ISSN 2782-2001
Системы анализа
и обработки данных
том 82, № 2, 2021, с. 83–94

http://journals.nstu.ru/vestnik
Analysis and data
processing systems
Vol. 82, No. 2, 2021, pp. 83–94

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND CONTROL

УДК 81+81.322+519.25 DOI: 10.17212/2782-2001-2021-2-83-94

Тестирование γ-классификатора, настроенного на распознавание языков произведений на основе латинского алфавита*

З.Д. УСМАНОВ[1,a], А.А. КОСИМОВ[2,b]

[1] *734063, Республика Таджикистан, г. Душанбе, пр. Айни, 299/1, Институт математики им. А. Джураева НАН РТ*
[2] *734042, г. Душанбе, пр. Акад. Раджабовых, 10, Таджикский технический университет имени академика М.С. Осими*

[a] *zafar-usmanov@rambler.ru* [b] *abdunabi_kbtut@mail.ru*

В статье на примере модельной коллекции из десяти текстов на пяти языках (английском, немецком, испанском, итальянском и французском) с использованием латинской графики устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 26 латинских алфавитных букв. Математическая модель γ-классификатора представляется в виде триады. Ее первым компонентом является цифровой портрет (ЦП) текста – распределение в тексте частотности буквенных униграмм; вторым компонентом служит формула для вычисления расстояний между ЦП текстов и третьим – алгоритм машинного обучения, реализующий гипотезу «однородности» произведений, написанных на одном языке, и «неоднородности» произведений, написанных на разных языках. 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». Обученный на текстах модельной коллекции γ-классификатор показал высокую, 100 %-ю точность в распознавании языков произведений. Для тестирования классификатора были выбраны дополнительно шесть случайных текстов, из которых пять на тех же языках, что и тексты модельной коллекции. Методом ближайшего (по расстоянию) соседа все новые тексты подтвердили свою однородность с соответствующими парами одноязычных произведений. Шестой текст на румынском языке показал свою неоднородность по отношению ко всем элементам коллекции. Вместе с тем проявил близость по минимальным расстояниям, прежде всего, к двум текстам на испанском языке и затем и к двум произведениям на итальянском языке.

Ключевые слова: текст, язык, латинская графика, алфавит, частотность униграмм, цифровой портрет текста, гипотеза однородности, классификатор, обучение, распознавание языков, тестирование классификатора, оценка эффективности

* *Статья получена 02 декабря 2020 г.*

ВВЕДЕНИЕ

В наше время письменность на основе латинского алфавита получила широкое распространение среди романской, германской, славянской, финно-угорской, тюркской, семитской и иранской групп языков, среди стран Индокитая, Зондского архипелага и Филиппин, Африки (южнее Сахары), Америки, Австралии и Океании [1]. За исключением современного английского, для большинства других языков латинский алфавит из 26 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z) оказался недостаточным, в связи с чем для отражения фонетических особенностей тех или иных языковых систем к базовой латинской графике были добавлены различные диакритические знаки, лигатуры и другие модификации букв.

Задача, решением которой будем заниматься в настоящей работе, состоит в том, возможно ли обойтись только лишь 26 латинским буквами для автоматического распознавания языка, на котором написана та или иная печатная продукция.

В качестве экспериментального материала, на котором развертывается наше исследование, выбрана небольшая коллекция ***C*** из десяти произведений (текстов), среди которых

на английском языке **(En)**:

У. Шекспир. Romeo and Juliet (Ромео и Джульетта, **en_1**, 25832 *слов*),

М. Твен. A Connecticut Yankee in King Arthur's Court (Янки из Коннектикута при дворе короля Артура, **en_2**, 117257 *слов*);

на немецком языке **(De)**:

Г. Пиз. Schiff ohne Mannschaft (Корабль без экипажа, **de_1**, 59695 *слов*),

Г. Диана. Das flammende Kreuz: Roman (Пылающий Крест: Роман, **de_2**, 70104 *слов*);

на испанском языке **(Es)**:

Д.Дж. Генрих. El ocaso de la magia (Сумерки магии, **es_1**, 73300 *слов*);

В.Ф. Альберто. Oceano (Океан, **es_2**, 103596 *слов*);

на итальянском языке **(It)**:

Г. Эд. Elminster: la nascita di un mago (Эльминстер: рождение волшебника, **it_1**, 127087 *слов*);

С. Роберт. Il paradosso del passato (Парадокс прошлого, **it_2**, 69697 *слов*);

на французском языке **(Fr)**:

С. Жорж. Lavinia (Лавиния, **fr_1**, 13151 *слов*);

Б. Мишель. Les Nymphéas noirs (Черные водяные лилии, **fr_2**, 108137 *слов*).

Отметим, что сведения о текстах содержат имена авторов, названия их произведений в оригинале и в переводе на русский язык, а также сокращенные обозначения произведений совместно с их размерами, определяемыми количеством слов. Особенность коллекции в том, что она охватывает всего лишь 5 европейских языков, и все ее тексты на основе латинской графики с использованием дополнительных специфических символов: четырех символов ä, ö, ß, ü – в немецком (**de**), одного символа ñ – в испанском (**es**), десяти символов à, è, é, ì, í, î, ò, ó, ù, ú – в итальянском (**it**) и четырнадцати символов â, à, ç, é, ê, è, ë, î, ï, ô, û, ù, ü, ÿ – во французском (**fr**) языках.

Приступая к решению поставленной задачи, отметим, что в качестве исследовательского инструмента мы будем использовать *математическую*

триаду в составе цифрового портрета текстов, представляемых распределениями частотности 26 латинских букв, формулы для вычисления расстояний между текстом и алгоритмом для выявления однородных текстов [2]. Упомянутая триада с момента своего появления в 2017 году применялась, прежде всего, при распознавании авторства для различных вариантов ЦП текстов [3–13]. В дополнение к сказанному уместно отметить, что в моногра-фии [14] представлен обширный обзор работ по идентификации авторов текста на основе разнообразных цифровых портретов текстов и применяемых методов классификации.

1. ЦИФРОВОЙ ПОРТРЕТ ПРОИЗВЕДЕНИЙ

В качестве учетных элементов для описания произведений взяты указанные для всех пяти языков 26 латинских букв.

Определение 1. *Цифровым портретом (ЦП) текста будем называть распределение в нем частотности 26 букв.*

ЦП текста T записывается в табличном виде:

$$\begin{aligned} N&: \quad 1 \quad 2 \quad \ldots \quad 26 \\ P&: \quad p_1 \quad p_2 \quad \ldots \quad p_{26}, \end{aligned} \tag{1}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частотности букв в тексте *T*, причем $\sum_{k=1}^{26} p_k = 1$.

Одновременно с (1) цифровой портрет представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{S} p_k \qquad (s = 1,\ldots,26). \tag{2}$$

2. РАССТОЯНИЯ МЕЖДУ ЦИФРОВЫМИ ПОРТРЕТАМИ ТЕКСТОВ

Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{S} p_k{}^{(\alpha)} \tag{3}$$

– соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $(s = 1,\ldots, 26)$.

Определение 2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho(T_1, T_2)$, *определяемое формулой*

$$\rho(T_1, T_2) = \sqrt{\frac{26}{2}} \max_S \left| F^{(1)}(s) - F^{(2)}(s) \right|. \tag{4}$$

3. ГИПОТЕЗА *H* «ОДНОРОДНОСТИ» ПРОИЗВЕДЕНИЙ

Она привлекается для того, чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Ее мы формулируем в следующем виде.

ГИПОТЕЗА *H*. *Произведения, написанные на одном языке, «однородны», а на разных языках – «не однородны».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т. п.

4. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ *H*-ГИПОТЕЗЫ

Пусть γ – некоторое положительное число.

Определение 3. *Тексты называются γ-однородными (написанными на одном языке), если*

$$\rho(T_1, T_2) \le \gamma, \tag{5}$$

и γ-неоднородными (написанными на разных языках), если

$$\rho(T_1, T_2) > \gamma. \tag{6}$$

Неравенства (5) и (6) являются математической интерпретацией (моделью) гипотезы *H*. Это значит, что в дальнейшем мы приступаем к распознаванию языков произведений с помощью математического аппарата, названного γ-классификатором [2].

Определение 4. *γ-классификатор – это зависящий от одного вещественного параметра γ алгоритм принятия решения об отнесении пары текстов T_1 и T_2 к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одному языку в рамках математической модели означает справедливость неравенства (5), а двум разным языкам – справедливость неравенства (6). Гипотеза *H* может нарушаться для каких-то пар текстов одного и того же языка в случае, когда вместо неравенства (5) имеет место неравенство (6), а также в случае, когда какие-то два текста на разных языках удовлетворяют неравенству (5) вместо того, чтобы выполнялось неравенство (6).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы *H* одновременно в двух случаях: при невыполнении неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнении неравенства «неоднородности» в случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma) / L,$$

где L – число взаимных расстояний между всеми парами текстов из коллекции $\boldsymbol{C}$ (в нашем случае $L = C_{10}^2 = 45$. Из этой формулы следует, что π может

принимать значения из отрезка [0, 1], причем $\pi = 0$, если $\tau = L(= 45)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу H следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

5. ИТОГОВЫЕ РЕЗУЛЬТАТЫ НА ПРИМЕРЕ КОЛЛЕКЦИИ *C*

В процессе настройки выполняются следующие операции:

– предобработка экспериментальных данных путем удаления из всех произведений коллекции дополнительных сверх латинского алфавита букв;

– вычисления цифровых портретов (1) (частотности 26 латинских букв) для всех десяти произведений модельной коллекции ***C***;

– вычисления по формулам (2), (3) и (4) сорока пяти парных расстояний $\rho(T_1, T_2)$ между произведениями коллекции ***C*** (результаты расчетов приведены в табл. 1);

Таблица 1

Table 1

Расстояния между текстами коллекции *C*

Distances between texts of the collection *C*

Тексты		En		De		Es		It		Fr	
		en_1	en_2	de_1	de_2	es_1	es_2	it_1	it_2	fr_1	fr_2
En	en_1										
	en_2	0.0832									
De	de_1	0.3949	0.3281								
	de_2	0.3817	0.3148	0.0287							
Es	es_1	0.3606	0.3030	0.2845	0.2963						
	es_2	0.3471	0.2895	0.3077	0.2945	0.0450					
It	it_1	0.2486	0.2302	0.2677	0.2579	0.1950	0.1814				
	it_2	0.2426	0.2243	0.2988	0.2928	0.2086	0.1951	0.0378			
Fr	fr_1	0.1354	0.1945	0.3982	0.3849	0.3205	0.2941	0.2628	0.2691		
	fr_2	0.1480	0.1833	0.4038	0.3920	0.3260	0.2995	0.2776	0.2773	0.0299	

– вычисление с помощью алгоритма настройки γ-классификатора оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы *H* достигает минимального значения, и, следовательно, величина π показателя выполнения гипотезы *H* принимает максимальное значение.

По данным табл. 1 получены следующие результаты:

– совокупность всех пар расстояний размещается на отрезке [0.0287, 0.4038], при этом минимальное расстояние реализуется между двумя произведениями **de_1** и **de_2** на немецком языке, а максимальное – между **de_1** на немецком и **fr_2** на французском языках;

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{\text{опт}} \in [0.0833;\ 0.1353); \tag{7}$$

в соответствии с определением 3 это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значения $\gamma^{\text{опт}}$ из указанного полуинтервала, то пара текстов принадлежит одному и тому же языку (соответствующие расстояния в таблице помечены серым цветом); если же превосходит, то принадлежит разным языкам (соответствующие расстояния оставлены непомеченными);

– отметим, что для всех (без исключений) произведений коллекции ***C*** полностью подтвердилась гипотеза *H* и ее математическая интерпретация в виде определения 3, и потому получено

$$\tau = \tau_{\min} = 0,$$

т. е. ни одно из неравенств (5) и (6) не было нарушено;

– вследствие этого показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{\max} = 1.$$

6. ТЕСТИРОВАНИЕ

Итак, настройка (обучение) γ-классификатора на данных модельной коллекции текстов ***C*** прошла успешно. Для тестирования классификатора выбраны случайным образом 6 текстов:

на английском языке **(En)**:

Дж. Лондон. The Call of the Wild (Зов предков) (Text_En, 31763 *слов*);

на немецком языке **(De)**:

М. Вилли. Die seltsamen Reisen des Marco Polo (Странные путешествия Марко Поло) (Text_De, 126607 *слов*);

на испанском языке **(Es)**:

Д. Арне. Misterioso (Таинственный) (Text_Es, 106835 *слов*);

на итальянском языке **(It)**:

Ш. Боб. Sfida al cielo (Вызов небу) (Text_It, 101154 *слов*);

на французском языке **(Fr)**:

К.С. Доминикович. Fantôme (Призрак) (Text_Fr, 46089 *слов*);

на румынском языке **(Ro)**:

Т.Р. Руэл. Întoarcerea regelui (Возвращение короля) (Text_Ro, 146266 *слов*).

Отметим, что сведения относительно выбранных произведений описаны по той же схеме, что и для элементов коллекции ***C***. В дополнение к предыдущему отметим, что в румынском языке (**ro**) латинский алфавит расширен на 5 символов, ă, â, î, ș. ț.

Для шести произведений, предназначенных для тестирования, построены цифровые портреты (1) и затем для каждого из них по формулам (2)–(4) вычислены расстояния до десяти объектов коллекции ***C***. Соответствующие значения записаны в ячейках табл. 2, расположенных на пересечениях строк и столбцов. Там же серым цветом выделены пары ячеек с минимальными расстояниями между тестируемыми и содержащимися в коллекции произведениями.

Таблица 2

Table 2

Расстояния между текстами коллекции *C* и шестью случайно выбранными тестируемыми произведениями

Distances between the texts of the collection *C* and the 6 random tested works

Тексты		Text_En	Text_De	Text_Es	Text_It	Text_Fr	Text_Ro
En	en_1	0.1592	0.4069	0.3235	0.2378	0.1477	0.2084
	en_2	0.0857	0.3400	0.2659	0.2194	0.1905	0.1760
De	de_1	0.2599	0.0305	0.2659	0.2866	0.4235	0.2723
	de_2	0.2467	0.0489	0.2526	0.2734	0.4103	0.2663
Es	es_1	0.2674	0.3010	0.0552	0.1874	0.3250	0.1707
	es_2	0.2538	0.3197	0.0430	0.1738	0.2985	0.1440
It	it_1	0.1987	0.2882	0.1579	0.0330	0.3050	0.1565
	it_2	0.2365	0.3260	0.1715	0.0281	0.3047	0.1563
Fr	fr_1	0.2802	0.4101	0.2712	0.2501	0.0460	0.1933
	fr_2	0.2690	0.4158	0.2767	0.2604	0.0448	0.2033

Полученные результаты показывают, что ближайшими соседями [15, 16] первых пяти произведений оказались как раз однородные с ними по языку пары текстов исходной коллекции. Что касается текста на румынском языке

(Text_Ro), то все его расстояния до десяти коллекционных текстов превзошли максимальное значения $\gamma^{\text{опт}}$ (см. формулу (7)). Следовательно, как и ожидалось, для Text_Ro в коллекции не оказалось ни одного однородного объекта. Интересно, однако, отметить, что γ-классификатор указал в качестве ее ближайших соседей два произведения es_1 и es_2 на испанском и два произведения it_1 и it_2 на итальянском языках.

ЗАКЛЮЧЕНИЕ

Итак, γ-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ на случайных выборках текстов с цифровыми портретами на основе частотности 26 базовых латинских букв подтвердил 100 %-ю статистическую способность к распознаванию языков произведений.

Таким образом, математическая триада в составе цифрового портрета текстов, представляемых распределениями частотности 26 латинских букв, формул (1)–(3) для вычисления расстояний между текстами и алгоритмами для выявления однородных текстов оказалась подходящей для эффективного решения поставленной задачи.

Авторы выражают уверенность в том, что увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания языков, но также и для самых разнообразных однородностей текстовых документов.

СПИСОК ЛИТЕРАТУРЫ

1. Список латинских букв // Википедия. – URL: https://ru.wikipedia.org/wiki/Список_латинских_букв (дата обращения: 13.11.2020).

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

3. *Усманов З.Д., Косимов А.А.* О распознавании авторства таджикского текста // Доклады Академии наук Республики Таджикистан. – 2016. – Т. 59, № 3–4. – С. 114–119.

4. *Косимов А.А.* Оценка эффективности использования биграмм при идентификации текста // Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60, № 5–6. – С. 224–229.

5. *Косимов А.А.* Оценка эффективности использования триграмм при идентификации текста // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2017. – № 1 (166). – С. 51–57.

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

7. *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.

8. *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.

9. *Ашурова Ш.Н.* Оценка эффективности использования словесных биграмм при идентификации текста // Материалы международной научно-практической конференции ТУТ «Роль ИКТ в инновационном развитии экономики Республики Таджикистан». – Душанбе: Баҳманрӯд, 2017. – С. 292–297.

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

11. *Ашурова Ш.Н., Тошходжаев Х.А.* О распознавании автора текста на основе частотности словесных биграмм // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2020. – № 2 (50). – С. 57–61.

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

13. *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.

14. *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.

15. *Воронцов К.В.* Математические методы обучения по прецедентам. – URL: http://www.ccas.ru/voron (дата обращения: 01.11.2020).

16. *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.

Усманов Зафар Джураевич, доктор физико-математических наук, профессор, академик НАН РТ, Институт математики им. А. Джураева НАН РТ. E-mail: zafar-usmanov@rambler.ru

Косимов Абдунаби Абдурауфович, кандидат технических наук, старший преподаватель кафедры АСУ факультета ИКТ, Таджикский технический университет имени академика М.С. Осими. E-mail: abdunabi_kbtut@mail.ru

Usmanov Zafar Dzh., Doctor of Physical and Mathematical Sciences, professor, Academician of the National Academy of Sciences of the Republic of Tajikistan, Institute of Mathematics. A. Dzhuraeva National Academy of Sciences of the Republic of Tajikistan. E-mail: zafar-usmanov@rambler.ru

Kosimov Abdunabi A., PhD (Eng.), senior lecturer at the department of ACS, Tajik Technical University named after academician M.S. Osimi. E-mail: abdunabi_kbtut@mail.ru

DOI: 10.17212/2782-2001-2021-2-83-94

Testing the γ-classifier adapted to recognize the languages of works based on the Latin alphabet[*]

Z.D. USMANOV[1,a], *A.A. KOSIMOV*[2,b]

[1] *734063, Republic of Tajikistan, Dushanbe, 299/1 Aini Prospekt, Academician of NAS RT; A. Juraev Institute of Mathematics, the National Academy of Sciences of Tajikistan*
[2] *734042, Dushanbe, 10 Acad. Radjabovykh Prospekt, Tajik Technical University named after acad. M.S. Osimi*

[a] *zafar-usmanov@rambler.ru* [b] *abdunabi_kbtut@mail.ru*

Abstract

Using the example of a model collection of 10 texts in five languages (English, German, Spanish, Italian, and French) using Latin graphics, the article establishes the applicability of the γ-classifier for automatic recognition of the language of a work based on the frequency of 26 common Latin alphabetic letters. 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 alphabetic unigrams in the text; the second component is formulas for calculating the distances between the DP texts and the third is a machine learning algorithm that implements the hypothesis of "homogeneity" of works written in one language and "heterogeneity" of works written in different languages. 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. The γ-classifier trained on the texts of the model collection showed a high, 100% accuracy in recognizing the languages of the works. For testing the classifier, an additional six random texts were selected, of which five were in the same languages as the texts of the model collection. By the method of the nearest (in terms of distance) neighbor, all new texts confirmed their homogeneity with the corresponding pairs of monolingual works. The sixth text in Romanian showed its heterogeneity in relation to all elements of the collection. At the same time, it showed closeness in minimum distances, first of all, to two texts in Spanish and then to two works in Italian.

Key words: text, language, Latin graphics, alphabet, frequency of unigrams, digital portrait of text, hypothesis of homogeneity, classifier, teaching, language recognition, testing of the classifier, performance evaluation

REFERENCES

1. List of Latin letters. *Wikipedia*. (In Russian). Available at: https://ru.wikipedia.org/wiki/Список_латинских_букв (accessed 13.11.2020).

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. Usmanov Z.D., Kosimov A.A. O raspoznavanii avtorstva tadzhikskogo teksta [On authorship identification of a text written in Tajik]. *Doklady Akademii nauk Respubliki Tadzhikistan = Reports of the Academy of Sciences of the Republic of Tajikistan*, 2016, vol. 59, no. 3–4, pp. 114–119.

[*] *Received 10 December 2020.*

4. Kosimov A.A. Otsenka effektivnosti ispol'zovaniya bigramm pri identifikatsii teksta [Evaluation of the efficiency of using bigrams in text identification]. *Doklady Akademii nauk Respubliki Tadzhikistan = Reports of the Academy of Sciences of the Republic of Tajikistan*, 2017, vol. 60, no. 5–6, pp. 224–229.

5. Kosimov A.A. Otsenka effektivnosti ispol'zovaniya trigramm pri identifikatsii teksta [Evaluation of trigramm use efficiency for a text identification]. *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*, 2017, no. 1 (166), pp. 51–57.

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

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

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

9. Ashurova Sh.N. [Assessment of the effectiveness of the use of verbal bigrams in the identification of text]. *Materialy mezhdunarodnoi nauchno-prakticheskoi konferentsii TUT "Rol' IKT v innovatsionnom razvitii ekonomiki Respubliki Tadzhikistan"* [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).

10. Ashurova Sh.N. Otsenka effektivnosti ispol'zovaniya slovesnykh trigramm pri identifikatsii teksta [Evaluation of the effectiveness of the use of verbal trigrams in text identification]. *Vestnik Tekhnologicheskogo universiteta Tadzhikistana = Bulletin of the Technological University of Tajikistan*, 2017, no. 4 (31), pp. 51–58. (In Russian).

11. Ashurova Sh.N., Toshkhudzhaev Kh.A. O raspoznavanii avtora teksta na osnove chastotnosti slovesnykh bigramm [On recognition of the author of the text based on the frequency of verbal bigrams]. *Politekhnicheskii vestnik. Seriya: Intellekt. Innovatsii. Investitsii = Polytechnic Bulletin. Series: Intelligence. Innovation. Investments*, 2020, no. 2 (50), pp. 57–61.

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

13. 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.

14. 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.

15. Vorontsov K.V. *Matematicheskie metody obucheniya po pretsedentam* [Mathematical methods of teaching by precedents]. Available at: http://www.ccas.ru/voron (accessed 01.11.2020).

16. 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 sys-

tems (Workshop on the computer of the Department of Mathematical Forecasting Methods)]. Moscow, Faculty of CMC, MSU Publ., 2010. 278 p.

Для цитирования:

Усманов З.Д., Косимов А.А. Тестирование γ-классификатора, настроенного на распознавание языков произведений на основе латинского алфавита // Системы анализа и обработки данных. – 2021. – № 2 (82). – С. 83–94. – DOI: 10.17212/2782-2001-2021-2-83-94.

For citation:

Usmanov Z.D., Kosimov A.A. Testirovanie γ-klassifikatora, nastroennogo na raspoznavanie yazykov proizvedenii na osnove latinskogo alfavita [Testing the γ-classifier adapted to recognize the languages of works based on the Latin alphabet]. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*, 2021, no. 2 (82), pp. 83–94. DOI: 10.17212/2782-2001-2021-2-83-94.