УДК 81'322.2

АНАЛИЗ НЕОДНОРОДНОСТЕЙ В ТЕКСТЕ НА ОСНОВЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ЧАСТЕЙ РЕЧИ

Седов А.В., Рогов А.А.

ГОУ ВПО «Петрозаводский государственный университет», Петрозаводск, Россия (185000, г. Петрозаводск, пр. Ленина, 31), e-mail: sedov_a@mail.ru, rogov@psu.karelia.ru

Рассматривается способ выделения в тексте неоднородных фрагментов. В качестве неоднородного фрагмента понимается часть текста, состоящая из нескольких предложений или абзацев, которая отличается от остального текста по некоторым признакам. В качестве признаков выступают различные последовательности частей речи. Рассматриваются последовательности как в рамках одного предложения, так и принадлежащие различным предложениям и абзацам. Предлагается алгоритм анализа фрагментов. Для его работы требуется произвести грамматический разбор текстов, осуществить выбор размера фрагментов и выбор признаков, на основании которых производится анализ. На основе подсчёта статистики хи-квадрат для всех возможных фрагментов и для оставшихся частей текста принимается решение об однородности/неоднородности фрагментов. В ходе исследования рассматриваются тексты Ф.М. Достоевского, А. Григорьева, В. Даля. В результате для текстов определяются последовательности, которые выделяют из текстов фрагменты. Данный алгоритм был применён для атрибуции текстов. Из всего набора признаков были выделены те, которые разделяли тексты Ф.М. Достоевского от остальных. Исключение составил текст В. Даля, который был достаточно мал (полтора размера фрагмента).

Ключевые слова: анализ текста, неоднородность, атрибуция текстов.

DETECTION HETEROGENEITY IN TEXTS USING PART OF SPEECH SEQUENCES

Sedov A.V., Rogov A.A.

Petrozavodsk State University, Petrozavodsk, Russia (185000, Petrozavodsk, Lenin street, 31), e-mail: sedov_a@mail.ru, rogov@psu.karelia.ru

In the article considered the way of selection heterogeneity fragments in the texts. As heterogeneous fragment understood part of the text, that consists of a few sentences or paragraphs, which differs from the rest of the text in some respects. The sequences of parts of speech are used as parameters of the selection. Considering sequences within a single sentence, and belonging to different sentences and paragraphs. The algorithm for the analysis of fragments is proposed. It requires performing parsing text, determination of the size of the fragments and select features on which to analyze. Using the calculation of chi-square statistic between frequencies of sequences in certain fragment and the whole text for all possible fragments the decision about homogeneity can be decided. For the investigation we used texts of Dostoevsky, A. Grigoriev, V. Dal. As a result, in the texts we have found sequences thathelped us to isolate fragments from the text. These sequencesalso have been applied for the attribution of texts. Of the entire set of attributes were identified by those who shared texts Dostoevsky from the rest. The exception was the text of the V. Dal, which was quite small (had half the size of the fragment).

Key words: text analysis, heterogeneity, text attribution.

Введение

В настоящее время существует большое количество текстов, которые содержат неоднородные включения. Это студенческие и псевдонаучные работы, скомпилированные из разных источников, а также тексты, подвергшиеся существенному редактированию. Зачастую данные тексты выдаются как собственные, уникальные, написанные непосредственно самим автором. Поэтому возникают задачи отделения фрагментов, написанных самостоятельно, от скопированных [7] фрагментов, а также нахождения автора, или первоисточника [3]. Кроме того, существует задача поиска в литературном тексте фрагментов с разной эмоциональной окраской [2; 8].

Существующие современные системы, такие как системы обнаружения плагиата, существенно опираются на базы текстов. Если по каким-либо причинам текст, который использовался при создании, не вошёл в поисковые базы, то система может принять фрагмент данного текста как уникальный. Следовательно, методы, которые позволяют выявлять неоднородные фрагменты в тексте, и тем самым указывать на возможность плагиата без привязки к базам данных, являются актуальными и своевременными.

В данной работе предлагается алгоритм поиска фрагментов, имеющих отличную от основного текста синтагматику, характеризующуюся определённой последовательностью составляющих элементов – слов с частеречной принадлежностью. В основе алгоритма лежит статистика частоты встречаемости последовательностей частей речи, состоящих из трёх или четырёх слов.

В данной статье рассматривается непосредственно сам алгоритм нахождения неоднородных фрагментов на основе последовательностей частей речи, предложен один из способов выбора последовательностей для анализа, рассмотрены примеры, а также возможность применения данного алгоритма к атрибуции текстов. Работа выполняется при финансовой поддержке Программы стратегического развития ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности.

Алгоритм поиска неоднородности фрагмента текста

1. *Разметка исходного текста.*

С помощью какого-либо грамматического анализатора производится разметка исходного текста. В качестве такого анализатора можно взять морфологический анализатор mystem [5] от компании «Яндекс». В результате исходный текст будет представлять собой последовательность частей речи, разделённых знаками препинания, концами абзацев и т.п. Поиск неоднородных фрагментов можно производить как с учётом этих знаков, так и без их учёта.

2. *Выбор размера фрагмента.*

Размер фрагмента можно задавать количеством слов или предложений. Выбор размера зависит от цели исследования. В производимых экспериментах величина фрагмента изменялась от одного до пятнадцати предложений.

3. *Выбор признаков.*

Для проведения исследования необходимо определиться с выбором исследуемой последовательности частей речи, которую в дальнейшем будем называть признаком. Среди множества вариантов последовательностей частей речи следует выбрать ту, которая обладает наибольшей информативностью.

Существует несколько способов выбора наиболее информативных признаков [1]. В качестве исследуемого признака можно выбрать тот, у которого статистика χ^2 имеет максимальную дисперсию. Для этого следует рассмотреть всевозможные наборы признаков (различные варианты последовательностей частей речи), исследуемый текст разбить на фрагменты. Для каждого выбранного признака и каждого фрагмента необходимо найти статистику χ^2 (способ вычисления статистики описан в следующем пункте), на основе которой высчитать дисперсию. В качестве исследуемого признака можно взять признак с максимальной дисперсией [1].

4. Вычисление статистики χ^2.

Для исследуемого фрагмента текста нужно сосчитать, сколько раз выбранная последовательность слов встречается в данном фрагменте и сколько в остальной части текста. Обозначим: p' – число выбранных последовательностей, p – общее число последовательностей во фрагменте, q – общее число последовательностей в оставшемся тексте, q' – число выбранных последовательностей в оставшемся тексте. Тогда статистика χ^2 имеет вид [6]:

$$\chi^2 = \frac{(p' - p \cdot r)^2}{p \cdot r} + \frac{(q' - q \cdot r)^2}{q \cdot r} + \frac{((p - p') - p \cdot (1 - r))^2}{p \cdot (1 - r)} + \frac{((q - q') - q(1 - r))^2}{q \cdot (1 - r)},$$

где $r = \frac{p' + q'}{p + q}$.

5. Поиск неоднородных фрагментов.

Исходный текст необходимо разбить на всевозможные фрагменты. Для каждого фрагмента вычислить статистику χ^2 (аналогично предыдущему пункту). Максимальное значение данной статистики будет соответствовать неоднородным фрагментам. При этом если значение превысит некоторое критическое значение, то отличие данного фрагмента от остальных будет статистически значимым с вероятностью P.

Пример выявления наиболее информативных признаков

Алгоритм поиска неоднородного фрагмента текста апробирован на разных текстах. В качестве примера приведем ниже выявление наиболее информативных признаков для произведения Ф.М. Достоевского «Дворянин». Рассмотрим все возможные фрагменты, состоящие из 5 предложений. Для каждого фрагмента и для каждой последовательности из четырёх частей речи была вычислена статистика χ^2. Для каждого признака подсчитана дисперсия. Часть результатов эксперимента, отсортированная по убыванию значения дисперсии, приведена в табл. 1. Максимальное значение дисперсии 142,96 соответствует четвёрке «Глагол – Существительное – Союз – Существительное». Этот признак для данного текста наиболее информативен.

Таблица 1 – Значения статистики χ^2 для различных признаков и фрагментов в произведении Ф.М. Достоевского «Дворянин»

Признаки				Номер фрагмента							Дисперсия
				1	2	3	4	5		70	
Глаг	Сущ	Союз	Сущ	0,1699	0,1772	0,1845	0,2234	0,2234	…	0,1555	142,9635
Союз	Прил	Сущ	Союз	0,1132	0,1180	0,1229	0,1488	0,1488	…	0,1036	113,8542
МодДис	Глаг	Предл	Мест	0,1132	0,1180	0,1229	0,1488	0,1488	…	0,1036	79,3861
Сущ	Глаг	Сущ	Союз	0,1132	0,1180	0,1229	0,1488	0,1488	…	0,1036	73,0946
Мест	МодДис	Глаг	Глаг	0,1699	0,1772	0,1845	0,2234	0,2234	…	0,1555	71,4637
Мест	Мест	МодДис	Глаг	0,1699	0,1772	0,1845	0,2234	0,2234	…	0,1555	70,6658
Без опис	Без опис	Без опис	Без опис	35,4120	33,9590	32,6147	26,9354	26,9354	…	0,1036	65,6831
МодДис	Глаг	Глаг	Предл	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	63,9424
Сущ	Мест	Мест	МодДис	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
Мест	Сущ	МодДис	Прил	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
Сущ	МодДис	Прил	Сущ	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
МодДис	Прил	Сущ	Мест	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
Прил	Сущ	Мест	Мест	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
МодДис	Глаг	Глаг	Мест	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
Глаг	Глаг	Мест	МодДис	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
Глаг	Мест	МодДис	Союз	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
Мест	МодДис	Союз	Прил	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	62,9922
…	…	…	…	…	…	…	…	…	…	…	…
Предик	Нар	Глаг	Нар	0,0566	0,0590	0,0614	0,0744	0,0744	…	0,0518	0,0011

Аналогичным образом в том же тексте был проведён эксперимент по выявлению наиболее информативных троек. Тройкой с наибольшей дисперсией стала «Глагол – Причастие – Существительное».

Примеры неоднородных фрагментов текстов

В результате применения алгоритма был найден фрагмент неоднородности. Далее приведен кусок текста из произведения Ф.М. Достоевского «Дворянин», в котором найденный фрагмент выделен курсивом.

Въ 45 № «Современной Лѣтописи Русскаго Вѣстника», за подписью г. К. Буха, напечатано было любопытное извѣстіе объ одномъ дворянинѣ, который хотѣлъ перечислиться въ государственные крестьяне. *Въ мензелинскомъ уѣздѣ оренбургской губерніи, въ селѣ Покровскомъ, по–народному Новая–Мазина, живетъ неслужащій дворянинъ симбирской губерніи Петръ Порфирьевичъ Мясоѣдовъ. Женившись на дочери государственнаго крестьянина того же села, Максима Андреева, онъ подалъ просьбу о причисленіи его въ государственные крестьяне, въ семейство тестя. Присутственное мѣсто, въ которое поступила его просьба, отказало ему, потому что по 619 статьѣ IX т. св. зак. гражд., изд. 1857 года, въ сельское состояніе могутъ быть причислены дѣти личныхъ дворянъ и приказнослужителей, не имѣющихъ оберъ–офицерскихъ чиновъ, а г. Мясоѣдовъ*

происходитъ отъ потомственной дворянской фамиліи. Отказъ этотъ въ точности соотвѣтствуетъ приведенной статьѣ закона. Но справедливо ли — спрашиваетъ далѣе авторъ — лишать человѣка возможности вступать въ ту среду, въ которой, по его понятіямъ и связямъ, онъ для себя находитъ болѣе выгодъ и удобствъ? Далѣе авторъ старается доказать, что не можетъ быть вреда отъ перечисленія человѣка, имѣющаго права дворянина, въ сословіе государственныхъ крестьянъ, и что не можетъ быть пользы отъ насильственнаго удержанія въ сословіи лица, которое отстало отъ того сословія и привычками, и образомъ жизни.

На обращеніе г. Буха къ юристамъ, на вызовъ его обсудить этотъ вопросъ, юристъ отыскался въ 49 № той же «Современной Лѣтописи Русскаго Вѣстника». Этотъ достопочтенный журналъ, соболѣзнующій о безжизненности нашей умственной среды, нашолъ возможнымъ и на этотъ новый вопросъ г. Буха отвѣтить отрицательно, отказомъ.

Приведем неоднородный фрагмент из 5 предложений этого же текста для четвёрок (признак «Глагол – Существительное – Союз – Существительное»):

Встрѣчаются напримѣръ на большой дорогѣ между Владиміромъ и Нижнимъ два обоза, на постояломъ дворѣ. Люди владимірскаго обоза распрашиваютъ, нѣтъ ли здѣсь въ обозѣ кого изъ новоторжскаго уѣзда тверской губерніи? Оказывается, что есть, и к тому же почти изъ той самой деревни, какая нужна. Выходитъ, что въ обозѣ дорогою изъ Москвы умеръ мужикъ, которому принадлежала тройка и товаръ. *Товарищи продали все это во Владимірѣ, выручили 600 рублей, и зная, что у покойника на селѣ остался отецъ и семья, ищутъ съ кѣмъ бы послать деньги. И встрѣчный мужикъ вѣшаетъ деньги къ себѣ на крестъ и черезъ полгода, съѣздивъ еще изъ Москвы въ Харьковъ и добравшись наконецъ до родного села, приноситъ семьѣ и деньги, и вѣсть о томъ, что Кирюха померъ, не доѣзжая какихъ–нибудь пятидесяти верстъ до Владиміра. «Такъ, сердечный, и не доѣхалъ; а до Владиміра всего одна какая–нибудь упряжка осталась, а много двѣ, такъ и померъ, не доѣхалъ». Разсказываются сотни подобныхъ примѣровъ. Напримѣръ еще на постояломъ дворѣ одинъ торговецъ, изъ крестьянъ, провожаетъ другого въ деревню, и проводы справляются обильнымъ чаепитіемъ.* «Поклонись ты батюшкѣ, да въ Москвѣ безпремѣнно купи женѣ платокъ въ два съ полтиной; да вотъ, какъ будешь въ Ярославлѣ, зайди къ Никанору Ѳедотову, знаешь? отдай ему вотъ тысячу двѣсти, чтобъ безпремѣнно по–прошлогоднему холстовъ мнѣ выслалъ, да чтобъ тѣхъ самыхъ рукъ холсты были; ты это ему накажи строго–настрого, а то онъ вѣдь мужикъ плутъ, пришлетъ пожалуй не тѣхъ». И отъѣзжающій суетъ за сапогъ деньги, завернутыя въ сальную бумагу, и деньги не пропадаютъ. И еще сотни подобныхъ примѣровъ разсказываются удивленными лицами верхняго лагеря, и въ тоже время тысячи есть примѣровъ крайней недобросовѣстности, совершеннаго отсутствія

самыхъ элементарныхъ понятій о чести въ поступкахъ лицъ нижняго лагеря относительно лицъ верхняго.

Отчего же это? Можетъ ли это понять, сумѣетъ ли догадаться г. Ростиславовъ? Причины, заставившія г. Мясоѣдова желать перечисляться въ крестьяне — чисто–психологическія, о которыхъ не можетъ быть рѣчи въ Сводѣ законовъ.

В ходе проведённых экспериментов для различных текстов было обнаружено, что наиболее информативными, с точки зрения максимальной дисперсии, являлись последовательности: для трёх частей речи «Местоимение – Наречие – Союз», «Глагол – Наречие – Союз» и «Предлог – Числительное – Существительное». Для последовательностей из четырёх частей речи наиболее информативными оказались: «Предлог – Прилагательное – Существительное – Глагол», «Существительное – Существительное – Существительное – Глагол» и «Предлог – Местоимение – Местоимение – Наречие».

Применение алгоритма к атрибуции текстов

Заметим, что алгоритм выявления неоднородных фрагментов можно использовать и для решения задачи атрибуции текстов. Рассмотрим следующую задачу. Имеется n текстов, которые будем считать однородными. В качестве таких текстов можно брать произведения, принадлежащие одному автору. Ставится задача определения степени близости неизвестного текста к этой группе. Решение данной задачи разобьём на несколько этапов. На первом этапе выбирается признак, затем по очереди выбирается один текст из группы однородных текстов. Оставшиеся произведения объединяются в один большой текст. Для каждого текста вычисляется статистика χ^2. Вычисление производится аналогично пункту 4 алгоритма поиска неоднородности для фрагмента. В качестве фрагмента будет выбранный текст. В качестве оставшегося текста будет выступать полученный объединенный. Будем обозначать значения статистик через χ_1^2, $\chi_2^2 \ldots, \chi_n^2$. На следующем этапе выбирается анализируемый текст, а в качестве второго берётся текст, полученный в результате объединения всех однородных. Для анализируемого текста вычисляется статистика χ^2. Обозначим её через χ_x^2. Если выполняется неравенство $\chi_x^2 \le \max_i \chi_i^2$, то искомый текст будет близок к данной группе по выбранному признаку. Близость текста к выбранной группе автоматически не означает решение задачи атрибуции. Для этого требуется критическая оценка полученного результата специалистом.

Для проверки работоспособности алгоритма в качестве однородных текстов использовались произведения, принадлежащие Ф.М. Достоевскому [4]. В качестве неоднородных текстов – произведения В.И. Даля, М.И. Владиславлева и А.А. Григорьева.

Для последовательностей из трёх частей речи (триад) «Местоимение – Наречие – Союз», «Существительное – Наречие – Модально-дискуссивное слово», «Глагол – Наречие – Существительное», «Предлог – Наречие – Существительное» статистика χ^2 для текстов Достоевского была меньше, чем для остальных текстов. Исключение составил текст «Загадки». Это может быть связано с тем, что размер данного текста был значительно меньше размеров остальных. Результаты эксперимента представлены в табл. 2. В колонке «Достоевский» выделены максимальные значения статистик χ^2 среди текстов Достоевского. А в колонке «Другие авторы» – значения, которые оказались меньше максимального значения из колонки «Достоевский». Для последовательностей из четырёх частей речи (четвёрки) такими последовательностями оказались: «Существительное – Существительное – Существительное – Глагол», «Предлог – Прилагательное – Существительное – Глагол» и «Местоимение – Модальное-дискуссивное слово – Предлог – Местоимение». Исключение составил всё тот же текст «Загадки». Результаты представлены в табл. 3.

Таблица 2 – Сравнение текстов Достоевского с другими авторами на основании триад

Признаки			Достоевский					Другие авторы							
I слово	II слово	III слово	Безцв. Явл.	Мелочи	Лит. Антик	Пожары	Дворянин	Г-жа Кох	Журн. Инт	Загадки	Заклад	Зап. Тайл	Князь Сереб	Панаев	Тарас Шев
Мест	Нар	Союз	0,00739	0,0184	0,0027	0,038205	**0,2239031**	1,186164	0,307	**0,039469**	0,2569	2,725	1,12441	7,63305	2,261
Сущ	Нар	МодДис	0,02323	0,09079	**0,1298**	0,04782	0,0048213	0,241474	0,1712	**0,054284**	0,35333	2,8027	0,17644	0,13127	1,552
Глаг	Нар	Сущ	0,02323	0,09079	0,1298	0,259088	**0,5054965**	2,172264	0,7112	**0,054284**	6,50973	0,6885	1,54642	5,28247	1,552
Предл	Числ	Сущ	0,29028	0,44865	**0,4771**	0,165546	0,3559647	3,878513	1,4378	**0,123528**	0,804	2,0652	1,73078	1,59332	1,743
Мест	Сущ	Союз	0,02297	0,04284	0,0413	**0,295**	0,0127209	0,339625	**0,1234**	0,387241	0,89729	0,6135	**0,00016**	0,93637	0,746
Прил	Сущ	Мест	0,33444	0,32076	**1,0847**	0,831074	0,1232635	2,720635	4,9062	**0,262536**	**0,28474**	5,9033	3,34656	8,44198	1,385
Сущ	Предл	Числ	0,01877	0,021	0,094	0,329957	**0,4498786**	5,009408	6,7869	**0,093831**	0,61072	4,3978	0,99288	2,51363	**0,156**
Союз	Нар	МодДис	0,01478	**0,17436**	0,0055	0,076464	0,0179109	3,926576	1,9893	**0,078994**	0,51416	0,927	0,6525	0,19102	**0,026**
Союз	Предл	Мест	**0,26497**	0,05764	0,0053	0,007128	0,2202288	**0,011242**	1,6944	**0,217778**	**0,23518**	2,2032	4,26073	0,52661	2,735
Сущ	Сущ	Прил	0,03025	0,13551	0,0496	**0,817632**	0,4401736	**0,043084**	3,9004	**0,12848**	0,83624	**0,0628**	3,96914	1,48276	1,21
Мест	МодДис	Предл	0,06988	0,06194	0,0271	**0,224908**	0,1062412	**0,080083**	0,3739	**0,044406**	0,28904	1,2522	1,26506	**0,10738**	0,345

Таблица 3 – Сравнение текстов Ф.М. Достоевского с другими авторами на основании четвёрок

Признаки				Достоевский					Другие авторы							
I слово	II слово	III слово	IV слово	Безцв. Явл	Мелочи	Лит. Антик	Пожары	Дворянин	Г-жа Кох	Журн. Инт	Загадки	Заклад	Зап. Тайл	Князь Сер	Панаев	Тарас Шев
Сущ	Сущ	Сущ	Глаг	0,016631	0,00087	0,003927	0,003218	0,0363439	0,0408	0,548416	0,033163	2,39794	0,0419	0,9961	0,07826	0,996791
Предл	Прил	Сущ	Глаг	0,033411	0,01611	0,001105	0,110861	0,0048404	0,4659	1,655269	0,104373	0,69718	0,21	1,3881	0,24632	0,204548
Мест	МодДис	Предл	Мест	0,03966	0,01292	0,042101	0,007508	0,0020795	2,2382	0,537308	0,028422	0,18986	0,4415	0,8538	0,06708	1,154495
Предл	Мест	Мест	Нар	0,016631	0,00087	0,003927	0,003218	0,0363439	0,8696	2,52656	0,033163	0,22152	2,3575	0,7888	0,07826	0,000012
Глаг	Мест	Предл	Мест	0,004344	0,02428	0,003944	0,033636	0,2053599	0,2313	0,21779	0,037904	1,96204	2,323	1,1385	0,08945	0,521901

При использовании последовательностей без учёта границ предложений для последовательностей «Предлог – Прилагательное – Существительное – Глагол» и

«Существительное – Прилагательное – Существительное – Модально-дискуссивное слово» максимальное значение статистики χ^2 для текстов Достоевского было меньше, чем для остальных текстов (табл. 4).

Таблица 4 – Сравнение текстов Ф.М. Достоевского с другими авторами на основании четвёрок без учёта границ предложений и абзацев

Признаки				Достоевский					Другие авторы							
I слово	II слово	III слово	IV слово	Безцв. Явл	Мелочи	Лит. Антик	Пожары	Дворянин	Г-жа Кох	Журн. Инт	Загадки	Заклад	Зап. Тайл	Князь Сер	Панаев	Тарас Шев
Предл	Прил	Сущ	Глаг	0,088154	0,0003	0,008382	0,080234	0,0267438	0,4971	1,486404	0,115191	0,71593	0,2593	1,2777	0,28532	0,255271
Сущ	Прил	Сущ	МодДи	0,00369	0,00113	0,009262	0,006705	0,0203678	0,4485	4,266243	22,21161	0,22753	2,413	0,9551	0,09067	0,855881
Предл	Мест	Мест	Нар	0,00369	0,00113	0,009262	0,006705	0,0203678	0,8897	2,440549	0,036608	0,22753	0,7236	0,8909	0,09067	0,000917
Мест	МодДис	Предл	Мест	0,066051	0,03105	0,056936	0,003812	2,69E-06	2,2612	0,488141	0,031376	0,19501	0,4726	0,8186	0,07771	1,238203
Предл	Сущ	Сущ	Мест	0,638806	0,38085	0,019899	0,104372	1,1319211	3,3923	3,996448	0,047075	0,29258	1,6274	1,2282	6,03388	1,245565

Заключение

Проведённые эксперименты показали, что с помощью описанного алгоритма можно выделить из текста фрагменты неоднородности, имеющие разные частоты встречаемости выбранной последовательности частей речи. Найденные фрагменты могут служить подсказкой для специалиста-филолога о том, что здесь может быть текст другого автора. Таким образом, данный алгоритм будет полезен при обнаружении плагиата: анализ может производиться не по всему тексту, а лишь по выделенным фрагментам неоднородности, что сократит размерность задачи. В дальнейшем планируется провести исследования по выявлению неоднородных фрагментов при условии выбора в качестве признаков линейной комбинации различных последовательностей.

Список литературы

1. Колесникова С.И. Методы анализа информативности разнотипных признаков // Вестник Томского государственного университета. Сер. «Управление, вычислительная техника и информатика». – 2009. – № 1 (6). – С. 69-80.

2. Котельников Е.В., Клековкина М.В. Автоматический анализ тональности текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). – Вып. 11 (18). – М. : Изд-во РГГУ, 2012.

3. Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. – Л. : Изд-во Ленингр. ун-та, 1990. – 164 с.

4. Полное собрание сочинений: канонические тексты / Ф.М. Достоевский; под редакцией проф. В.Н. Захарова. – Петрозаводск : Изд. ПетрГУ, 2004. – Т. 5.

5. Программа морфологического анализа текста на русском языке [Электронный ресурс] : [сайт]. – URL: http://company.yandex.ru/technologies/mystem (дата обращения: 14.11.2012).
6. Справочник по прикладной статистике. В 2-х т. / пер. с англ. ; под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. – М. : Финансы и статистика, 1989.
7. Bouville M. Plagiarism: Words and ideas (англ.) // Science and Engineering Ethics. – 2008.
8. Language independent approach to sentimental analysis [Электронный ресурс]. Компьютерная лингвистика и интеллектуальные технологии // Материалы международной конференции по компьютерной лингвистике. 2012. – URL: http://dialog-21.ru/digests/dialog2012/materials/pdf/70.pdf (дата обращения: 14.11.2012).

Рецензенты:

Питухин Евгений Александрович, доктор технических наук, профессор кафедры прикладной математики и кибернетики, ФГБОУ ВПО «Петрозаводский государственный университет», г. Петрозаводск.

Печников Андрей Анатольевич, доктор технических наук, ведущий научный сотрудник лаборатории телекоммуникационных систем, ФГБУН «Институт прикладных математических исследований» Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН), г. Петрозаводск.