Правительство Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет»

Кафедра информационно-аналитических систем

Дюрдева Полина Сергеевна

Автоматическое определение автора текста на основе распределения частот буквосочетаний

Бакалаврская работа

Допущена к защите.
ИО Зав. кафедрой:
к. ф.-м. н., доцент Михайлова Е. Г.

Научный руководитель:
к. ф.-м. н., доцент Михайлова Е. Г.

Рецензент:
научный сотрудник Центра речевых технологий Бояров А. А.

Санкт-Петербург
2016

SAINT-PETERSBURG STATE UNIVERSITY

Analytical Information Systems Chair

Polina Dyurdeva

Automated writer identification based on letter frequency distribution

Bachelor's Thesis

Admitted for defence.
Acting head of the chair:
associate professor Mikhaylova Elena

Scientific supervisor:
associate professor Mikhaylova Elena

Reviewer:
research scientist of Speech technology center Andrei Boiarov

Saint-Petersburg
2016

Оглавление

Введение

В связи с увеличением доступности и распространения текстовых документов в электронной форме увеличилась важность использования автоматических методов для анализа содержания документов. К задачам анализа текста можно отнести задачи классификации и кластеризации документов по различным критериям, например, жанру, эпохе написания, формату (роман, эссе, очерк), эмоциональной окраске, стилю речи, а также задачу определения автора текста.

С упрощением доступа к различным данным, расширением возможности поиска, копирования и распространения данных в сетях становится актуальной задача идентификации автора. Равным образом, вопросы, связанные с установлением авторства, являются важными в лингвистических, исторических и криминалистических исследованиях. Общедоступность электронных устройств позволяет отодвинуть распознавание автора с привлечением большого числа экспертов на второй план, ускорить и упростить этот процесс посредством его автоматизации.

Понятие идентификация автора определяется как процесс установления автора по множеству общих и частных признаков текста, составляющих авторский стиль.[6]

В существующих системах определения авторства текста пользуются популярностью статистические методы, основанные на поиске «авторского инварианта». «Авторский инвариант» характеризует языковую особенность (лексическую, грамматическую, фразеологическую и другую) текста. В качестве инварианта могут выступать: доля гласных или согласных, частота употребления определенной части речи, вероятность переходов от одной части речи к другой, «любимые» слова, информационная энтропия и так далее. В 2010 году, Ю.Н. Орловым и К.П. Осмининым был предложен статистический метод определения автора и жанра текста, основанный на распределении частот буквосочетаний (n-грамм)[7],[8]. Этот метод показал достойные результаты для произведений русской литературы. К сожалению, точность статистических методов определения авторства сильно зависит от специфики

используемых данных: от языка, на котором написаны тексты, от стиля речи текста, и, прежде всего, от длин текстов, на которых проводят исследование. В силу этого затруднительно делать вывод о точности такого подхода на данных другой природы. По этой причине целью настоящей работы являлся анализ применимости такого математического аппарата, как распределение частот буквосочетаний для разных языков при решении задачи установления авторства текстов, имеющих различные длины и написанных в разном стиле речи.

Данное исследование не предполагает решение задачи идентификации автора в полном объеме по той причине, что отличие авторских черт носит субъективный характер и зависит от ограничений, накладываемых на творческий процесс автора. Однако в итоге программная система, реализующая подобные методы, способна давать рекомендации о возможном авторстве текста.

1. Постановка задачи

В рамках дипломной работы были поставлены следующие задачи:

- Изучение метода идентификации автора на основе статьи [7]
- Выделение особенностей (численных характеристик), присущих определенному автору/стилю
- Реализация классификации текстовых документов по авторскому стилю
- Установление зависимости точности классификации от длин текстов, числа тренировочных данных, числа распознаваемых авторов
- Апробация и анализ результатов классификации текстов, написанных на русском, английском и немецком языках
- Сравнение результатов классификации текстов, написанных в художественном и публицистическом стилях
- Реализация кластеризации текстовых документов с помощью методов K-Means, Global k-means, PAM
- Оценка результатов кластеризации по мерам F-Measure, Purity, Rand-Index, NMI
- Оценка применимости подхода на основе частот буквосочетаний для задач идентификации и кластеризации текстовых документов.

2. Обзор существующих аналогов

В существующих на сегодняшний день системах определения авторства текста применяются различные подходы из теории математической статистики, распознавания образов и теории вероятностей, алгоритмы кластерного анализа, нейронных сетей и другие. Системы отличаются методом идетификации автора, средством анализа текста, требуемым объемом текста и точностью. Ниже приведен краткий обзор некоторых существующих программных продуктов [11].

- «Лингвоанализатор»
  Методы: марковские цепи, информационная энтропия
  Средство анализа текста: статистический анализ частот встречаемости графем
  Требуемый объем текста: 40 – 100 тысяч символов
  Точность: 84 – 89 %

- «Автороведˮ
  Методы: нейронные сети, метод опорных векторов
  Средство анализа текста: статистический анализ наиболее часто встречающихся триграмм и слов русского языка
  Требуемый объем текста: 20 – 25 тысяч символов
  Точность: 95 – 98 %

- «Стилеанализатор»
  Методы: нейронные сети, марковские цепи, деревья решения, меры расстояний
  Средство анализа текста: статистический анализ частот встречаемости букв, слов и других элементов текста
  Требуемый объем текста: 30 – 40 тысяч символов
  Точность: 90 – 98 %

Методы установления авторства текста, основанные на подсчете появления каких-либо характеристик текста (служебных частей речи (предлогов, союзов, частиц), самостоятельных частей речи (существительных, глаголов, прилагательных), длин слов, длин предложений) также

различаются по мерам сравнения полученных частот. Наиболее употребляемыми мерами сравнения текстов являются:

- Информационная энтропия
- Информация Фишера
- Мера «хи-квадрат»
- Мера Кульбака-Лейблера.

3. Описание используемого метода

При идентификации автора текста предполагается, что текст отображает индивидуальную манеру письма автора, которая позволяет отличить его от других. Чтобы сравнивать тексты между собой необходимо сопоставить тексту некоторую числовую характеристику, которая была бы близка для текстов одного и того же автора, и заметно различалась бы для произведений разных авторов. В качестве такой характеристики Ю.Н. Орлов и К.П. Осминин в статье [7] используют плотность функции распределения (ПФР) буквосочетаний из трех подряд идущих символов (3-грамм). ПФР определяется как совокупность эмпирических частот встречаемости букв или их сочетаний. При анализе текста с помощью ПФР не учитываются вхождения знаков препинания, пробелов и цифр.

Задача идентификации автора неизвестного текста в терминах ПФР формулируется следующим образом.

Дан некоторый набор текстов, в котором содержатся произведения $\mathcal{A}$ известных авторов. Пусть K_a – количество произведений a-го автор. $N_{i,a}$ – количество символов в i-ом произведении a-го автора, $i = 1, \ldots, K_a$. Все тексты в данном наборе будут представлены в виде ПФР. ПФР текста, объем которого равен $N_{i,a}$, задается как множество значений $f_{i,a}(j) = k_j/N_{i,a}$, k_j – число встречаемости n-граммы под номером j. Аргумент $j = 1, \ldots, \alpha(n, M)$, соответствует номеру буквосочетания (n-граммы) при алфавитном упорядочивании, где M- мощность алфавита языка, на котором написан текст, n – порядок n-граммы, т.е количество символов в буквосочетании. $\alpha(n, M) = M^n$ - число n-грамм в данном алфавите.

Каждый автор отождествляется с его средневзвешенной ПФР, которая задается по формуле (1):

$$F_a(j) = \frac{1}{N_a} \sum_{i=1}^{K_a} f_{i,a} N_{i,a} \quad (1)$$

$$N_a = \sum_{i=1}^{K_a} N_{i,a} \quad (2)$$

Эти ПФР будут играть роль авторских эталонов [7][8][9].

Чтобы сравнивать либо два текста, либо текст и авторский эталон, необходимо задать расстояние между соответствующими функциями распределения. В качестве метрики расстояния используется норма в пространстве суммируемых функций. Так, например, расстояние $p_{0,a}$ между ПФР неизвестного текста f_0 и какой-либо авторской ПФР F_a будет вычислено по формуле (3):

$$p_{o,a} = ||f_0 - F_a|| = \sum_{j=1}^{\alpha(n,M)} |f_0(j) - F_a(j)| \quad (3)$$

Соответственно, текст «0» будет принадлежать тому автору, расстояние до ПФР которого будет наименьшим.

При решении задачи классификации набор данных не разбивался явно на тестовое и тренировочное множество. Средневзвешенные ПФР строились по всему множеству книг одного автора. Расстояние от книги i до «своего» автора a вычислялось по формуле (4):

$$p_{i,a} = \frac{||f_{i,a} - F_a||}{1 - N_{i,a}/N_a} \quad (4)$$

Формула (4) позволяет исключить участие ПФР книги i в средней ПФР «своего» автора.[7]

4. Тестовые данные

Настоящая работа включала в себя несколько серий экспериментов.

Первая серия экспериментов проводилась на литературных произведениях, написанных на русском, английском и немецком языках. Далее более подробно описан каждый набор данных, соответствующий одному из вышеперечисленных языков.

В Таблицах 1, 2, 2 представлена информация по авторам русской, английской, немецкой литературы. Для каждого автора представлены следующие характеристики его произведений:

- max – максимальная длина текста (тыс. знаков)
- min - минимальная длина текста (тыс. знаков)
- avg – средняя длина текста (тыс. знаков)
- std – среднеквадратическое отклонение (тыс. знаков)
- ls – среднее расстояние от книги до «своего» авторского эталона
- lo – среднее расстояние от книги до «чужого» авторского эталона

Русские текстовые данные включали в себя произведения 10 авторов IX-XX веков. У каждого автора случайным образом было выбрано 10 произведений (на данную выборку не было наложено никаких ограничений, в ней присутствуют произведения разного объема, жанра и типа).

Тестирование алгоритма на англоязычных текстах производилось на 10 авторах. Для эксперимента использовались 6 книг каждого автора, выбранные случайным образом. В данной выборке были произведения преимущественно жанров фантастика, приключения и роман.

Объем тестовых данных для каждого автора при тестировании немецкой литературы был равен 6 книгам. Были использованы тексты 5 писателей.

Автор	avg	max	min	std	ls	lo
Ч.Т. Айтматов	263	677	48	235	0.142	0.162
М. А. Булгаков	207	593	42	172	0.147	0.169
Ф.М. Достоевский	702	1338	208	417	0.092	0.127
Н.В. Гоголь	163	391	38	124	0.152	0.172
Н.С. Лесков	223	966	32	292	0.162	0.176
Д.Н. Мамин-Сибиряк	336	679	45	240	0.129	0.153
А.Н. Островский	84	131	59	232	0.193	0.219
Л.Н. Толстой	620	1336	184	392	0.105	0.136
И.С. Тургенев	256	517	92	121	0.111	0.136
А.П. Чехов	72	265	8	88	0.2342	0.2343

Таблица 1: Характеристики произведений русских писателей

Автор	avg	max	min	std	ls	lo
J.K. Rowling	725	1147	377	277	0.07	0.118
C. Clarke	363	464	295	69	0.086	0.117
R. J. Zelazny	246	300	187	43	0.091	0.117
S. Meyer	579	835	166	241	0.065	0.119
T. Pratchet	271	385	150	111	0.103	0.133
W. Burroughs	329	411	253	64	0.088	0.126
M. Twain	281	497	10	190	0.132	0.152
A. Doyle	151	334	29	137	0.128	0.149
C. Dickens	402	682	35	267	0.09	0.121
R. Stevenson	306	439	146	105	0.1	0.123

Таблица 2: Характеристики произведений английских писателей

Author	avg(тыс.зн)	max	min	std	ls	lo
J. Wassermann	346	677	195	185	0.079	0.108
J.Goethe	363	569	37	207	0.106	0.123
R. Walser	206	433	60	185	0.103	0.118
P. Heyse	76	158	31	45	0.125	0.135
T. Storm	67	118	38	32	0.140	0.149

Таблица 3: Характеристики произведений немецких писателей

Вторая серия экспериментов проводилась на англоязычных публицистических текстах. В качестве исходных данных был взят корпус

«Reuter_50_50 Data Set» [4]. Этот набор данных содержит новостные статьи на тему «Промышленность», что позволяет свести к минимуму фактор тематического различия между текстами. Этот корпус составлен из 5000 статей 50 авторов. Объем статей в среднем составляет 2455 знаков со стандартным отклонением 832 знака. Нужно отметить, что разброс длин текстов достаточно большой: в корпусе присутствуют статьи длиной более 4 тысяч и менее 300 символов.

5. Решение задачи классификации

Алгоритм «близости ПФР» для решения задачи классификации показал достойные результаты на литературных произведениях. В Таблице 4 приведены результаты классификации для русских, английских и немецких текстов. R – результат, который представлен в виде: c/n, где c – количество книг автора, которые были верно распознаны, а n – общее число книг автора.

Автор	R	Автор	R	Автор	R
Айтматов	9/10	Rowling	6/6	Wassermann	6/6
Булгаков	9/10	Clarke	6/6	Goethe	5/6
Достоевский	10/10	Zelazny	6/6	Walser	6/6
Гоголь	10/10	S. Meyer	6/6	P. Heyse	5/6
Лесков	8/10	T. Pratchet	6/6	T. Storm	2/6
Мамин-Сибиряк	9/10	Burroughs	6/6		
Островский	10/10	Twain	4/6		
Толстой	9/10	Doyle	5/6		
Тургенев	10/10	Dickens	6/6		
Чехов	1/10	Stevenson	5/6		

Таблица 4: Результаты классификации

На основе экспериментов были получены оценки точности для русского, английского и немецкого языков, представленные в Таблице 5. $r(w)$ – число книг, автор которых был определен верно (неверно), точность – процент верно распознанных книг.

Язык	r	w	точность (%)
Русский	85	15	85
Английский	56	4	93
Немецкий	22	8	73

Таблица 5: Общий результат классификации

Как видно из Таблицы 4, алгоритм продемонстрировал наименьшую точность при распознавании автора произведений А.П.Чехова, Т. Шторма и М. Твена, что может быть объяснено следующими фактами:

- Средние длины текстов Чехова, Шторма меньше средних длин текстов остальных авторов

- Минимальные длины текстов Чехова, Шторма и Твена не превосходят минимальных длин текстов остальных авторов
- Наблюдается значительный разброс длин рассматриваемых текстов этих авторов.

Ввиду вышеизложенных фактов возникает необходимость установления взаимосвязи между точностью классификации и длинами текстов. Для установления взаимосвязи были выбраны следующие данные:

- 55 русских текстов длиной более 60 тысяч знаков
- 42 английский текстов длиной более 95 тысяч знаков
- 30 немецких текстов длиной более 30 тысяч знаков.

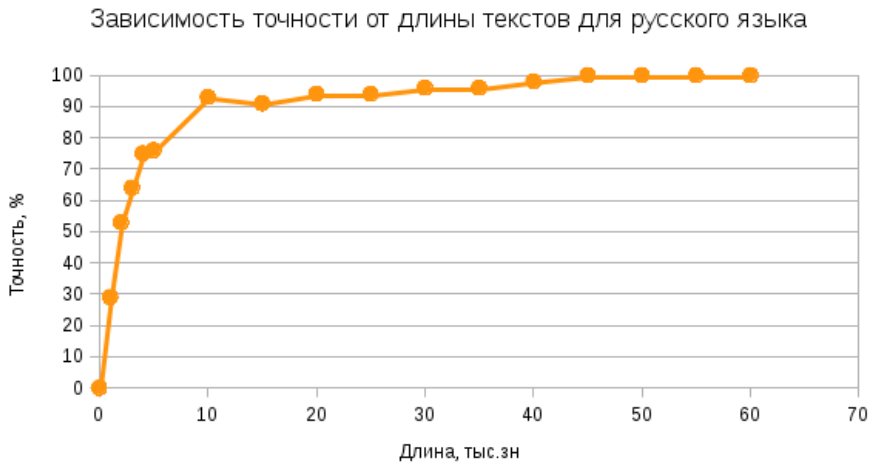
Точность на объемах, превосходящих заданные значения, имеет незначительное отклонение от точности классификации на полных объемах произведений.

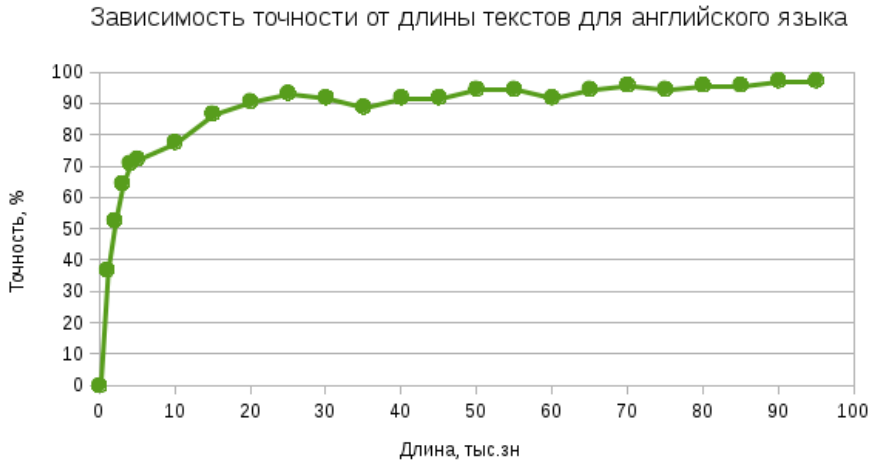
Взаимосвязь длин текстов и результатов классификации представлена на Рисунке 1.

Из графиков на Рисунке 1 видно, что:

1. Приведение текстов к одной длине позволяет увеличить точность классификации
2. Точность 80 – 100 % достигается на объемах, превышающих 10 тысяч знаков
3. Точность 60 – 75 % достигает на текстах длины 2 – 4 тысячи знаков.

Замечание 2 является достаточным основанием для проведения экспериментов на относительно небольших текстах публицистического жанра.





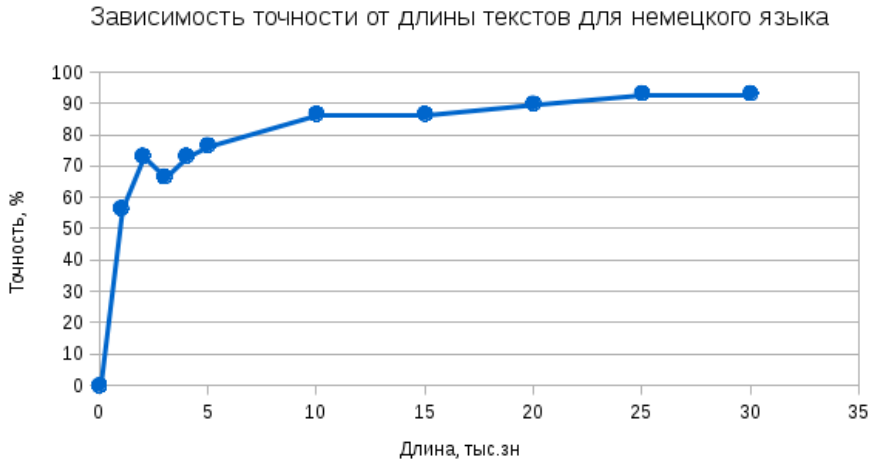


Рис. 1: Зависимость точности классификации от длины текста

Результаты второй серии экспериментов представлены в Таблице 6, где A – число авторов, N – число текстов для одного автора, $r(w)$ – число книг, автор которых был определен верно (неверно), «точность» – процент верно распознанных книг.

A	N	r	w	точность (%)
50	50	1695	805	67,8
50	100	3370	1630	67,4

Таблица 6: Результаты классификации для публицистических текстов

В результате эксперимента была получена точность в 67 %. Этот факт подтверждает, что точность зависит преимущественно от длин классифицируемых текстов, а не от стиля речи.

Еще один фактор, который может повлиять на точность классификации – число тренировочных данных, то есть число текстов, составляющих авторский эталон.
Как видно из Таблицы 6, при экспериментах на публицистических текстах точность в 67 % была получена при числе книг равном 50 и 100 для каждого автора.
Следовательно, число книг, составляющих авторский эталон, было равно 49 и 99 при сравнении текста с эталоном «своего» автора, 50 и 100 при сравнении текста с эталоном «чужого» автора.
Последний факт создает проблему трудоемкого поиска авторских данных в вышеуказанных объемах, так как число публицистических текстов, принадлежащих одному автору при решении реальных задач, может быть намного меньше требуемых.

Поэтому возникла необходимость в исследовании зависимости точности классификации от числа текстов, составляющих авторский эталон. На Рисунке 2 представлен график этой зависимости. Исследование проводилось на 50 и 10 авторах. При проведении данного эксперимента наборы данных были поделены на тренировочное и тестовое множества. Размер тестового множества изменялся в ходе эксперимента. Размеры тренировочных наборов составляли 1000 книг для 50 авторов и 200 книг для 10 авторов.

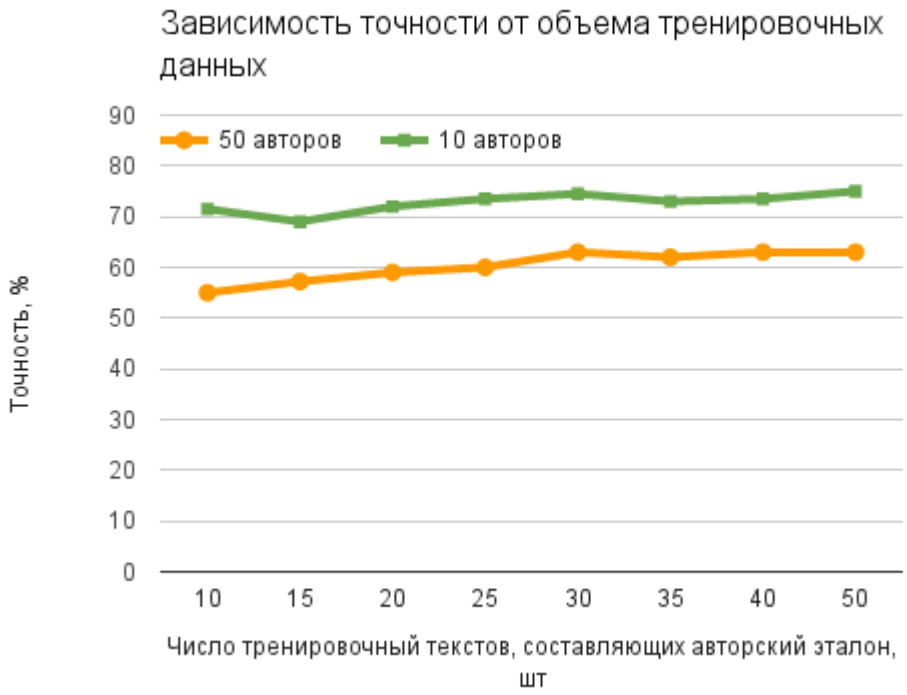


Рис. 2: Зависимость точности классификации от числа текстов, входящих в авторский эталон. Публицистические тексты

В ходе эксперимента было обнаружено, что:

1. С увеличением числа авторов точность заметно падает
2. Максимальная точность достигается при наличии 30 тренировочных текстов для каждого автора.

Некоторые из существующих систем установления авторства обладают низкой разделительной способностью в случае большого числа авторов. Следующий эксперимент был проведен для того, чтобы установить потенциальное число авторских стилей, которое метод сравнения ПФР способен различать при классификации.
Для эксперимента были выбраны публицистические тексты длиной более 2 тысяч символов. Был проведен ряд тестов для 5, 10, 15 и 20 авторов при различном количестве текстов на одного автора. Начальное число текстов для каждого автора было равно 5, затем число книг у каждого автора последовательно увеличивалось. На Рисунке 3 представлены результаты.

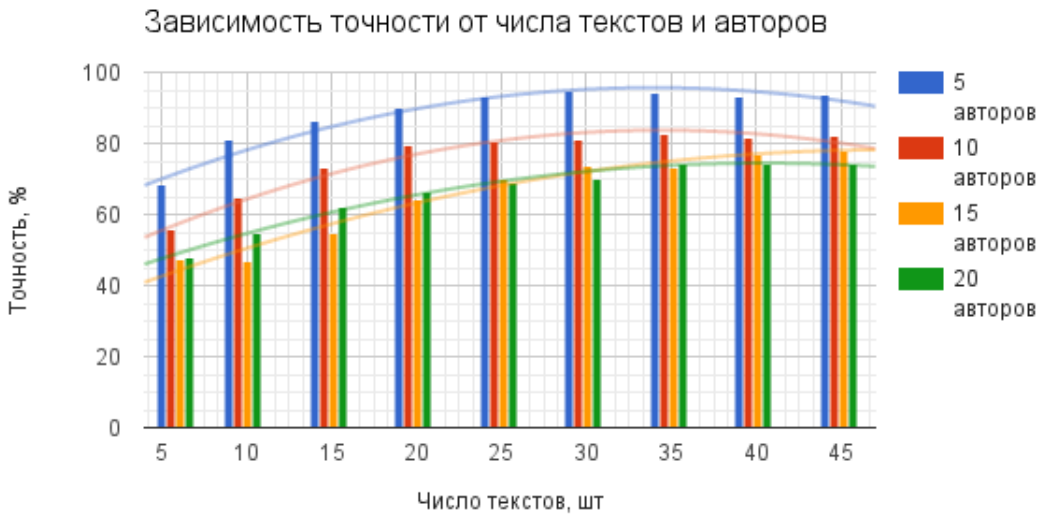


Рис. 3: Зависимость точности классификации от числа текстов

Проведенный эксперимент показал, что тексты длиной 2 – 4 тысячи знаков распознаются с точностью 80 – 95 % для небольшого числа авторов (5 – 10). Оптимальное число тренировочных данных равно 30 – 35 текстам для одного автора, при этом дальнейшее увеличение объема тренировочного множества не приводит к значительному увеличению точности классификации.

По проведенным выше экспериментам можно сделать следующие выводы о методе сравнения ПФР для решения задачи определения автора:

- Метод сравнения ПФР слабо зависит от языка и стиля текста
- Метод чувствителен к числу тренировочных данных и к длинам текстов
- Авторство текстов, длины которых мало отличаются между собой, методом сравнения ПФР определяется точнее.

Наиболее показательные результаты по проведенным экспериментам вынесены в Таблицу 7.

Длины текстов	Число авторов	Число трен.текстов	Точность (%)
$>$ 20 000	-	$>$ 5	90 – 100
$>$ 10 000	-	$>$ 5	80 – 100
2 000 – 4 000	5	25 – 30	90 - 95
2 000 – 4 000	5 – 10	30 – 35	80 – 95
2 000 – 4 000	15 – 20	25 – 45	60 – 80

Таблица 7: Результаты тестирования метода сравнения ПФР для решения задачи определения автора текста

6. Решение задачи кластеризации

Процесс кластеризации подразумевает группирование входных данных по группам таким образом, чтобы все элементы внутри каждой группы были похожи друг на друга в определенном смысле. Рассмотрим набор текстовых документов C. Задача кластеризации текстов заключается в разбиении C на группы C_1,C_2,...,C_k так, что каждый C_i содержит как можно больше произведений определенного писателя и как можно меньше произведений иных авторов.[10]

Разбиение на группы основано на расчете расстояния между элементами кластеров. Расстояние между двумя произведениями будет определяться, как и ранее, в пространстве суммируемых функций между ПФР двух текстов:

$$p_{i,k} = \| f_i - f_k \| = \sum_{j=1}^{\alpha(n,M)} |f_i(j) - f_k(j)|$$

Где i и k какие либо произведения из набора C.

Во многих алгоритмах кластеризации используется понятие центроид – «центр тяжести» кластеров. Поскольку в качестве объектов кластеризации используются ПФР, в роли центроида будет выступать средневзвешенная ПФР, вычисленная по ПФР входящим в один кластер. В использованных в настоящей работе алгоритмах кластеризации число кластеров k инициализируется до начала работы алгоритма числом равным количеству авторов.

В ходе работы был протестирован процесс кластеризации посредством алгоритмов:

1. K-means: Основная идея алгоритма заключается в том, что на каждой итерации перевычисляется центроид для каждого кластера. Затем элементы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике расстояния. Алгоритм завершается, когда не происходит изменений в центроидах.[5]

2. Global-K-Means: Данный алгоритм не зависит от случайного выбора начальных элементов в качестве центроидов, в отличие от алгоритма k-means. Каждый шаг GlobalK-Means подразумевает решение задачи кластеризации с числом кластеров $n = 1, \ldots, k$ методом k-means. Лучшее решение шага $n - 1$ становится инициализирующим значение центроидов для шага n. [1]

3. PAM: В отличие от k-means кластеризация происходит на основе попарных расстояний между элементами. Перераспределение элементов происходит не относительно центров, а относительно медоидов кластеров. [2]

Каждый из этих алгоритмов, ввиду специфичности кластеризуемых данных, был реализован и адаптирован под ПФР.

По причине отсутсвия универсальной метрики кластеризации, для оценивания результатов работы алгоритмов было выбрано несколько метрик[3]:
Пусть $Q = \{q_1, q_2, \ldots, q_k\}$ – набор кластеров, $C = \{_1, _2, \ldots, _j\}$ - набор классов (текстов авторов), N – число кластеризуемых текстов.

- Purity Measure

$$purity(Q; C) = \frac{1}{N} \sum_k \max_j |q_k \cap c_j|$$

  Метрика чистоты (Purity Measure) показательна при небольшом количестве кластеров.

- NMI (normalized mutual information)

$$NMI(Q; C) = \frac{2I(Q, C)}{H(Q) + H(C)}$$

  Взаимная информация $I(Q; C)$ и энтропия $H(Q)$ определяются следующим образом:

$$I(Q; C) = \sum_k \sum_j \frac{|q_k \cap c_j|}{N} \log \frac{N|q_k \cap c_j|}{|q_k||c_j|}$$

$$H(Q) = -\sum_k \frac{|q_k|}{N} \log \frac{|q_k|}{N}$$

Нормализованную взаимную информацию (NMI) можно использовать для сравнения кластеризации при разном числе кластеров, так как эта метрика нормализуемая.

- Rand Index

$$RI = \frac{TP + TN}{TP + TF + TN + FN}$$

$TP(true\ positive)$ – число пар текстов одного автора лежащих в одном кластере
$TN(true\ negative)$ – число пар текстов разных авторов лежащих в разных кластерах
$FP(false\ positive)$ – число пар текстов разных авторов лежащих в одном кластере
$FN(false\ negative)$ – число пар текстов одного автора лежащих в разных кластерах
Коэффициент Рэнд (Rand Index) измеряет процент решений, которые являются правильными. При его вычислении учитываются как ложно-позитивные, так и ложно негативные решения.

- F-Measure

$$F = \frac{Precision * Recall}{Precision + Recall}$$

Точность (Precision) и полнота (Recall) определяются следующим образом:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

F-мера (F-measure) дополнительно позволяет расставлять разные приоритеты для различного рода ошибок, например, более строго относиться к ложно-негативным, чем ложно-позитивным.

Все метрики имеют близкое к 0 значение при плохой кластеризации, при правильной кластеризации они равны 1. Специфика метрик позволяет оценить кластеризацию с разных сторон.

При тестировании кластеризации было произведено несколько серий экспериментов для всех алгоритмов, упомянутых выше. Каждая серия проходила с определенным числом авторов N.

N	Algorithm	Purity	RandIndex	NMI	F-Measure
Русские тексты					
5	k-means	90	75	81	78
5	Global k-means	93	88	91	90
5	PAM	50	45	66	50
8	k-means	73	36	55	53
8	Global k-means	78	40	59	55
8	PAM	73	32	41	43
10	k-means	76	30	44	42
10	Global k-means	78	33	45	50
10	PAM	63	23	30	40
Английские тексты					
5	k-means	90	72	79	83
5	Global k-means	95	88	94	92
5	PAM	92	80	89	83
8	k-means	90	60	78	75
8	Global k-means	91	64	80	76
8	PAM	89	57	73	73
10	k-means	90	53	74	63
10	Global k-means	89	55	73	69
10	PAM	83	40	69	55
Немецкие тексты					
5	k-means	81	55	65	67
5	Global k-means	85	69	67	70
5	PAM	76	43	73	55

Таблица 8: Результаты кластеризации

В Таблице 8 приведены результаты кластеризации на литературных текстах с использованием метода сравнения ПФР. Все три алгоритма показали высокую точность для небольшого числа (5–8) кластеров. Алгоритм Global k-means показал наибольшую точность. Однако, с возрастанием числа авторов точность кластеризации падает.

7. Заключение

В результате проделанной работы была создана программная реализация алгоритма для автоматической идентификации автора текста в электронной форме. Программа использовала алгоритм, в основе которого лежал метод, который в своей работе[7] описали Ю.Н. Орлов и К.П. Осминин.

Программа была реализована на языке программирования Python с использованием оптимизированных библиотек. Программа содержит следующие основные модули: модуль загрузки текстов, модуль обработки тексов, модуль классификации, 3 модуля кластризации, соответсвующие выбранным алгоритмам, модуль сбора статистики по набору текстов (характеристики текстов, характеристики распределений расстояний между текстами и другие), модуль оценки кластеризации, тестировочный модуль и другие вспомогательные модули. Система имеет гибкую программную архитектуру, представляющую возможность быстро подменять/добавлять алгоритмы классификации и кластеризации, изменять порядок n (n-грамм) и другие параметры. Была проведена настройка системы для работы с русскими, английскими, немецкими текстами. В качестве данных для тестирования были взяты литературные произведения на соответствующих языках и публицистические англоязычные тексты. Было проведено порядка 200 экспериментов для анализа работы алгоритма для решения задачи классификации.

Были выявлены зависимости точности классификации от длин используемых текстов, числа тренировочных (эталонных) данных, числа распознаваемых авторов.

Также была реализована функциональность кластеризации входного множества документов с использованием алгоритмов кластеризации K-Means, Global K-Means, PAM. Была произведена оценка результатов кластеризации с помощью метрик Rand Index, NMI, F-Measure, Purity measure.

Таким образом, проведенный анализ работы метода на основе распределения частот буквосочетаний для решения задачи установления

авторства показал, что метод дает довольно хорошие результаты не только на длинных, но и на достаточно коротких текстах, что выгодно отличает его от большинства других методов. Ограничением этого метода при установлении авторства небольших текстов является низкая разделительная способность в случае большого числа авторов (высокая точность достигается не более чем на 10 авторах). К ограничениям также можно отнести требуемое количество тренировочных данных, которое является достаточно большим.

Для увеличения показателей точности алгоритма при решении задачи классификации можно предложить использование текстов приблизительно одинаковых длин. Таким образом, если тексты имеют большой разброс в длинах, то сокращение длин текстов до приемлемой длины (10 000 - 20 000 символов) может улучшить качество классификации.

Список литературы

[1] A. Likasa N. Vlassisb J. J. Verbeekb. The global k-means clustering algorithm. — Pattern Recognition, Vol. 36(2), 2003.

[2] A. P. Reynolds G. Richards, Rayward-Smith V. J. The Application of K-medoids and PAM to the Clustering of Rules. — Lecture Notes in Computer Science, Vol. 3177, 2004.

[3] Balani Naveen. Evaluation of clustering. — http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html : Cambridge University Press, 2009.

[4] Reuter_50_50 Data Set. — archive.ics.uci.edu/ml/datasets/Reuter_50_50, 2011.

[5] Teknomo K. K-Means Clustering Tutorials. — people.revoledu.com tutorial 2007.

[6] А.С. Романов. Методика и программный комплекс для идентификации автора неизвестного текста. — 2010.

[7] Борисов Л.А. Орлов Ю.Н. Осминин К.П. Идентификация автора текста по распределению частот буквосочетаний. — Прикладная информатика Т. 26. No 2. С. 95-108, 2013.

[8] Орлов Ю.Н. Осминин К.П. Определение жанра и автора литературного произведения статистическими методами. — http://library.keldysh.ru/preprint.asp?id=2013-27 : Препринты ИПМ им.М.В.Келдыша No27. 26с, 2010.

[9] Орлов Ю.Н. Осминин К.П. Методы статистического анализа литературных текстов. — ЭдиториалУРСС/Книжныйдом «ЛИБРОКОМ», 2012.

[10] Павлов В.А. Дюрдева П.С. Шалымов Д.С. Кластеризация русскоязычных рукописей на основе графа отношения особенностей. — Компьютерные инструменты в образовании № 1: 24–35, 2016.

[11] Т.В. Батура. Формальные методы определения авторства текстов. — SSN 1818-7900. Вестник НГУ. Серия Информационные технологии. Том 10, выпуск 4, 2012.