ISSN 2782-2001
Системы анализа
и обработки данных
том 85, № 1, 2022, с. 73–82

http://journals.nstu.ru/vestnik
Analysis and data
processing systems
Vol. 85, No. 1, 2022, pp. 73–82

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND MANAGEMENT

УДК 81’322+811.222.8+519.25 DOI: 10.17212/2782-2001-2022-1-73-82

О распознавании автора текстового фрагмента на основе частотности буквенных биграмм*

А.А. КОСИМОВ

734042, г. Душанбе, пр. Акад. Раджабовых, 10, Таджикский технический университет имени академика М.С. Осими

abdunabi_kbtut@mail.ru

На примере модельной коллекции таджикских литературных произведений изучается задача о возможности определения авторства фрагмента текста минимального размера, извлеченного из коллекции. Рассматривается модельная коллекция текстов таджикского языка, составленная из произведений классической поэзии и современной прозы на кириллической графике. Каждому произведению сопоставлен цифровой портрет – распределения частотностей символьных биграмм. Для решения проблемы идентификации авторов текстов биграммы вполне приемлемы как количественные характеристики. В качестве инструмента реализации задачи используется γ-классификатор, позволяющий по частотности элементов алфавитно-буквенных биграмм с достаточно высокой степенью эффективности идентифицировать авторов текстовой информации. Математическая модель γ-классификатора представляется в виде триады. Ее первым компонентом является цифровой портрет (ЦП) текста – распределение в тексте частотности буквенных биграмм; вторым компонентом служит формула для вычисления расстояний между ЦП текстов и третьим – алгоритм машинного обучения. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями модельной коллекции, заключалась в определении оптимального значения вещественного параметра γ, для которого минимизируется ошибка нарушения гипотезы «однородности». Также установлено, что с помощью γ-классификатора по цифровому портрету удается идентифицировать авторов произведений на таджикском языке. Путем применения метрического классификатора и методом ближайшего (по расстоянию) соседа удалось идентифицировать авторов убывающих по размерам последовательности текстовых фрагментов от величины в 7000 слов (40 000 символов) вплоть до 20 слов (100 символов). Определен минимальный объем выборки слов или символов для распознавания автора таджикского текста. Описаны результаты экспериментов с минимальным объемом выборки слов (символов) для распознавания автора текста.

Ключевые слова: текст, фрагмент, символ, слова, биграмм, цифровой портрет текста, частотность, ближайший сосед, классификатор, идентификация

* *Статья получена 27 августа 2021 г.*

ВВЕДЕНИЕ

В настоящей статье с использованием γ-классификатора [1, 2] и цифрового текстового портрета [3], характеризующего распределение частотности буквенных биграмм, приводится описание процесса идентификации авторов произведений таджикского текста. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм, биграмм и триграмм с учетом пробела [4]. Существенным моментом в сравнении с нашим предыдущим исследованием [5, 8–15] является изучение вопроса о минимально допустимом размере текстового фрагмента, для которого удается получить удовлетворительный результат решения рассматриваемой задачи. Отметим также, что в сравнении с исследованием [4] выбора фрагментов из текстов, извлеченных из «начала», «середины» и «окончания» произведений, решается задача по отдельности для поэзий и прозы.

В дополнение уместно отметить, что в монографии [16] представлен обширный обзор работ по идентификации авторов текстов на основе разнообразных цифровых портретов текстов и применяемых методов классификации.

Модельная коллекция текстов, на которой разворачивается наше исследование, та же самая, что и в [5].

Обработка коллекционного материала происходила в 4 этапа.

Этап 1. Выбор двух произведений различных авторов – «Рустам ва Сухроб» А. Фирдоуси и «Дохунда» С. Айни. Из каждого произведения извлекалось по 9 случайных выборок текстовых фрагментов, размеры которых в словах и символах показаны в табл. 1.

Таблица 1

Table 1

Информация о размерах фрагментов в словах и символах

Information on the sizes of fragments in words and symbols

Номер фрагмента	1	2	3	4	5	6	7	8	9
Число слов	7000	3500	1700	900	450	250	130	60	20
Число символов	40 000	20 000	10 000	5000	2500	1200	600	300	100

Как видно из таблицы, размеры фрагментов уменьшаются от номера к номеру.

Замечание. Формальное деление числа символов на соответствующее ему число слов не будет означать среднюю длину слова, исчисляемую количеством букв, поскольку к символам помимо букв относятся также знаки препинания и арифметических операций, цифры, обозначения.

Этап 2. Для каждого фрагмента выбранных произведений строится цифровой портрет, который определяется распределением частотности буквенных биграмм, содержащихся в рассматриваемом фрагменте.

Цифровой портрет представляется в табличном виде:

$$N: \quad 1 \quad 2 \quad \ldots \quad 1225$$
$$P: \quad p_1 \quad p_2 \quad \ldots \quad p_{1225},$$

в котором первая строка – номера биграмм, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости буквенных биграмм в тексте T, причем $\sum_{i=1}^{1225} p_i = 1$.

Этап 3. Вычисление расстояний $\rho(T_1, T_2)$ между цифровыми портретами 9 фрагментов T_1 и 12 произведениями T_2 рассматриваемой коллекции текстов. Соответствующие вычисления производились по формуле

$$\rho(T_1, T_2) = \sqrt{\frac{1225}{2}} \max_s \left| \sum_{i=1}^{s} p_i^{(1)} - p_i^{(2)} \right|, \qquad (1)$$

где $p_i^{(1)}$ и $p_k^{(2)}$ – частоты встречаемости в фрагментах T_1 и в произведениях T_2 буквенных биграмм i (i = 1,…, 1225) и (s = 1,…, 1225).

Этап 4. Определение автора текстового фрагмента производится методом ближайшего соседа [6, 7]. Сущность метода заключается в том, что классифицируемый фрагмент T_1 объявляется принадлежащим тому автору, чье произведение T_2 в сравнении с другими произведениями оказывается наиболее «сходным» с фрагментом. Иными словами, рассматриваемые объекты являются ближайшими соседями, и расстояния между их цифровыми портретами минимальны в сравнении с прочими расстояниями.

Полученные результаты сведены в две группы таблиц (табл. 2–4 и 5–7). В ячейках таблиц представлены расстояния от 12 произведений модельной коллекции текстов до девяти фрагментов из романа С. Айни «Дохунда» (в 1-й группе) и до 9 фрагментов из поэмы А. Фирдоуси «Рустам ва Сухроб» (во 2-й группе).

В этих таблицах используются те же сокращения для имен авторов и названий произведений, что и в статье [5], а именно: А. Фирдоуси «Бежан бо Манижа» (АФ, Б&М, 14799) и «Рустам ва Сухроб» (АФ, Р&С, 16355); Дж. Руми «Маснавии Маънави, Дафтари 1» (ЧР, ММ1, 48713) и «Маснавии Маънави, Дафтари 2» (ЧР, ММ2, 41661); А. Суруш «Дафтари 1» (АС, Д1, 7890) и «Дафтари 2» (АС, Д2, 9322); С. Айни «Одина» (СА, О, 25446), «Ахмади Девбанд» (СА, АД, 7480), «Дохунда» (СА, Д, 71134) и «Марги судхур» (СА, МС, 48801); С. Турсун «Нисфирузи» (СТ, Н, 9936) и «Повести Камони Рустам» (СТ, ПКР, 4041). Для авторов и их произведений приняты обозначения, указываемые в скобках: первые две буквы – это инициалы авторов, вторые – сокращенные шифры текстов, третьи – число слов в произведениях.

В последующих таблицах первые 2 колонки указывают авторов и их произведения, а ячейки девяти других колонок – значения расстояний фрагментов различных длин до соответствующих произведений, вычисленные по формуле (1). Отметим также, что в названиях таблиц используются фразы о фрагментах, извлеченных из ***«начала»***, ***«середины»*** и ***«окончания»*** произведений, тем самым в определенном смысле подчеркивается случайный характер выбора фрагментов из текстов произведений.

Закрашенные ячейки этой таблицы показывают, что из четырех фрагментов, взятых из «начала» романа С. Айни «Дохунда», все 4 оказались ближайшими соседями для самой произведения. Кроме того, еще 3 фрагмента (размерами в 2500, 1200 и 600) оказались ближайшими соседями для поэмы Дж. Руми «Маснавии Маънави, Дафтари 1». Интересно, что 2 самых маленьких

фрагмента (в 300 и 100 символов) оказались ближайшими соседями для поэм А. Фирдоуси «Рустам ва Сухроб» и «Бежан бо Манижа».

Таблица 2

Table 2

Расстояния между цифровыми портретами произведений из коллекции текстов и фрагментами, извлеченными из «начала» романа С. Айни «Дохунда»

Distances between digital portraits of works from the collection of texts and fragments extracted from the "beginning" of the Dohunda novel by S. Aini

Авторы (произв.)		Длина фрагментов (в символах)								
		40 000	**20 000**	**10 000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	0.5573	0.5982	0.6717	0.7503	1.0922	1.2611	1.6813	1.5618	2.3129
	Б&М	0.6544	0.6716	0.6682	0.7212	1.1462	1.3140	1.7159	1.6335	2.2528
ЧР	**ММ1**	0.7584	0.8357	0.7952	1.0484	0.9381	1.1166	1.4725	1.7174	2.4685
	ММ2	0.8130	0.8841	0.8559	1.1104	1.0265	1.2275	1.5834	1.7827	2.5338
АС	**Д1**	0.4965	0.5156	0.9310	0.9761	1.2956	1.4943	1.8959	2.1686	2.9197
	Д2	0.3919	0.4851	0.7735	0.7849	1.1836	1.3003	1.7000	1.9239	2.6749
СТ	**Н**	0.4577	0.4234	0.7571	0.7876	1.2746	1.4555	1.8483	2.1166	2.8677
	ПКР	0.5441	0.5874	0.8607	0.8800	1.2978	1.4787	1.8927	2.2428	2.9939
СА	**О**	**0.3130**	**0.3175**	**0.6645**	**0.7096**	**1.0283**	**1.2285**	**1.6214**	**2.0498**	**2.8009**
	АД	**0.5506**	**0.6049**	**1.0013**	**0.9192**	**1.3278**	**1.4543**	**1.8821**	**2.2721**	**3.0232**
	Д	**0.1232**	**0.1482**	**0.6078**	**0.6529**	**0.9750**	**1.1686**	**1.5704**	**1.9388**	**2.6899**
	МС	**0.2562**	**0.2456**	**0.6564**	**0.7016**	**1.0170**	**1.2144**	**1.5983**	**1.9371**	**2.6882**

Таблица 3

Table 3

Расстояния между цифровыми портретами произведений из коллекции текстов и фрагментами, извлеченными из «середины» романа С. Айни «Дохунда»

Distances between digital portraits of works from the collection of texts and fragments extracted from the "middle" of the Dohunda novel by S. Aini

Авторы (произв.)		Длина фрагментов (в символах)								
		40 000	**20 000**	**10 000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	0.5732	0.8167	0.7396	1.0583	0.9825	1.0452	1.1623	1.7480	1.6993
	Б&М	0.7022	0.9456	0.8792	1.1253	1.1185	1.1427	1.2877	1.7797	1.6941
ЧР	**ММ1**	0.7356	1.0244	0.9692	1.1792	1.0078	1.1868	1.3460	2.0187	1.9645
	ММ2	0.7907	1.0878	1.0124	1.2333	1.0619	1.2722	1.3939	2.0629	2.0087
АС	**Д1**	0.5232	0.6087	0.6365	0.7520	0.6976	0.8247	0.9005	1.5000	1.6306
	Д2	0.4695	0.6444	0.6177	0.8370	0.8024	0.9197	0.9955	1.5950	1.5978
СТ	**Н**	0.4576	0.6337	0.5411	0.7787	0.7073	0.8797	0.9555	1.5550	1.6652
	ПКР	0.5479	0.5887	0.5823	0.6048	0.6376	0.6899	0.7863	1.4214	1.8266
СА	**О**	**0.2970**	**0.2827**	**0.3377**	**0.3843**	**0.3892**	**0.5474**	**1.2148**	**1.2685**	**1.4150**
	АД	**0.6741**	**0.7069**	**0.6588**	**0.5567**	**0.7036**	**0.6747**	**0.9854**	**1.4066**	**1.9524**
	Д	**0.1766**	**0.4375**	**0.3604**	**0.5681**	**0.5244**	**0.7299**	**1.0766**	**1.4519**	**1.4266**
	МС	**0.2846**	**0.3704**	**0.3081**	**0.4759**	**0.4246**	**0.6376**	**1.0314**	**1.3129**	**1.4253**

Как видно из табл. 2, для восьми фрагментов, взятых из «середины» романа «Дохунда», один фрагмент оказался ближайшим соседом для самого произведения, один фрагмент – ближайшим соседом для «Марги судхур» и 6 фрагментов – ближайшими соседями для «Одина». Кроме того, всего лишь один фрагмент (размером в 600 символов) оказался ближайшим соседом для произведения С. Турсуна «Повести Камони Рустам».

Таблица 4

Table 4

Расстояния между цифровыми портретами произведений из коллекции текстов и фрагментами, извлеченными из «окончания» романа С. Айни «Дохунда»

Distances between digital portraits of works from the collection of texts and fragments extracted from the "end" of the Dohunda novel by S. Aini

Авторы (произв.)		Длина фрагментов (в символах)								
		40 000	**20 000**	**10 000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	0.5628	0.6208	0.5537	0.6202	0.6871	0.9209	0.8734	1.7209	1.9824
	Б&М	0.6697	0.7422	0.6638	0.7522	0.8067	0.9392	0.8917	1.7392	2.0540
ЧР	**ММ1**	0.7293	0.7597	0.7709	0.8398	0.9750	1.1842	1.1367	1.9842	1.8407
	ММ2	0.7834	0.8120	0.8192	0.8353	1.0301	1.1866	1.1438	1.9587	1.8086
АС	**Д1**	0.5978	0.5670	0.7384	0.5152	0.8140	1.0133	0.9722	1.8133	2.1899
	Д2	0.4811	0.3955	0.4918	0.5168	0.8001	1.0124	0.9649	1.8124	2.0716
СТ	**Н**	0.5052	0.4906	0.6405	0.4039	0.6616	0.9098	0.9120	1.7098	2.2713
	ПКР	0.6484	0.6176	0.7889	0.5160	0.4772	0.6742	1.0508	1.8593	2.3876
СА	**О**	**0.3654**	**0.3270**	**0.4934**	**0.4218**	**0.4954**	**0.7064**	**0.9380**	**1.9927**	**2.5028**
	АД	**0.5793**	**0.4900**	**0.6613**	**0.5182**	**0.5652**	**0.8035**	**1.0971**	**2.2048**	**2.7042**
	Д	**0.2219**	**0.1888**	**0.3601**	**0.3788**	**0.5171**	**0.7438**	**0.7492**	**1.8338**	**2.3522**
	МС	**0.2999**	**0.2691**	**0.4404**	**0.4816**	**0.5185**	**0.7634**	**0.9229**	**1.9314**	**2.4918**

Для фрагментов из «конца» романа «Дохунда» (размерами не менее 5000 и 600 символов) основной результат тот же, что и в двух предыдущих случаях: ближайшими для них соседями служат только произведения С. Айни. Кроме того, всего лишь один фрагмент (размером в 100 символов) оказался ближайшим соседом для поэмы Дж. Руми «Маснавии Маънави, Дафтари 2». Интересно, что 3 других фрагмента (размерами в 2500, 1200 и 300) оказались ближайшими соседями для произведений С. Турсуна «Повести Камони Рустам» и «Нисфирузи».

Таблица 5

Table 5

Расстояния между цифровыми портретами произведений из коллекции текстов и фрагментами, извлеченными из «начала» поэмы А. Фирдоуси «Рустам ва Сухроб»

Distances between digital portraits of works from the collection of texts and fragments extracted from the "beginning" of the Rustam & Suhrob poem by A. Firdousi

Авторы (произв.)		Длина фрагментов (в символах)								
		40 000	**20 000**	**10 000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.1546**	**0.1875**	**0.4082**	**0.4992**	**0.5481**	**0.8051**	**1.1439**	**1.3656**	**4.5971**
	Б&М	**0.2449**	**0.2787**	**0.4306**	**0.4558**	**0.5225**	**0.7216**	**0.9941**	**1.3299**	**4.6489**
ЧР	**ММ1**	0.4687	0.5032	0.6811	0.8072	0.8467	1.1006	1.3661	1.5885	4.2512
	ММ2	0.5336	0.5681	0.7458	0.8851	0.9092	1.1616	1.4271	1.6495	4.2329
АС	**Д1**	0.9457	0.9818	1.1572	1.2691	1.3285	1.5824	1.8478	2.0703	4.8321
	Д2	0.7414	0.7701	0.9494	1.0648	1.1243	1.3782	1.6435	1.8661	4.6062
СТ	**Н**	0.8734	0.9063	1.0663	1.2237	1.2471	1.4468	1.7121	1.9346	4.7704
	ПКР	1.0088	1.0431	1.2001	1.3605	1.3809	1.5375	1.8316	2.0216	4.8087
СА	**О**	0.8587	0.8447	1.0526	1.1802	1.2274	1.2817	1.6181	1.7696	5.0472
	АД	0.9738	0.9976	1.1677	1.3137	1.3856	1.5231	1.8619	1.9381	5.1542
	Д	0.6547	0.6429	0.8486	0.9774	1.0246	1.2165	1.5178	1.7043	4.8671
	МС	0.7363	0.7583	0.9302	1.0455	1.1163	1.2878	1.5532	1.7757	4.9891

В табл. 5–7 девять фрагментов выбираются из поэмы А. Фирдоуси «Рустам ва Сухроб». Закрашенные ячейки этой таблицы показывают, что из восьми фрагментов, взятых из «начала» поэмы А. Фирдоуси «Рустам ва Сухроб», 3 оказались ближайшими соседями для самой поэмы, а 5 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь один фрагмент (размером в 100 символов) оказался ближайшими соседом для поэмы Дж. Руми «Маснавии Маънави, Дафтари 2».

Таблица 6

Table 6

Расстояния между цифровыми портретами произведений из коллекции текстов и фрагментами, извлеченными из «середины» поэмы А. Фирдоуси «Рустам ва Сухроб»

Distances between digital portraits of works from the collection of texts and fragments extracted from the "middle" of the Rustam & Suhrob poem by A. Firdousi

Авторы (произв.)		Длина фрагментов (в символах)								
		40 000	**20 000**	**10 000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.0801**	**0.2232**	**0.3465**	**0.7885**	**0.6762**	**0.6125**	**0.7236**	**2.4113**	**1.5991**
	Б&М	**0.2058**	**0.2326**	**0.3856**	**0.7948**	**0.6072**	**0.5566**	**0.7761**	**2.3231**	**1.7191**
ЧР	**ММ1**	0.5624	0.5867	0.7125	1.1195	1.1087	0.9869	1.1738	2.5881	1.3655
	ММ2	0.6155	0.6345	0.7721	1.1617	1.1349	1.0401	1.2224	2.6382	1.4321
АС	**Д1**	0.8553	0.8839	0.7788	0.6378	0.7845	0.8601	1.0698	3.0289	1.8648
	Д2	0.6511	0.6599	0.5745	0.7457	0.7249	0.7512	0.8533	2.8102	1.6723
СТ	**Н**	0.7526	0.8194	0.6739	0.6135	0.6888	0.7802	0.9861	2.9363	1.7576
	ПКР	0.8871	0.9548	0.8106	0.6946	0.8351	0.9169	1.1228	3.0518	1.8553
СА	**О**	0.8573	0.8623	0.6692	0.4569	0.5918	0.7255	0.9591	2.9402	1.5882
	АД	0.9826	0.9774	0.7842	0.5721	0.7092	0.8406	1.0741	3.1706	1.9954
	Д	0.6533	0.6583	0.4652	0.5612	0.4701	0.5329	0.7551	2.8259	1.5417
	МС	0.7349	0.7399	0.5467	0.4473	0.4866	0.6031	0.8366	2.8478	1.5855

Закрашенные ячейки этой таблицы показывают, что из пяти фрагментов, взятых из «середины» поэмы А. Фирдоуси «Рустам ва Сухроб», 4 оказались ближайшими соседями для самой поэмы, а один фрагмент – ближайшим соседом для «Бежан бо Манижа». Кроме того, всего лишь один фрагмент (размером в 100 символов) оказался ближайшим соседом для поэмы Дж. Руми «Маснавии Маънави, Дафтари 1». Интересно, что три других фрагмента (размерами в 5000, 2500 и 1200) оказались ближайшими соседями произведений С. Айни «Марги судхур» и «Дохунда».

Как видно из табл. 6, для восьми фрагментов, взятых из «конца» поэмы А. Фирдауси «Рустам ва Сухроб», 6 оказались ближайшими соседями для самой поэмы, а 2 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь один фрагмент (размером в 300 символов) оказался ближайшим соседом для произведения С. Айни «Ахмади Девбанд».

Таблица 7

Table 7

Расстояния между цифровыми портретами произведений из коллекции текстов и фрагментами, извлеченными из «окончания» поэмы А. Фирдоуси «Рустам ва Сухроб»

Distances between digital portraits of works from the collection of texts and fragments extracted from the "end" of the Rustam & Suhrob poem by A. Firdousi

Авторы (произв.)		Длина фрагментов (в символах)								
		40 000	**20 000**	**10 000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.1246**	**0.1531**	**0.1984**	**0.4345**	**0.4677**	**0.4627**	**0.6423**	**1.3389**	**4.0598**
	Б&М	**0.2475**	**0.2311**	**0.2388**	**0.4905**	**0.3773**	**0.4905**	**0.6784**	**1.3789**	**3.9715**
ЧР	**ММ1**	0.6055	0.5878	0.5971	0.7289	0.5801	0.9481	1.0529	1.5929	4.2365
	ММ2	0.5417	0.5269	0.5818	0.7849	0.6361	0.9959	0.9873	1.6322	4.2867
АС	**Д1**	0.7065	0.7092	0.8487	0.9051	0.8217	0.6754	1.0173	1.1852	4.6725
	Д2	0.5022	0.4627	0.6021	0.6585	0.7147	0.5423	0.8131	1.2518	4.5101
СТ	**Н**	0.6137	0.6284	0.7771	0.8261	0.7274	0.5776	0.9403	1.2659	4.5848
	ПКР	0.7473	0.7637	0.9105	0.9556	0.8722	0.7261	1.0225	1.1005	4.7781
СА	**О**	0.7797	0.7396	0.7767	1.0118	0.9397	0.8122	1.0347	1.3253	4.8335
	АД	0.9465	0.9064	0.9214	1.1816	1.1531	0.9964	1.0721	1.0502	4.9397
	Д	0.6013	0.5612	0.5762	0.8271	0.7557	0.7469	1.0266	1.3011	4.6307
	МС	0.7167	0.6766	0.6916	0.9486	0.8765	0.7351	0.9432	1.2235	4.6988

ЗАКЛЮЧЕНИЕ

Таким образом, результаты, представленные в таблицах, показывают, что ближайшими соседями по отношению к выбранным фрагментам являются в основном произведения именно того же автора, из произведения которого извлекались сами фрагменты. В иной интерпретации это значит, что методом ближайшего соседа путем вычисления расстояний по формуле (1) представляется возможным установить авторство достаточно малого фрагмента литературного произведения, причем для поэтических произведений (в сравнении с прозаическими) более успешно.

Для художественных текстов можно предложить оценку эффективности применяемого метода, опираясь на вполне естественную гипотезу, согласно которой фрагмент, извлекаемый из какого-либо произведения, должен быть «однородным» с любыми произведениями одного и того же автора. На языке «расстояний» этому соответствует утверждение о том, что ближайшими соседями искомого фрагмента являются, прежде всего, произведения того же автора.

Для прозаического произведения по данным табл. 2 метод ближайшего соседа безошибочно определяет автора фрагментов, состоящих из не менее 5000 символов.

По данным табл. 3 метод безошибочно определяет автора восьми фрагментов из девяти и для одного фрагмента (размером 600 символов) допускает ошибку.

По данным табл. 4 метод ближайшего соседа безошибочно определяет автора пяти фрагментов из девяти и для четырех фрагментов (размерами 2500, 1200, 300 и 100 символов) допускает ошибку.

Для поэтического произведения по данным табл. 5 метод ближайшего соседа безошибочно определяет автора восьми фрагментов из девяти и для одного фрагмента (размером 100 символов) допускает ошибку.

По данным табл. 6 метод безошибочно определяет автора пяти фрагментов из девяти и для четырех фрагментов размерами 5000, 2500, 1200 и 300 символов допускает ошибку.

По данным табл. 7 метод ближайшего соседа безошибочно определяет автора восьми фрагментов из девяти и для одного фрагмента размером 300 символов допускает ошибку.

СПИСОК ЛИТЕРАТУРЫ

1. *Усманов З.Д.* Классификатор дискретных случайных величин // Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60, № 7–8. – С. 291–300.

2. *Усманов З.Д.* Алгоритм настройки кластеризатора дискретных случайных величин // Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60, № 9. – С. 392–397.

3. *Косимов А.А., Рахмонов Ф.А.* О распознавании автора текста на основе частотности буквенных биграмм // Конференсияи илми-амалии омузгорон, мухаккикони чавон, докторантон PhD, магистрантон ва донишчуён бахшида ба эълон гардидани солхои 2019–2021 «Солхои рушди дехот, сайёхи ва хунархои мардуми», солхои 2020–2040 «Бистсолаи омузиш ва рушди фанхои табиатшиноси, дакик ва риёзи дар сохаи илму маориф», Рузи илми точик ва 30-солагии Истиклолияти давлатии Чумхурии Точикистон, ДПДТТХ ба номи М.С. Осими. – Хучанд, 2020. – 11 с.

4. *Косимов А.А.* О минимальном объеме текста, необходимого для распознавания его автора // Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60, № 9. – С. 398–401.

5. О распознавании автора текста на основе частотности буквенных униграмм / А.А. Косимов, Р.Ш. Умарализода, А.А. Хасанов, Ш.С. Саидов // Конференсияи чумхуриявии илми-амалии «Илм – асоси рушди инноватсиони», Донишгохи техникии Точикистон ба номи академик М.С. Осими, 27–28 апрели соли 2021. – Душанбе, 2021. – С. 322–326.

6. *Воронцов К.В.* Математические методы обучения по прецедентам. – URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 11.02.2022).

7. *Дьяконов А.Г.* Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (Практикум на ЭВМ кафедры математических методов прогнозирования): учебное пособие. – М.: ВМК МГУ им. М.В. Ломоносова, 2010. – 278 с.

8. *Каримов А.А.* О цифровом портрете текстовой информации // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2019. – № 1 (45). – С. 7–10.

9. *Каюмов М.М.* О цифровом портрете текстовой информации, основанном на частотности знаков пунктуации // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2019. – № 1 (45). – С. 20–23.

10. *Каюмов М.М.* О распознавании автора текста на основе частотности αβ-кодов словоформ // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2020. – № 2 (50). – С. 29–36.

11. *Ашурова Ш.Н.* Оценка эффективности использования словесных биграмм при идентификации текста // Роль ИКТ в инновационном развитии экономики Республики Таджикистан: материалы международной научно-практической конференции. – Душанбе: Бахманруд, 2017. – С. 292–297.

12. *Ашурова Ш.Н.* Оценка эффективности использования словесных триграмм при идентификации текста // Вестник Технологического университета Таджикистана. – 2017. – № 4 (31). – С. 51–58.

13. *Ашурова Ш.Н., Тошхуджаев Х.А.* О распознавании автора текста на основе частотности словесных биграмм // Политехнический вестник. Серия: интеллект, инновации, инвестиции. – 2020. – 2(50). – С. 57–61.

14. *Бахтеев К.С.* О применимости укороченных цифровых портретов для идентификации автора текста // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2020. – № 2 (50). – С. 25–28.

15. *Бахтеев К.С.* О распознавании авторства по усеченным цифровым портретам текста // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2018. – № 4 (173). – С. 82–92.

16. *Романов А.С., Шелупанов А.А., Мещеряков Р.В.* Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста. – Томск: В-Спектр, 2011. – 188 с.

DOI: 10.17212/2782-2001-2022-1-73-82

On the recognition of the author of a text fragment based on the frequency of alphabetic bigrams[*]

A.A. KOSIMOV

Tajik Technical University named after Acad. M.S. Osimi, 10 Acad. Radjabov Prospekt, Dushanbe, 734042

abdunabi_kbtut@mail.ru

Abstract

On the example of a model collection of Tajik literary works, the problem of the possibility of determining the authorship of a fragment of the text of the minimum size extracted from the collection is studied. A model collection of texts in the Tajik language composed of works of classical poetry and modern prose in Cyrillic graphics is considered. Each piece is associated with a digital portrait - the distribution of the frequencies of symbolic bigrams. To solve the problem of identifying the authors of texts, bigrams are quite acceptable quantitative characteristics. A γ-classifier is used as a tool for implementing the task, which allows the authors of textual information to be identified by the frequency of elements of alphabetic bigrams with a sufficiently high degree of efficiency. The mathematical model of the γ-classifier is represented as a triad. Its first component is a digital portrait (DP) of the text - the distribution of the frequency of bigrams in the text; the second component is formulas for calculating the distances between the DP texts and the third is a machine learning algorithm. The tuning of the algorithm using a table of paired distances between all products of the model collection consisted in determining an optimal value of the real parameter γ, for which the error of violation of the "homogeneity" hypothesis is minimized. It was also found that with the help of a γ-classifier by a digital portrait, it is possible to identify the authors of works in the Tajik language. By using the metric classifier and the method of the nearest (in terms of distance) neighbor, it was possible to identify the authors of decreasing sequences of text fragments from 7000 words (40,000 characters) up to 20 words (100 characters). The minimum volume of a sample of words or symbols for recognition of the author of a Tajik text has been determined. The results of experiments with a minimum sample size of words (characters) for recognizing the author of a text are described.

Keywords: text, fragment, symbol, words, bigram, digital portrait of text, frequency, nearest neighbor, classifier, identification

REFERENCES

1. Usmanov Z.D. Klassifikator diskretnykh sluchainykh velichin [The classifier of discrete random variables]. *Doklady Akademii nauk Respubliki Tadzhikistan = Reports of the Academy of Sciences of the Republic of Tajikistan*, 2017, vol. 60, no. 7–8, pp. 291–300.

2. Usmanov Z.D. Algoritm nastroiki klasterizatora diskretnykh sluchainykh velichin [Tuning the algorithm of the clasifier of discrete random variables]. *Doklady Akademii nauk Respubliki Tadzhikistan = Reports of the Academy of Sciences of the Republic of Tajikistan*, 2017, vol. 60, no. 9, pp. 392–397.

3. Kosimov A.A., Rakhmonov F.A. [On the recognition of the author of the text based on the frequency of alphabetic bigrams]. *Scientific-practical conference of teachers, young researchers, doctoral students PhD, undergraduates and students dedicated to the proclamation of 2019–2021 "Years of rural development, tourism and folk crafts", 2020–2040 "Twentieth anniversary of teaching and development of natural sciences, exact and mathematical sciences in the field of science and education", Tajik Science Day and 30th anniversary of the State Independence of the Republic of Tajikistan, Tajik Technical University named after M.S. Osimi*. Khujand, 2020. 11 p. (In Russian).

4. Kosimov A.A. O minimal'nom ob"eme teksta, neobkhodimogo dlya raspoznavaniya ego avtora [On the minimum amount of text required to recognize its author]. *Doklady Akademii nauk Respubliki Tadzhikistan = Reports of the Academy of Sciences of the Republic of Tajikistan*, 2017, vol. 60, no. 9, pp. 398–401. (In Russian).

[*] *Received 27 August 2021.*

5. Kosimov A.A., Umaralizoda R.Sh., Khasanov A.A., Saidov Sh.S. [*On recognition of the author of a text based on the frequency of alphabetic unigrams]. Republican scientific-practical conference "Science – the basis of innovative development"*. Tajik Technical University named after M.S. Osimi. Dushanbe, April 27–28, 2021, pp. 322–326. (In Russian).

6. Vorontsov K.V. *Matematicheskie metody obucheniya po pretsedentam* [Mathematical methods of teaching by precedents]. Available at: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (accessed 11.02.2022).

7. D'yakonov A.G. *Analiz dannykh, obuchenie po pretsedentam, logicheskie igry, sistemy WEKA, RapidMiner i MatLab (Praktikum na EVM kafedry matematicheskikh metodov prognozirovaniya)* [Data analysis, training on precedents, logic games, WEKA, RapidMiner and MatLab systems (Workshop on the computer of the Department of Mathematical Forecasting Methods)]. Moscow, MSU Publ., 2010. 278 p.

8. Kayumov M.M. O tsifrovom portrete tekstovoi informatsii [On the digital portrait of text information]. *Politekhnicheskii vestnik. Seriya: Intellekt. Innovatsii. Investitsii = Polytechnic Bulletin. Series: Intelligence. Innovation. Investments*, 2019, no. 1 (45), pp. 7–10.

9. Kayumov M.M. O tsifrovom portrete tekstovoi informatsii, osnovannom na chastotnosti znakov punktuatsii [On the digital portrait of text information based on the frequency of punctuation marks]. *Politekhnicheskii vestnik. Seriya: Intellekt. Innovatsii. Investitsii = Polytechnic Bulletin. Series: Intelligence. Innovation. Investments*, 2019, no. 1 (45), pp. 20–23.

10. Kayumov M.M. O raspoznavanii avtora teksta na osnove chastotnosti αβ-kodov slovoform [On recognition of the author of a text based on the frequency of *αβ*-codes of word forms]. *Politekhnicheskii vestnik. Seriya: Intellekt. Innovatsii. Investitsii = Polytechnic Bulletin. Series: Intelligence. Innovation. Investments*, 2020, no. 2 (50), pp. 29–36.

11. Ashurova Sh.N. [Assessment of the effectiveness of the use of verbal bigrams in the identification of text]. *Rol' IKT v innovatsionnom razvitii ekonomiki Respubliki Tadzhikistan*: materialy mezhdunarodnoi nauchno-prakticheskoi konferentsii [Materials of the international scientific-practical conference HER "The role of ICT in the innovative development of the economy of the Republic of Tajikistan"]. Dushanbe, Bahmanrud Publ., 2017, pp. 292–297. (In Russian).

12. Ashurova Sh.N. Otsenka effektivnosti ispol'zovaniya slovesnykh trigramm pri identifikatsii teksta [Efficiency evaluation of using word trigrams for a text identification]. *Vestnik Tekhnologicheskogo universiteta Tadzhikistana = Bulletin of the Technological University of Tajikistan*, 2017, no. 4 (31), pp. 51–58. (In Russian).

13. Ashurova Sh.N., Toshkhudzhaev Kh.A. On recognition of the author of the text based on the frequency of verbal bigrams // Polytechnic Bulletin, Series: intelligence, innovation, investment. 2020. 2 (50). pp. 57–61 (in Russian).

14. Bakhteev K.S. O primenimosti ukorochennykh tsifrovykh portretov dlya identifikatsii avtora teksta [About the applicability of shortened digital portraits to identify the author's text]. *Politekhnicheskii vestnik. Seriya: Intellekt. Innovatsii. Investitsii = Polytechnic Bulletin. Series: Intelligence. Innovation. Investments*, 2020, no. 2 (50), pp. 25–28.

15. Bakhteev K.S. O raspoznavanii avtorstva po usechennym tsifrovym portretam teksta [On the recognition of authorship by truncated digital portraits of text]. *Izvestiya Akademii nauk Respubliki Tadzhikistan. Otdelenie fiziko-matematicheskikh, khimicheskikh, geologicheskikh i tekhnicheskikh nauk = News of the Academy of Sciences of the Republic of Tajikistan. Department of physical, mathematical, chemical, geological and technical sciences*, 2018, no. 4 (173), pp. 82–92.

16. Romanov A.S., Shelupanov A.A., Meshcheryakov R.V. *Razrabotka i issledovanie matematicheskikh modelei, metodik i programmnykh sredstv informatsionnykh protsessov pri identifikatsii avtora teksta* [Development and research of mathematical models, methods and software for information processes in the identification of the author of the text]. Tomsk, V-Spektr Publ., 2011. 188 p.

Для цитирования:

Косимов А.А. О распознавании автора текстового фрагмента на основе частотности буквенных биграмм // Системы анализа и обработки данных. – 2022. – № 1 (85). – С. 73–82. – DOI: 10.17212/2782-2001-2022-1-73-82.

For citation:

Kosimov A.A. O raspoznavanii avtora tekstovogo fragmenta na osnove chastotnosti bukvennykh bigramm [On the recognition of the author of a text fragment based on the frequency of alphabetic bigrams]. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*, 2022, no. 1 (85), pp. 73–82. DOI: 10.17212/2782-2001-2022-1-73-82.