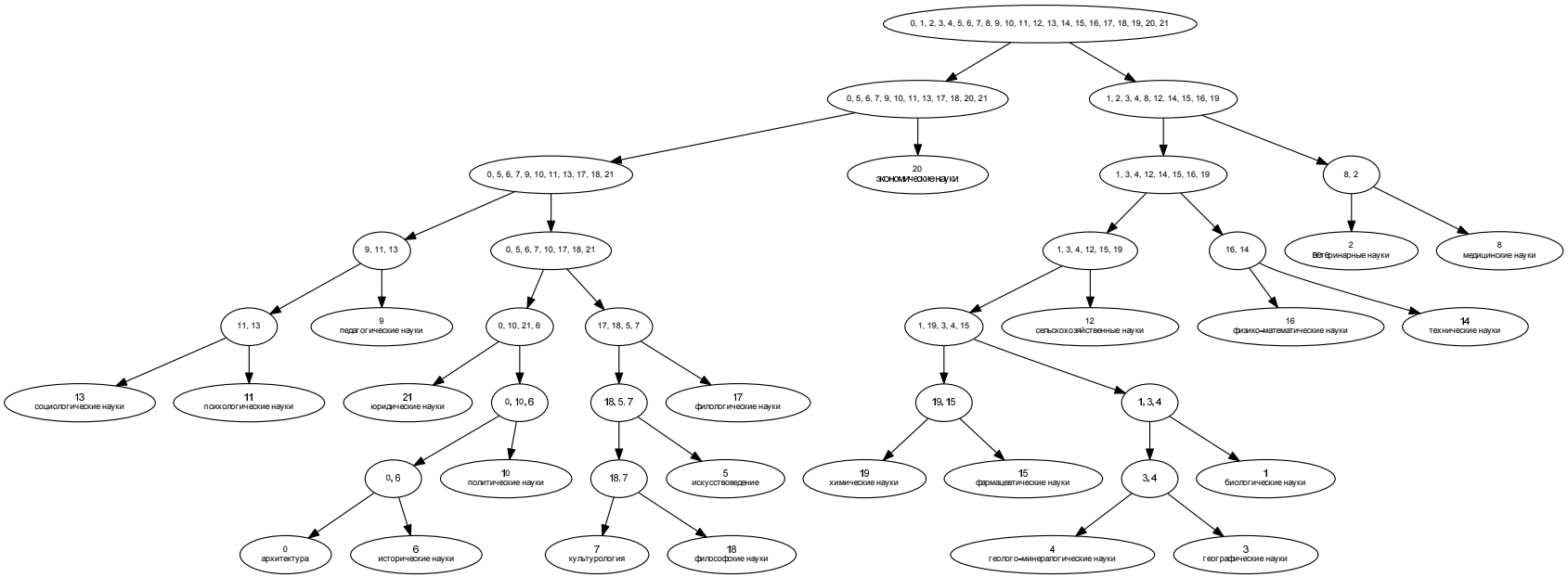

(c) Класс ветеринарные науки.




(d) Класс культурология.

# Сравнение униграммной и $n$-граммной моделей

## NB иерархическая стратегия




Для более чувствительной оценки методов выделения терминов было решено классифицировать фрагменты авторефератов.

# SVM, ARTM и композиция NB

Лучшее качество классификации фрагментов авторефератов показала тематическая модель классификации, построенная с помощью подхода ARTM и проинициализированная признаками, отобранными откалиброванной композицией NB.

macro average AUC

| Метод | NB | Композиция NB | SVM | ARTM | ARTM (иниц. комп. NB) |
|---|---|---|---|---|---|
| 1 этап | 0,7842 | 0,8865 | 0,9276 | 0,9415 | 0,9914 |
| TF-IDF | 0,7444 | 0,8382 | 0,8795 | 0,9360 | 0,9902 |
| Termhood | 0,8003 | 0,8941 | 0,9090 | 0,9496 | 0,9987 |

micro average AUC

| Метод | NB | Композиция NB | SVM | ARTM | ARTM (иниц. комп. NB) |
|---|---|---|---|---|---|
| 1 этап | 0,6756 | 0,8477 | 0,9461 | 0,9554 | 0,9898 |
| TF-IDF | 0,6788 | 0,8412 | 0,9139 | 0,9479 | 0,9970 |
| Termhood | 0,6334 | 0,8597 | 0,9321 | 0,9488 | 0,9990 |

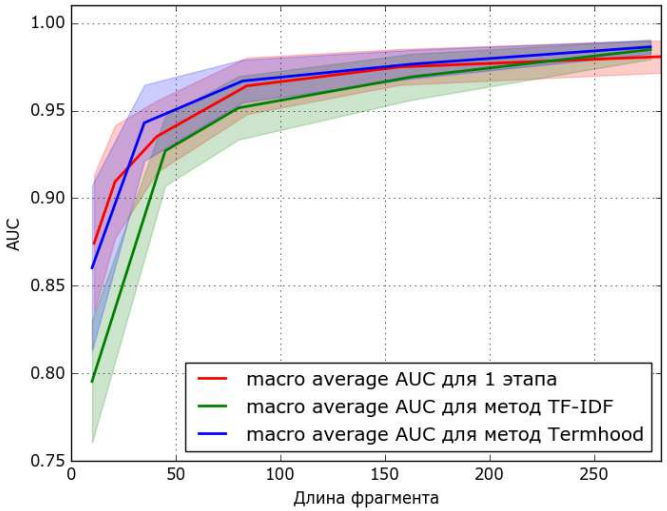
Рис.: Сравнение моделей классификации на подвыборке объектов.

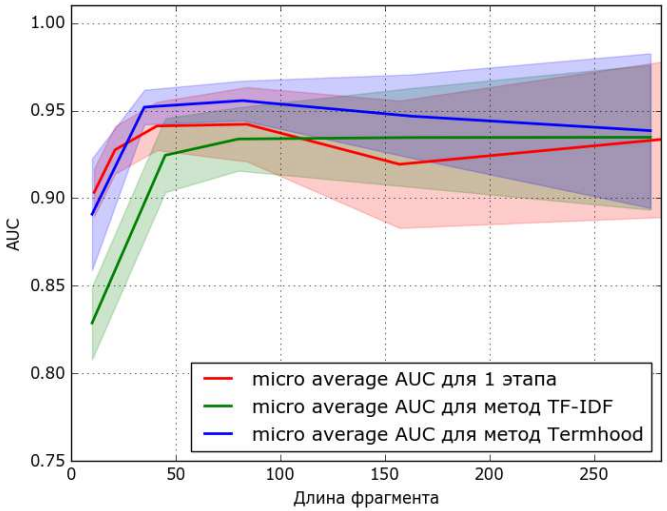
# Композиция NB

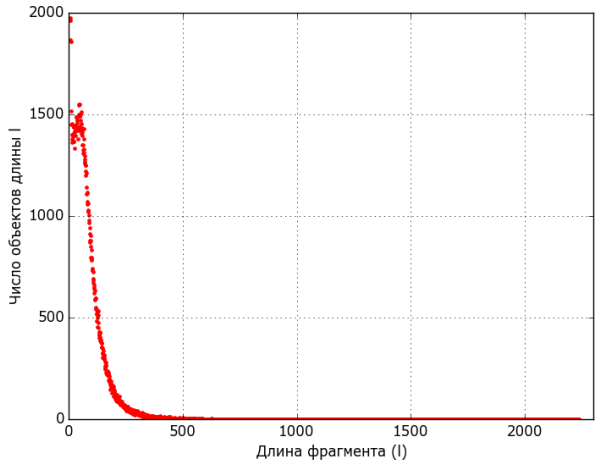
Оценка качества выделения терминов фрагментов с помощью композиции NB:

| Метод | 1 этап | TF-IDF | Termhood |
|---|---|---|---|
| macro average AUC | 0,8792 | 0,8267 | 0,9018 |
| micro average AUC | 0,8526 | 0,8468 | 0,8773 |

# Зависимость *AUC* от длины фрагмента



# Результаты, выносимые на защиту

- Предложен критерий для оценивания и сравнения алгоритмов выделения терминов, основанный на качестве классификации композиции NB.
- Проведены вычислительные эксперименты по сравнению униграммной и мультиграммной моделей классификации.
- Выполнена программная реализация алгоритмов классификации: NB, композиция NB, тематическая модель классификации (с помощью библиотеки BigARTM).