Компьютерные инструменты в образовании, 2015
№ 5: 14–23
УДК: 51-78, 519.2
http://ipo.spb.ru/journal

ОПРЕДЕЛЕНИЕ АВТОРСКОГО СТИЛЯ ТЕКСТОВ НА ОСНОВЕ СТАТИСТИЧЕСКОГО ПОДХОДА ДВУХВЫБОРОЧНОГО ТЕСТИРОВАНИЯ И МЕТОДА K-БЛИЖАЙШИХ СОСЕДЕЙ*

Кижаева Наталья Александровна, Шалымов Дмитрий Сергеевич

Аннотация

В статье рассматривается задача определения авторского стиля текста. Разработан метод, основанный на процессе генерации повторной выборки. Тексты произведений рассматриваются как последовательности символов, сгенерированные различными случайными источниками. Процедура генерации повторных выборок применена для получения тестовых фрагментов текста. Для того чтобы проверить, принадлежат ли выборки одной генеральной совокупности, используется двухвыборочный критерий. Представлены результаты численных экспериментов для текстов на английском и русском языках.

Ключевые слова: *авторский стиль, определение авторства текста, сравнение текстов, двухвыборочный критерий.*

1. ВВЕДЕНИЕ

Растущее количество доступных текстовых источников данных, таких как блоги, посты в социальных сетях и т. п., оказывает большое влияние на развитие методов интеллектуального анализа текстов, к которым традиционно относятся методы определения авторства и стиля. Под авторским стилем понимается характерный выбор слов, грамматических конструкций или любой другой набор признаков, благодаря которым произведение становится уникальным [1].

На сегодняшний день многочисленные сферы применения таких методов включают в себя уголовное и гражданское право (анализ улик в электронном виде, споры об авторских правах [2]), компьютерную безопасность (анализ вредоносного кода), дополнения к исследованиям классических литературных произведений (анализ работ неизвестного или спорного авторства). Таким образом, развитие новых численных методов определения авторства является актуальной задачей. Среди задач определения авторства можно выделить следующие: верификация автора (определить, написан ли текст конкретным автором) [3], выявление плагиата (выявить сходство двух текстов) [4], профилирование автора (получить информацию о возрасте, образовании и т. п. автора текста) [5] и т. д.

Одна из первых значимых работ в этой сфере — исследование Мостеллера и Уоллеса 1964 года, посвященное анализу «Федеральных газет» (The Federalist Papers — сборник из 85 статей, написанных А. Гамильтоном, Дж. Мэдисоном и Дж. Джеем) [6]. Метод, основанный на байесовском статистическом анализе частот распространенных слов, выда-

* Работа выполнена при поддержке гранта СПбГУ 6.37.181.2014.

вал отличительные результаты для предполагаемых авторов. Статья положила начало исследованию новых стилометрических признаков и методов моделирования [1].

С тех пор были предложены различные текстовые метрики. Самые простые берут начало в обычной описательной статистике: средняя длина слов, относительная частота, количество слов в предложении, распределение частей речи и др. Такие величины легко вычисляются. Более того, их легко можно получить для любого языка и корпуса текстов (если доступен подходящий парсер), и, несомненно, они важны для оценки стиля письма. Однако ни одна из этих метрик не отделяет гарантированно одного автора от другого в большинстве случаев [7].

Следующий уровень метрик — это символьные признаки, согласно которым, текст рассматривается просто как последовательность символов [1]. Такими метриками могут выступать количество буквенных и численных символов, количество строчных и заглавных символов, частота букв, количество знаков препинания и т. д. [2]. Более сложный подход заключается в исследовании частот n-грамм (последовательностей n символов). Среди преимуществ этого подхода можно перечислить способность улавливать контекст, использование пунктуации, устойчивость к орфографическим ошибкам.

Методы машинного обучения внесли заметный вклад в область анализа авторства. С точки зрения машинного обучения, задачу определения авторства можно рассматривать как задачу категоризации текстов, когда необходимо отнести текст к определенной категории (автору) на основе тренировочных данных (текстов известного авторства) [8]. Несмотря на то, что множество алгоритмов обучения применимы к такой задаче, их эффективность зависит от выбора признаков. Известно, что использование различных подходов теории оптимального управления также эффективно в создании новых методов интеллектуального анализа данных (data mining) [9].

Текущая работа является продолжением исследований, начатых в [10]. Задача сравнения литературного стиля рассматривается как сравнение распределения двух текстов. Одним из распространенных подходов, который логично применим к поставленной задаче, является использование двухвыборочного критерия, разработанного для определения того, получены ли две выборки из одной и той же генеральной совокупности.

Нормальное распределение не может быть выбрано в качестве предельного в этой задаче, так как текст, написанный одним или несколькими соавторами, едва ли порожден одним случайным источником. С целью стабилизировать этот процесс разработан следующий подход. На первом шаге производится оценка функции распределения нулевой гипотезы в предположении, что стили рассматриваемых текстов совпадают. Далее выборки, сгенерированные из разных текстов, используются для подсчета p-значений в соответствии с построенным распределением в нулевой гипотезе. В случае совпадающих стилей эти p-значений равномерно распределены на отрезке $[0, 1]$. Сравнение полученных значений с равномерным распределением осуществляется при помощи двухвыборочного критерия типа Колмогорова-Смирнова.

2. МЕТОД СРАВНЕНИЯ СТИЛЕЙ ТЕКСТОВ

2.1. Двухвыборочный критерий

Двухвыборочные критерии применяются в задаче проверки гипотезы о принадлежности двух независимых выборок одному закону распределения в евклидовом пространстве $\mathbf{R^d}$.

Пусть $X = X_1, X_2, ..., X_m$ и $Y = Y_1, Y_2, ..., Y_n$ — две независимые случайные величины с неизвестными функциями распределения F и G. Задача состоит в проверке гипотезы:

$$H_0 : F(x) = G(x).$$

Альтернативная гипотеза:

$$H_1 : F(x) \neq G(x).$$

Критерий типа Колмогорова-Смирнова — один из самых распространенных непараметрических критериев проверки однородности двух выборок.

Статистика критерия для эмпирических функций распределений $\tilde{F}(x)$ и $\tilde{G}(x)$ определяется следующим образом:

$$D = \sup_x |\tilde{F}(x) - \tilde{G}(x)|$$

Асимптотический критерий не зависит от распределения, поэтому для достаточно больших выборок тестовая статистика не зависит от их исходных распределений. Он также применим и к задаче проверки соответствия выборки некоторому закону распределения (так называемый критерий согласия Колмогорова). В этом случае статистика измеряет расстояние между эмпирической функцией распределения выборки и предполагаемым кумулятивным распределением.

В работах [12] и [13] представлен обзор непараметрических критериев для многомерного случая, дан их сравнительный анализ. Обобщение критерия Смирнова в многомерном случае дано в статье [14].

Статистика двухвыборочного критерия призвана описывать качество «смешения» элементов, принадлежащих двум независимым одинаково распределенным выборкам S_1 и S_2. Это качество можно измерить с помощью отношений K ближайших соседей, посчитанных для каждого элемента выборок. Если выборки хорошо «замешаны», то такие пропорции приблизительно одинаковые.

Обозначим $|\cdot|$ — произвольная норма в пространстве R^d и положим

$$Z_i = \begin{cases} X_i & 1 \le i \le m, \\ Y_{i-m} & m+1 \le i \le l, \end{cases}$$

где $l = m + n$ –размер выборки. Ближайший r-й сосед для Z_i – это такой элемент Z_j , что $|Z_\nu - Z_i| < |Z_j - Z_i|$ ровно для $r - 1$ значений ν, $1 < \nu < l$, $\nu \neq i, j$.

В статье рассматривается следующая статистика:

$$T_K(S_1 \cup S_2) = \sum_{x \in S_1 \cup S_2} \sum_{r=1}^{K} I\begin{pmatrix} x \text{ и } r\text{-й сосед} \\ \text{принадл. одной выборке} \end{pmatrix}$$

которая представляет все совпадения с K ближайшими соседями.

В условиях гипотезы о равенстве распределений в объединенной выборке в среднем меньше совпадений с ближайшими соседями, чем при альтернативной, поэтому по критерию отвергается нулевая гипотеза для больших значений статистики T_K.

Доказано, что предельное распределение такого рода статистик — нормальное для произвольной нормы в пространстве R^d [15]. При сравнении двух реальных текстов из-за их неоднородности невозможно использовать асимптотически нормальное распределение. В то же время распределение для нулевой гипотезы можно смоделировать в духе бутстрапа (англ. *bootstrap*) [16]. Построение эмпирической функции распределения объединенных выборок подразумевает равенство теоретических распределений согласно

нулевой гипотезе. В то же время в случае разных законов распределения выше описанная процедура (использование только «первоначального смешивания») может выдать зашумленное распределение. Поэтому ниже приводится точное описание процесса генерации выборок.

2.2. Алгоритм

Для начала рассматриваемые тексты преобразуются в бинарные файлы F_1, F_2 и вводится $F_0 = F_1 \cup F_2$. Важной частью алгоритма, нацеленного на различение распределений этих файлов, является процедура создания повторной выборки (англ. *re-sampling*). Выборки формируются с помощью N-грамм, последовательностей N символов из текста, как показано в Алгоритме 1.

Algorithm 1 Процедура сэмплирования

Require:

- F — текстовый файл;
- N — размер атрибута (N-граммы);
- $NWORD$ — количество атрибутов в векторе (размерность вектора);
- $NVEC$ — количество векторов в выборке (размер выборки).

repeat $NVEC$ раз

1. Сгенерировать случайное число – начальную позицию вектора в файле;
2. Построить вектор из $NWORD$ последовательных атрибутов.

Как уже было упомянуто, в рассматриваемой задаче нормальное распределение не может быть выбрано как распределение нулевой гипотезы. Поэтому закон распределения рассчитывается методом бутстрапа, то есть многократной генерацией выборок без замещения из F_0. Далее подсчитываются значения статистики T_K.

Затем вычисляются p-значения относительно закона распределения в нулевой гипотезе, полученного на предыдущем шаге. Если нулевая гипотеза верна и файлы нельзя различить, то это распределение –равномерное на отрезке $[0, 1]$.

Проверка гипотезы осуществляется по двухвыборочному критерию, и каждая оценка рассматривается как испытание по схеме Бернулли. Схемой Бернулли называется последовательность независимых испытаний, в каждом из которых возможны лишь два исхода — «успех» и «неудача», при этом успех в каждом испытании происходит с одной и той же вероятностью $p \in (0, 1)$, а неудача — с вероятностью $q = 1 - p$ [11]. Согласно модели, два текста отличаются по стилю, если соотношение отклонений гипотезы (1, «успех») в последовательности результатов испытаний значительно больше 0.5.

2.3. Комментарии к Алгоритму 2

1. В строке 15 алгоритма p-значения вычисляются по формуле

$$PV(U_i) = \frac{\sum_{perm=1}^{NPER} I(V_{perm} > U_i)}{NPER}, \; i = 1 : NPER.$$

2. Для того чтобы определить, что пропорция отказов в последовательности H значительно больше 0.5, используется одновыборочный z-тест. Соответствующие

p-значения вычисляются по формуле:

$$pp = 1 - \Phi\left(\frac{\widehat{P} - 0.5}{\sqrt{\left(\widehat{P}\left(1-\widehat{P}\right)\right)}}\right), \tag{1}$$

где Ф — кумулятивная функция стандартного нормального распределения, и

$$\widehat{P} = \frac{sum\{H\}}{NPER}.$$

Нулевая гипотеза отклоняется, если $pp < TR$.

Algorithm 2 Основной алгоритм

Require:

- F_1, F_2 – сравниваемые файлы;
- $ITER$ – количество итераций;
- $NPER$ – количество случайных перестановок в процедуре создания повторной выборки;
- K – количество KNN ;
- TR_KS – пороговое значение в одновыборочном критерии К.-С. ;
- TR – пороговое значение, ниже которого нулевая гипотеза о равенстве стилей F_1 и F_2 отвергается;

Ввести $F_0 = F_1 \cup F_2$

for $iter = 1 : ITER$ **do**

for $perm = 1 : NPER$ **do**

$F = random\_permutation(F_0)$;

$S_1 = Sample(N, NWORD, NVEC, F)$;

$S_2 = Sample(N, NWORD, NVEC, F)$;

Вычислить $V_{perm} = T_K(S_1 \cup S_2)$;

end for

Построить эмпирическую функцию распределения P_0 для $\{V_{perm}\}, perm = 1 : NPER$.

for $perm = 1 : NPER$ **do**

$S_1 = Sample(NA, NWORD, NVEC, F_1)$;

$S_2 = Sample(NA, NWORD, NVEC, F_2)$;

Вычислить: $U_{perm} = T_K(S_1 \cup S_2)$;

end for

Вычислить $NPER$ p-значений: $PV = \{pval_{perm}\}, perm = 1 : NPER$ от $\{U_{perm}\}, perm = 1 : NPER$ согласно P_0;

С помощью критерия согласия Колмогорова сравнить PV с равномерным распределением на $[0, 1]$ и получить $h_{iter} = 1$, если нулевая гипотеза отвергается, и $h_{iter} = 0$ в противном случае;

end for

Проверить гипотезу о том, что пропорциональное отношение непринятий гипотезы в последовательности $H = \{h_{iter}\}, iter = 1 : ITER$ меньше, чем TR. В случае если нулевая гипотеза отклоняется, стили файлов F_1 и F_2 считаются разными.

3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Далее приведены несколько экспериментов, демонстрирующие работу алгоритма. Предварительная обработка текстов заключалась в удалении пробелов и символов переноса строки.

3.1. Сравнение текстов на английском языке

Для этого эксперимента были выбраны два романа американских писателей, известных своей дружбой и литературным соперничеством — Ф.С.Фицджеральда и Э.Хемингуэя. Первый файл, обозначенный как F, роман *Великий Гэтсби* Ф.С.Фицджеральда, второй – *Праздник, который всегда с тобой* Э.Хемингуэя (обозначен как H).

В таблицах ниже приведены значения pp (1). Здесь и далее текст, используемый для генерации распределения нулевой гипотезы, стоит в первом столбце (обозначенный как F_1 в Алгоритме 2). Стили двух текстов считаются разными, если нулевая гипотеза не принимается, то есть если $pp < TR$.

Сравнения производились при разных значениях параметров $NWORD$, $NVEC$, посчитано среднее время исполнения (алгоритм реализован в системе MATLAB). Увеличение числа атрибутов и размера выборки приводит к улучшению результата, но увеличивает время исполнения t.

Таблица 1. $NWORD = 8$, $NVEC = 16$, $t = 92.6c$

	F	H
F	0.99	**0.99**
H	**0.76**	0.99

При параметрах $NWORD = 8$, $NVEC = 16$ алгоритму не удается отличить стили разных произведений, хотя он корректно определяет одинаковые работы. Случаи, когда нулевая гипотеза была принята ошибочным образом, выделены в таблицах с результатами жирным шрифтом.

Таблица 2. $NWORD = 16$, $NVEC = 32$, $t = 96.4$ñ

	F	H
F	0.99	**0.07**
H	0.01	0.99

В случае $NWORD = 16$, $NVEC = 32$ нулевая гипотеза ошибочным образом не была отклонена только раз, что означает улучшение результата работы алгоритма.

Таблица 3. $NWORD = 32$, $NVEC = 64$, $t = 107.8$ñ

	F	H
F	0.99	0
H	0	0.99

Таблица 4. $NWORD = 64$, $NVEC = 128$, $t = 130.5$ с

	F	H
F	0.99	0
H	0	0.99

При значениях параметров $NWORD = 32$, $NVEC = 64$ метод точно разделяет авторов и выявляет одинаковые произведения. Дальнейшее увеличение значений не приводит к улучшению результата.

3.2. Сравнение текстов на русском языке

Эксперименты также были проведены на текстах на русском языке. Было выбрано по два произведения трех авторов XIX, XX и XXI века. Поскольку авторы жили в разное время, мы предполагаем, что их стили значительно отличаются друг от друга.

Список исследуемых произведений :

- *Бесы* Ф.М. Достоевского
- *Братья Карамазовы* Ф.М. Достоевского
- *Защита Лужина* В.В. Набокова
- *Приглашение на казнь* В.В. Набокова
- *Generation* П В.О. Пелевина
- *Жизнь насекомых* В.О. Пелевина

Таблица 5. Сравнение книг на русском языке

	Дост. 1	Дост. 2	Набоков 1	Набоков 2	Пелевин 1	Пелевин 2
Дост. 1	1	0.99	0	**0.99**	0	0
Дост. 2	0.99	1	0	**0.76**	0	0
Набоков 1	0	0	1	0.12	0	0
Набоков 2	**0.99**	**0.76**	0.12	1	0	0
Пелевин 1	0	0	0	0	1	0.07
Пелевин 2	0	0	0	0	0.07	1

Все сравнения были проведены при следующих значениях параметров: $ITER = 50$, $N = 32bit$, $NWORD = 16$, $NVEC = 32$, $NPER = 50$, $K = 10$ и $threshold = threshold_{KS} = 0.05$.

Нулевая гипотеза была ошибочно принята при сравнении книги *Приглашение на казнь* В.В. Набокова с романами Ф.М. Достоевского (помечены жирным шрифтом в таблице 5). Это возникло по причине большой разницы в объеме текстов (233595 символов в *Приглашении на казнь*, 872561 в *Бесах* и 1511460 в *Братьях Карамазовых*).

Примеры значений статистики T_K приведены на рис. 1 и рис. 2. По оси x отложены номера итераций $ITER$, по оси y - значения U_{perm} и V_{perm}. Напомним, что V_{perm} вычисляется для выборок из объединенного файла F_0, а U_{perm} — для выборок из файлов F_1 и F_2. В случае разных авторов диаграммы расположены далеко друг от друга, в случае же совпадения стилей они пересекаются.

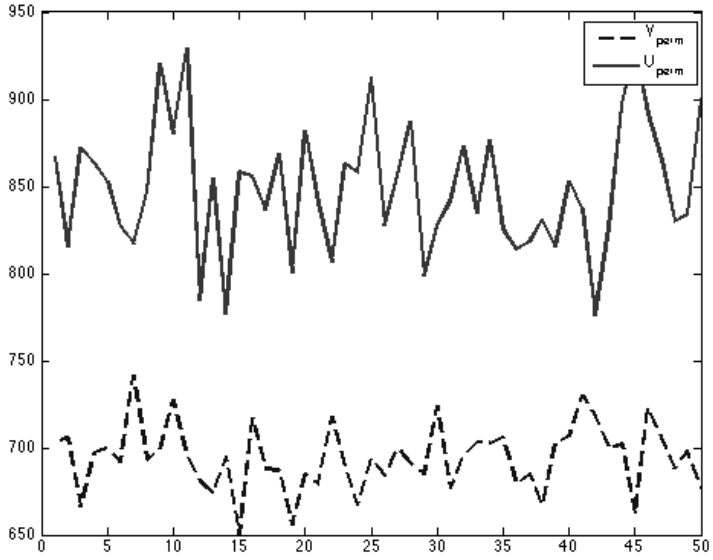

Рис. 1. Сравнение произведений В.В. Набокова и В.О. Пелевина. Значения V_{perm} и U_{perm}

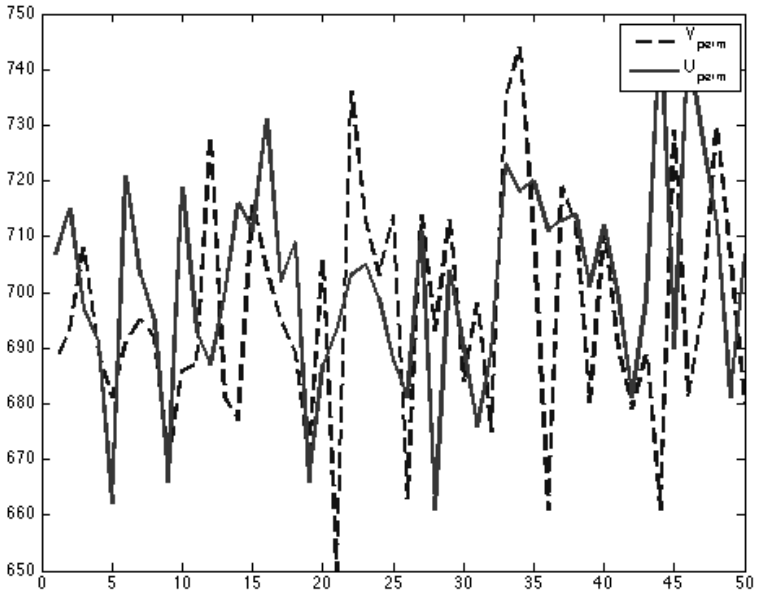

Рис. 2. Сравнение двух романов Ф.М. Достоевского. Значения V_{perm} и U_{perm}

4. ЗАКЛЮЧЕНИЕ

Представленный в работе метод позволяет отличать тексты, обладающие разным авторским стилем. В его основе лежит идея сравнения эмпирических функций распределения, построенных для выборок, сгенерированных из разных источников. Для решения этой задачи был выбран двухвыборочный критерий типа Колмогорова-Смирнова. Моделирование функции распределения нулевой гипотезы осуществлялось методом бутстрапа, то есть многократной генерацией выборок.

Работа развивает идеи, изложенные в [10]. Рассмотрены текстовые коллекции как на английском, так и на русском языках. Проанализировано время работы, а также точность результатов при различных значениях параметров. Метод не требует серьезной предварительной обработки текстов и справляется с анализом коллекций значительных объемов. Несмотря на большое количество вычислительных операций, алгоритм легко масштабируется и может быть распараллелен. Приведенные результаты экспериментов демонстрируют работоспособность алгоритма и его независимость от языка, на котором написаны произведения.

В будущем планируется анализ текстов на восточных языках (например на арабском, иврите и т. д.) и сравнение предложенного метода с распространенными подходами к определению авторства, такими как дельта-преобразование Барроуза [17], модели сжатия [18], ANOVA тест [19], латентное размещение Дирихле [20].

Список литературы

1. E. Stamatatos, "A survey of modern authorship attribution methods *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.
2. R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.
3. M. Koppel, and J. Schler, "Authorship verification as a one-class classification problem *Proc. of the 21st International Conference on Machine Learning*, NewYork: ACM Press, p. 62, 2004.
4. S. Meyer zu Eissen, B. Stein, and M. Kulig, "Plagiarism detection without reference collections *Advances in Data Analysis*, Berlin, Germany: Springer, pp. 359–366, 2007.
5. M. Koppel., S. Argamon, A.R. Shimoni, "Automatically categorizing written texts by author gender *Literary and Linguistic Computing*, vol. 17 no. 4, pp. 401–412, 2002.
6. F. Mosteller, D.L. Wallace, "Inference in an authorship problem - a comparative-study of discrimination methods applied to authorship of disputed Federalist Papers *Journal of the American Statistical Association*, vol. 58(302), p. 275, 1963.
7. P. Juola, "Authorship attribution *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2006.
8. F. Sebastiani, "Machine learning in automated text categorization *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.
9. O. Granichin, Z. V. Volkovich, and D. Toledano-Kitai, *Randomized Algorithms in Automatic Control and Data Mining*, Springer, 2015.
10. Granichin O., Kizhaeva N., Shalymov D., Volkovich Z. "Writing style determination using the KNN text model"// In: Proc. of the 2015 IEEE International Symposium on Intelligent Control, September 21-23, 2015, Sydney, Australia, pp. 900-905.
11. Ширяев А.Н., Вероятность-1: элементарная теория вероятностей, математические основания, предельные теоремы. МЦНМО, Москва, 2011.
12. B.S. Duran, "A survey of nonparametric tests for scale *Communications in statistics - Theory and Methods*, vol. 5, pp. 1287–1312, 1976.
13. W.J. Conover, M.E. Johnson, and M.M. Johnson, "Comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data *Technometrics*, vol.23, pp. 351–361, 1981.
14. J.H. Friedman and L.C. Rafsky, "Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests *Annals of Statistics*, vol.7, pp. 697–717, 1979.
15. N. Henze, "A multivariate two-sample test based on the number of nearest neighbor type coincidences *Annals of Statistics*, vol.16, pp. 772–783, 1988.
16. B. Efron, R. Tibshirani, "An Introduction to the Bootstrap". Boca Raton, FL: Chapman & Hall/CRC, 1993.

17. S. Stein and S. Argamon, "A mathematical explanation of burrows' delta *In Proceedings of Digital Humanities 2006*, Paris, France, 2006.
18. W. Oliveira Jr., E. Justino, L.S. Oliveira, "Comparing compression models for authorship attribution *Forensic Science International*, vol. 228, pp. 100–104, 2013.
19. D.I. Holmes, R. Forsyth, "The Federalist revisited: New directions in authorship attribution *Literary and Linguistic Computing*, vol. 10, no.2, pp. 111–127, 1995.
20. J. Savoy, "Authorship attribution based on a probabilistic topic model *Information Processing and Management*, vol. 49, pp. 341–354, 2013.

LITERARY STYLE DETERMINATION DASED ON STATISTICAL HYPOTHESIS TESTING AND KNN APPROACH

Kizhaeva N. A., Shalymov D. S.

Abstract

The paper presents a method for the literary style determination. It is based on a re-sampling approach and character level features. A text is considered as a sequence of characters (n-grams) generated by different random sources. Bootstap-like approach is used to draw samples from the texts. Kolmogorov-Smirnov two-sample test and KNN based statistic are applied. Experiments with texts in English and Russian are given, illustrating the algorithm operation.

Keywords: *writing style, authorship attribution, two-sample test, re-sampling.*

Кижаева Наталья Александровна,
аспирант кафедры системного программирования СПбГУ,
natalia.kizhaeva@gmail.com

Шалымов Дмитрий Сергеевич,
кандидат физико-математических наук,
shalydim@gmail.com