# *Term frequency*, референтный текстовый корпус и оценивание близости коротких текстов смыслосому эталону

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

14-я Международная конференция
«Интеллектуализация обработки информации» (ИОИ-2022),

6–9 декабря 2022 г.

г. Москва

## Основные требования к представительной (референтной) коллекции

1. Максимально полное раскрытие интересующей пользователя темы в каждом тексте формируемой подборки.
2. Тексты подборки (коллекции) должны быть максимально релевантны заданной предметной области как по лексическому составу, так и по внутритекстовым связям (синтаксическим, семантическим и т. п.).
3. Максимум среднего числа наиболее значимых терминов в расчёте на одно простое распространённое предложение (фразу) при минимуме его длины (в словах), что соответствует *эталонной* передаче смысла.

## Минимизация ручного труда эксперта: основные идеи

1. Использование экспертом коротких текстов, которые сопоставлялись бы по лексическому составу и (возможно) по связям между словами с документами, добавляемыми в референтную коллекцию.
2. В их роли могут быть аннотации научных статей или иные тексты, резюмирующие значимые факты заданной предметной области.
3. Отбор в референтную коллекцию — задача, обратная *абстрактивной суммаризации*: нужно найти текст, в котором описанные в аннотации (коллекции аннотаций) общие идеи отражены наиболее полно.

## «Классическая» постановка задачи [Еремеев М. А., 2019]:

1. Для каждого уровня языка определяется свой *алфавит токенов*. Например, для лексического уровня токенами будут слова, для синтаксического — типы и длины синтаксических связей.
2. Частота токена считается аномально высокой, если она превышает *95%-й квантиль* его частоты в референтном корпусе текстов, не являющихся сложными для выбранной аудитории читателей.
3. Предполагается, что *95% токенов* в референтном корпусе на каждом из языковых уровней *не превышают* своей *фиксированной частоты*, которая определяется экспериментальным путём (по результатам нейрофизиологических и психофизиологических исследований).

Квантиль — некоторое значение, которое исследуемая случайная величина не превышает с фиксированной вероятностью.

## Отбор документов в референтный корпус на основе коллекции аннотаций

Поскольку здесь речь идёт о *минимально необходимой* представленности слов (терминов) из аннотаций в анализируемом документе, то логично предположить, что в данной задаче следует рассматривать *5%-й квантиль* частотной характеристики слова относительно заданного документа.

# Выбор частотной характеристики

Основное требование — независимость от числа слов документа.

Будем для каждой фразы в составе каждой аннотации вычислять долю ненулевых значений *TF-меры* для входящих во фразу слов относительно анализируемого документа.

*TF-мера* оценивает важность слова $t_i$ в пределах отдельного документа $d$ и определяется как

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

где $n_i$ — число вхождений слова $t_i$ в документ $d$,

а в знаменателе — общее число слов в документе.

**Замечания**

- одна фраза здесь соответствует простому распространённому предложению (в терминологии теории «Смысл⇔Текст»);
- при этом допускается, что одна и та же фраза может встречаться в нескольких аннотациях коллекции (например, если это статьи одного и того же автора);
- в любом случае каждая фраза принимается к рассмотрению только один раз;
- использование именно доли ненулевых значений TF-меры, а не самих значений *term frequency* для слов фразы, позволяет решить проблему зависимости оценки значимости документа от числа слов в нём;
- важно лишь присутствие максимального числа слов из аннотаций в анализируемом документе, частота же отдельных слов здесь не принципиальна.

# Оценка важности документа для включения в референтную коллекцию

Пусть $d$ — документ, оцениваемый на предмет включения в референтную коллекцию (корпус). Для каждого слова $w$ в составе каждой фразы $Ts$ каждой аннотации из сформированной экспертом коллекции вычисляется значение TF-меры относительно документа $d$, $\mathrm{tf}\,(w, d)$. При этом доля ненулевых значений TF-меры по фразе $Ts$ формально определяется как

$$c(Ts, d) = \frac{\left|w\colon\ (w \in Ts) \wedge (\mathrm{tf}\,(w, d) > 0)\right|}{\left|w\colon w \in Ts\right|}. \tag{2}$$

Обозначим 5%-й квантиль эмпирического распределения величины (2) по документу $d$ для заданной коллекции аннотаций $\mathbb{T}\mathrm{s}$ как $C_5\,(\mathbb{T}\mathrm{s}, d)$.

Сама $\mathbb{T}\mathrm{s}$ при этом есть объединение множеств фраз для отдельных аннотаций.

Отсортируем документы-кандидаты на включение в референтный корпус по убыванию $C_5\,(\mathbb{T}\mathrm{s}, d)$. Пусть $d_{\max}$ — документ с максимальным по множеству документов-кандидатов $D$ значением $C_5$ для фраз аннотаций из $\mathbb{T}\mathrm{s}$.

Введём для каждого $d \in D$ вектор значений квантилей

$$\overline{V}\,(\mathbb{T}\mathrm{s}, d) = \Big(C_\gamma\,(\mathbb{T}\mathrm{s}, d)\Big)_{\gamma \in [5,10,20,25,30,40,50,60,70,75,80,90,95]}, \tag{3}$$

куда помимо упомянутых выше 5-го и 95-го перцентилей войдут децили, а также первый и третий квартили.

# Оценка важности документа для включения в референтную коллекцию

Пусть $\overline{V}(\mathbb{T}\mathrm{s}, d_{\max})$ — вектор вида (3) для документа $d_{\max}$.

Обозначим последовательность векторов вида (3) по коллекции $\mathbb{T}\mathrm{s}$ для документов $d_j \in D\colon d_j \neq d_{\max}$, отсортированную по убыванию величины Евклидова расстояния до $\overline{V}(\mathbb{T}\mathrm{s}, d_{\max})$, как $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$.

Разобьём последовательность $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$ на кластеры $H_1, \ldots, H_r$ с применением алгоритма, близкого алгоритмам класса FOREL. При этом кластер $H_r$ будет отвечать документам с наименьшим расстоянием до $d_{\max}$.

**Утверждение 1**

Наибольшую значимость для целевой коллекции будут иметь документы $d \in D$, отнесённые к кластеру $H_r$, плюс сам документ $d_{\max}$.

**Замечания**

1. Описанную выше классификацию документов $d \in D$ следует провести независимо по нескольким коллекциям аннотаций статей близких тематик в целях повышения полноты *(recall)* поиска документов, значимых для референтного корпуса.
2. *Полнота поиска* здесь определяется отношением числа документов, отвечающих условию *Утверждения 1* и признанных экспертом значимыми для референтного корпуса, к общему числу документов $d \in D$ из признанных экспертом значимыми.

# Значимость аннотации при отборе документов в целевую коллекцию

Пусть $\mathbb{T}\mathrm{s}_i \subset \mathbb{T}\mathrm{s}$ — множество фраз $i$-й аннотации в составе коллекции $\mathbb{T}\mathrm{s}$,

$C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ — 5%-й квантиль эмпирического распределения величины (2) по документу $d_{\max}$ относительно фраз этой аннотации.

Обозначим документ с максимальным среди документов-кандидатов множества $D$ значением величины $C_5$ для фраз в составе $\mathbb{T}\mathrm{s}_i$ как $d_{\max(i)}$.

## Утверждение 2

По степени важности для вычисления $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$ среди аннотаций из $\mathbb{T}\mathrm{s}$ можно выделить следующие пять групп:

- *группа 1:* аннотации, где $d_{\max} = d_{\max(i)}$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$;
- *группа 2:* аннотации, где $d_{\max} \neq d_{\max(i)}$, но $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ относимы к одному кластеру;
- *группа 3:* аннотации, где $d_{\max} \neq d_{\max(i)}$, но $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ лежат в разных кластерах;
- *группа 4:* аннотации, где $d_{\max} \neq d_{\max(i)}$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ относимы к одному кластеру;
- *группа 5:* аннотации, где $d_{\max} \neq d_{\max(i)}$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ лежат в разных кластерах.

При этом наибольшую точность поиска значимых документов дают аннотации из *группы 1* (содержательно — максимально близкие смысловому эталону), *группы 2* и *3*.

# Исходные данные эксперимента: множество документов-кандидатов

- 3 статьи в журнале «Таврический вестник информатики и математики» (ТВИМ);
- 2 статьи в сборниках трудов 8-й и 9-й международных конференций «Интеллектуализация обработки информации» 2010 и 2012 гг.;
- 1 статья в сборнике трудов 15-й Всероссийской конференции «Математические методы распознавания образов» (ММРО, 2011 г.);
- материалы тезисов двух докладов на 13-й Всероссийской конференции «Математические методы распознавания образов» (2007 г.);
- материалы тезисов четырнадцати докладов на 16-й Всероссийской конференции «Математические методы распознавания образов» (2013 г.);
- материалы тезисов двух докладов на международной конференции «Интеллектуализация обработки информации» (ИОИ) 2014 г.;
- материалы одного научного отчёта (Михайлов Д. В., 2003 г.).

**Примечание**

Число слов в документах здесь варьировалось от $218$ до $6298$, число фраз — от $9$ до $587$.

# Исходные данные эксперимента: тематика документов-кандидатов

- математические методы обучения по прецедентам (К. В. Воронцов, М. Ю. Хачай, Е. В. Дюкова, Н. Г. Загоруйко, Ю. Ю. Дюличева, И. Е. Генрихов, А. А. Ивахненко);
- модели и методы распознавания и прогнозирования (В. В. Моттль, О. С. Середин, А. И. Татарчук, П. А. Турков, М. А. Суворов, А. И. Майсурадзе);
- интеллектуальный анализ экспериментальных данных (С. Д. Двоенко, Н. И. Боровых);
- обработка, анализ, классификация и распознавание изображений (А. Л. Жизняков, К. В. Жукова, И. А. Рейер, Д. М. Мурашов, Н. Г. Федотов, В. Ю. Мартьянов, М. В. Харинов).

# Исходные данные формирования коллекций аннотаций

- сборник трудов конференции «Интеллектуализация обработки информации» 2012 г., раздел «Математическая теория и методы классификации» (14 статей);
- сборник трудов 14-й Всероссийской конференции «Математические методы распознавания образов» (2009 г.), раздел «Методы и модели распознавания и прогнозирования» (35 статей);
- сборник трудов 15-й Всероссийской конференции «Математические методы распознавания образов», разделы «Математическая теория и методы классификации» (18 статей) и «Статистическая теория обучения» (10 статей).

## Некоторые технические детали

- Вычисление значений *term frequency* — без учёта предлогов и союзов.
- Извлечение текста из PDF-файла — с помощью функций классов *pdfinterp*, *converter*, *layout* и *pdfpage* в составе пакета *PDFMiner*.
- В целях корректности распознавания все формулы из анализируемых документов переводились экспертом вручную в формат, близкий используемому в LaTeX.
- Для выделения границ предложений в тексте по знакам препинания был задействован метод *sent_tokenize()* класса *tokenize* из входящих в *NLTK*.
- Приведение слов к начальной форме — с помощью *PyMorphy2*.
- При более одном варианте разбора слова для определения его начальной формы берётся ближайший выдаваемому $n$-граммным теггером в составе *nltk4russian*.

Программная реализация на Python 2.7 и результаты экспериментов

# Отбор документов в референтный корпус

Таблица 1. Наиболее значимые для целевой коллекции документы.

| № | Автор(ы), название и выходные данные работы, $d \in D$ | $N_1$ | $N_2$ | $N_3$ |
|---|---|---|---|---|
| 1 | Воронцов К. В. Обзор современных исследований по проблеме качества обучения алгоритмов // ТВИМ. 2004. №1. С. 5–24. | 667 | 6299 | 4 |
| 2 | Воронцов К. В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // 15-я Всерос. конф. «Математические методы распознавания образов» (ММРО-15): Сб. докл. М., 2011. С. 40–43. | 230 | 2345 | 4 |
| 3 | Дюличева Ю. Ю. Стратегии редукции решающих деревьев (обзор) // ТВИМ. 2002. №1. С. 10–16. | 153 | 2360 | 1 |
| 4 | Дюличева Ю. Ю. О программной реализации и апробации алгоритма DFBSA синтеза эмпирического решающего леса // ТВИМ. 2003. №2. С. 35–44. | 174 | 2075 | 2 |
| 5 | Мартьянов В. Ю., Половинкин А. Н., Тув Е. В. Классификация изображений с использованием словаря кодовых слов на основе ансамблей деревьев решений // 9-я Междунар. конф. «Интеллектуализация обработки информации» (ИОИ-9): Сб. докл. М., 2012. С. 480–482. | 139 | 1602 | 1 |

Таблица 2. Документы, не отвечающие *Утверждению 1*, и пофразная близость эталону.

| Автор(ы), название и выходные данные работы | $N_1$ | $N_2$ |
|---|---|---|
| Дюкова Е. В., Песков Н. В. Об алгоритме классификации на основе полного решающего дерева // Всерос. конф. ММРО-13: Сб. докл. М., 2007. С. 125–126. | 23 | 348 |
| Ишкина Ш. Х., Ивахненко А. А. Комбинаторные оценки переобучения пороговых решающих правил // Всерос. конф. ММРО-16: Тез. докл. М., 2013. С. 23. | 14 | 278 |

Здесь: $N_1$ — число фраз документа; $N_2$ — общее число слов с учётом всех вхождений каждого слова; $N_3$ — число коллекций аннотаций, где документ отвечает условию *Утверждению 1*.

Таблица 3. Аннотации в порядке убывания их рейтинга согласно условиям *Утверждению 2*.

| $i$ | Автор (ы) и заголовок статьи | $N_{gr}$ |
|---|---|---|
| 1 | Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска | 1 |
| 2 | Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа | 1 |
| 3 | Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения | 1 |
| 4 | Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов | 1 |
| 5 | Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля | 1 |
| 6 | Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации | 1 |
| 7 | Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов | 2 |
| 8 | Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии | 2 |
| 9 | Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения | 3 |
| 10 | Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа | 5 |

Здесь $N_{gr}$ — номер группы из определяемых *Утверждением 2*; в качестве $d_{\max(i)}$ далее на слайдах указывается номер соответствующего документа по *Таблице 1*. Для сравнения: документ $d_{\max}$ здесь имеет порядковый номер 2, а $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right) = 0{,}53409091$.

Введём в рассмотрение последовательность $X$, состоящую из значений $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ для аннотаций $\mathbb{T}\mathrm{s}_i$ в составе $\mathbb{T}\mathrm{s}$.

Упорядочим $X$ по убыванию с разбиением на кластеры $H_1^X, \ldots, H_{r(X)}^X$.

Таблица 4. Вычисляемые оценки для аннотаций.

| Автор (ы) | $d_{\max(i)}$ | $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ | $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ | принадл. к $H_1^X$ |
|---|---|---|---|---|
| Фрей А. И. | 1 | 0,93571429 | 0,93571429 | true |
| Воронцов К. В., Махина Г. А. | 1 | 0,93461539 | 0,93461539 | true |
| Ботов П. В. | 1 | 0,79114286 | 0,79114286 | true |
| Ивахненко А. А., Воронцов К. В. | 1 | 0,75500000 | 0,75500000 | true |
| Животовский Н. К. | 1 | 0,74787879 | 0,74787879 | true |
| Неделько В. М. | 1 | 0,61312500 | 0,61312500 | true |
| Гуз И. С. | 2 | 0,79000000 | 0,90000000 | true |
| Каневский Д. Ю. | 2 | 0,62222222 | 0,78222222 | true |
| Хачай М. Ю. | 2 | 0,58214286 | 0,93214286 | true |
| Сенько О. В., Кузнецова А. В. | 2 | 0,50000000 | 0,62500000 | false |

**Отметим, что:**

- представленные в *Таблицах 3* и *4* статьи относятся к одному кластеру по величине $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ за исключением работы с порядковым номером 10;
- при их разбиении на кластеры по величине $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ с добавлением в разбиваемую последовательность значения $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$ получим два кластера: к первому будут отнесены статьи с номерами 1 и 2, а ко второму — все остальные.

Документы в составе множества $D$ сортируются по убыванию произведения оценок:

$$val_1 = -1/\log_{10}\left(\Sigma_{H_1}\right), \tag{4}$$

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=\{1,\, r/2,\, r\}\right|\right)}, \tag{5}$$

и, соответственно,

$$val_3 = \left|H_1 \setminus H_{r/2} \setminus H_r\right| / \mathrm{len}\left(Ts\right), \tag{6}$$

где $\Sigma_{H_1}$ есть сумма величин TF-IDF слов, отнесённых к кластеру $H_1$ относительно $d \in D$;
$\sigma\left(\left|H_i,\, i=\{1,\, r/2,\, r\}\right|\right)$ — СКО числа элементов в кластере из списка $\{H_1,\, H_{r/2},\, H_r\}$;
$\mathrm{len}\left(Ts\right)$ — длина фразы $Ts$ в составе группы $\mathbb{T}$s «заголовок + аннотация статьи».

*Первый вариант оценки:*

$$\mathrm{N}_1\left(\mathbb{T}\mathrm{s}, D\right) = \frac{\max\limits_{d\in D}\left(val_1\left(Ts_1, d\right) \cdot val_2\left(Ts_1, d\right) \cdot val_3\left(Ts_1, d\right)\right)}{\sigma\left(\max\limits_{d\in D}\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathbb{T}\mathrm{s}\right)+1}. \tag{7}$$

Здесь:

в числителе — оценка близости эталону заголовка статьи ($Ts_1$);
первое слагаемое в знаменателе — СКО значения близости эталону по всем $Ts_i \in \mathbb{T}$s.

*Второй вариант оценки:*

$$\mathrm{N}_2\left(\mathbb{T}\mathrm{s}, D\right) = \frac{\max\limits_{d\in D}\left(val_1\left(Ts_{\max}, d\right) \cdot val_2\left(Ts_{\max}, d\right) \cdot val_3\left(Ts_{\max}, d\right)\right)}{\sigma\left(\max\limits_{d\in D}\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathbb{T}\mathrm{s}\right)+1}, \tag{8}$$

где $Ts_{\max} \in \mathbb{T}$s — фраза, по которой получен максимум близости эталону.

**Утверждение 3**

*Максимальный итоговый* рейтинг по коллекции получает статья с наибольшим значением оценки (7), попадающим в один кластер со значением оценки (8) для той же статьи.

**Замечания**

- элементы упорядоченной по убыванию числовой последовательности $X$ *принадлежат одному кластеру*, если

$$\begin{cases} \left|\,\mathrm{mc}(X) - \mathrm{first}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \\ \left|\,\mathrm{mc}(X) - \mathrm{last}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \end{cases}, \tag{9}$$

  где $\mathrm{mc}(X)$ — центр масс последовательности как единого кластера, в качестве центра масс здесь берётся среднее арифметическое всех $x_j \in X$;

- корректное применение *Утверждения 3* предполагает отнесение к одному кластеру значений оценки (7) для статьи с максимальным итоговым рейтингом и максимального значения оценки (7) по коллекции, из которой ведётся отбор;
- в случае отсутствия в коллекции статьи, удовлетворяющей данному требованию, *максимальный итоговый рейтинг* получает статья с наибольшим значением оценки (7) по анализируемой коллекции;
- поскольку заголовок и фразы аннотации (по определению) несут некий единый смысловой образ, то допустима мена местами оценок (7) и (8) в *Утверждении 3*.

**Вход:** $S$; // последовательность текстов исходной коллекции,
// отсортированная по убыванию оценки (7)

**Выход:** $S_{res}$; // результат её ранжирования применением *Утверждения 3*

1: $S_{res} := \varnothing$;
2: **пока** $S \neq \varnothing$
3: $Flag := \text{false}$;
4: **для всех** $\mathbb{T}\text{s} \in S$
5: $Tmp := \{\text{N}_1(\text{first}(S), D), \text{N}_1(\mathbb{T}\text{s}, D), \text{N}_2(\text{first}(S), D)\}$;
6: отсортировать $Tmp$ по убыванию;
7: **если** $\text{good}(Tmp) = \text{true}$ **то**
8: $Flag := \text{true}$;
9: $S_{res} := S_{res} \odot \{\mathbb{T}\text{s}\}$; // $\odot$ — операция конкатенации
10: $S := S \setminus \{\mathbb{T}\text{s}\}$;
11: **выход из цикла** *{для}*
12: **конец если**
13: **конец для**
14: **если** $Flag = \text{false}$ **то**
15: $S_{res} := S_{res} \odot \{\text{first}(S)\}$;
16: $S := S \setminus \{\text{first}(S)\}$;
17: **конец если**
18: **конец пока**

Здесь:

good — функция, выдающая true/false в зависимости от выполнения условия (9);

first — функция, возвращающая первый элемент заданной последовательности.

# Для сравнения: результаты ранжирования

Таблица 5. Ранжирование статей согласно алгоритму на *Слайде 16* относительно оценки (7).

| № | Автор (ы) и заголовок статьи | Оценка (7) | Оценка (8) |
|---|---|---|---|
| 1 | Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа | 0,07112036 | 0,07112036 |
| 2 | Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов | 0,05185727 | 0,05185727 |
| 3 | Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения | 0,05169631 | 0,05169631 |
| 4 | Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска | 0,03992817 | 0,03992817 |
| 5 | Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля | 0,02178213 | 0,02178213 |
| 6 | Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии | 0,01969541 | 0,01969541 |
| 7 | Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации | 0,01851287 | 0,01851287 |
| 8 | Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения | 0,01731464 | 0,01731464 |
| 9 | Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов | 0,01591723 | 0,01591723 |
| 10 | Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа | 0,00285329 | 0,03573024 |

# Для сравнения: результаты ранжирования

Таблица 6. Ранжирование статей согласно алгоритму на *Слайде 16* относительно оценки (8).

| № | Автор (ы) и заголовок статьи | Оценка (8) | Оценка (7) |
|---|---|---|---|
| 1 | Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа | 0,07112036 | 0,07112036 |
| 2 | Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов | 0,05185727 | 0,05185727 |
| 3 | Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения | 0,05169631 | 0,05169631 |
| 4 | Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска | 0,03992817 | 0,03992817 |
| 5 | Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа | 0,03573024 | 0,00285329 |
| 6 | Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля | 0,02178213 | 0,02178213 |
| 7 | Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии | 0,01969541 | 0,01969541 |
| 8 | Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации | 0,01851287 | 0,01851287 |
| 9 | Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения | 0,01731464 | 0,01731464 |
| 10 | Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов | 0,01591723 | 0,01591723 |

# Основные выводы и открытые проблемы

1. Основной *результат* настоящей работы — *метод* формирования референтной текстовой коллекции для выявления зависимостей внутри текстов заданной тематики. При этом *зависимости могут быть любыми* и не ограничиваются характерной для эталонной передачи смысла сочетаемостью лексических единиц и их связей.
2. Предложенное решение даёт *минимум пятикратное* сокращение числа документов из минимально релевантных заданной предметной области при отборе в состав референтной коллекции.
3. *Более высокую оценку значимости* для референтной коллекции получают документы, которые *при большем числе фраз* имеют *большее среднее число наиболее значимых терминов* в расчёте на одну фразу *при минимуме её длины*. Содержательно это соответствует более *краткому и ёмкому* изложению — правилу «хорошего тона» изданий по физико-математическим и техническим наукам.
4. В целях повышения точности поиска значимых документов заслуживает внимания *адаптация предложенных оценок* к другим уровням языка, помимо лексики. Сравнение классификаций по разным уровням позволит сделать вывод о значимости документа в спорных случаях, например, невыполнения условия *Утверждения 1* на одном из уровней.