

## ПРИМЕНЕНИЕ МЕТОДА РАСПОЗНАВАНИЯ ОБРАЗОВ ДЛЯ ОЦЕНКИ КАЧЕСТВА УЧЕБНЫХ ТЕКСТОВ

УДК 655.5

М.А. Зильберглейт, А.С. Малюкевич,  
БГТУ, г. Минск

### Аннотация

В статье описано применение методов распознавания образов для анализа использования учебных текстов при подготовке специалистов по редакционно-издательскому профилю. В качестве основных методов распознавания были использованы: метод дискриминантного анализа, метод дерева решений.

### Введение

Содержание современных учебных пособий для высшего образования должно не только соответствовать всем установленным для данного типа и вида литературы требованиям, но и быть, прежде всего, понятным тем, кому оно адресовано.

Среди основных требований, предъявляемых к учебному изданию для студентов высших учебных заведений, можно указать информативность, терминологическую полноту, логичность изложения, лаконичность подачи текстового и изобразительного материала, а также удобочитаемость.

На сегодняшний день можно констатировать, что, к сожалению, не все издания подходят той группе читателей, на которую они рассчитаны. Это связано с тем, что уровень подготовки авторов намного выше уровня подготовки обучающихся, а первая оценка издания происходит, как правило, высококвалифицированным рецензентом и редактором. Конечно, использование каждый раз читателей для оценки качества текста требует значительных материальных и временных затрат. В этой связи мы предлагаем подход, который использует классические приемы методов распознавания образов для решения задачи по оценке качества текста. Основой данного метода является формирование так называемого «решающего правила» [1]. Оно формулируется на основе анализа двух или более выборок с известным качеством. С помощью сформулированного решающего правила можно дать оценку вновь изучаемому объекту.

Цель работы – оценить качество учебных изданий методом распознавания образов, получить решающее правило в виде дискриминантных функций для классификации объектов по заранее установленным признакам.

Основные задачи: провести дискриминантный анализ статистических данных, полученных в ходе обработки 79 текстовых фрагментов; оценить влияние каждого из признаков, на основе которых происходит классификация объектов исследования. Предварительная оценка материала учебного издания для высшей школы еще на стадии доредакционной подготовки позволит не только правильно организовать работу редакционно-издательского и авторского состава, но и повысит уровень самого издания, сделает его читабельным, интересным и отвечающим основным требованиям, предъявляемым к той или иной группе изданий.

В качестве основных этапов работы были определены: отбор текстовых фрагментов учебных изданий, анализ единиц выборки на установление основных статистических

показателей, опрос студентов, формулировка решающего правила методом линейных дискриминантных функций и методом дерева решений, проверка результатов исследования.

### Подготовительный этап

На данном этапе работы было установлено, что статистические параметры текстового фрагмента выполняют атрибутивную функцию, то есть помогают установить авторскую принадлежность того или иного текстового фрагмента. На основе использования метода дисперсионного анализа нами было также показано, что объем текстовой выборки порядка 1800–2000 символов является достаточным для использования статистических характеристик текста в исследованиях по читабельности [2].

Данный этап заключался в установлении основных статистических показателей текстовых фрагментов. В качестве объектов исследования были определены 79 тематических текстовых блоков учебных изданий по издательскому делу в объеме 1800–2000 символов каждый.

Для реализации данного этапа работы была использована программа SuperCounter 2.1. Основными статистическими показателями являются: средняя длина слов в слогах, средняя длина слов в буквах, средняя длина слов по Деверу, процент слов в 3–7 слогах, процент односложных слов, средняя длина предложения в словах, средняя длина предложения в слогах, процент неповторяющихся в тексте слов, процент чисел от общего количества слов, процент иностранных слов в тексте, процент опорных слов, найденных с использованием программного средства [3], а также процент опорных слов, выявленных самостоятельно.

В ходе анализа, для получения более репрезентативных результатов нами дополнительно были введены два параметра: отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более», а также отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более». По результатам проведенных в программе SuperCounter 2.1 расчетов была составлена итоговая таблица значений статистических параметров текста, фрагмент которой приведен ниже (таблица 1).

### Основная часть

Оценка текстовых фрагментов проводилась методом опроса. В качестве респондентов выступили 135 студентов, обучающихся на 3–4 курсе специальности «Издательское дело». Для реализации данного этапа работы была разработана 9-балльная шкала оценки.

Средний балл по всем текстам составил 4,89, что соответствует «базовому» уровню восприятия текстовой информации читателями.

Для дальнейшей работы с данными дополнительно было проведено разделение полученных в ходе опроса оценок на две группы объектов. Первая группа – объекты, имеющие легкий

Таблица 1 – Показатели статистических параметров текстовых фрагментов

№	Ср. дл. слов в слогах	Ср. дл. слов в буквах	Ср. дл. слов по Деверу	% слов в 3 сл. и более	% слов в 4 сл. и более	% слов в 5 сл. и более	% слов в 6 сл. и более	% слов в 7 сл. и более	% односложных слов	Ср. дл. предл. в словах	Ср. дл. предл. в слогах	% чисел от общ. кол. слов	% опорных слов (самостоятельно)	% опорных слов (программное средство)	3/6	4/6
1	2,38	5,95	7,38	43,10	23,20	11,10	3,03	0,34	29,30	21,20	50,50	1,01	79,18	51,54	14,22	7,66
2	2,77	6,48	7,77	56,70	33,80	13,50	2,55	0,73	24,40	13,10	36,30	0,00	71,94	58,63	22,24	13,25
3	2,65	6,17	7,55	48,80	29,90	16,30	3,65	0,66	26,90	13,10	34,70	0,00	58,80	43,19	13,37	8,19
4	2,84	6,53	7,65	54,70	37,60	16,40	6,57	0,37	25,20	24,90	70,80	0,37	67,18	46,95	8,33	5,72
5	2,70	6,38	7,53	51,90	25,40	14,40	5,15	2,06	25,10	22,40	60,50	0,00	66,79	51,29	10,08	4,93
6	2,65	6,41	7,64	52,30	26,90	13,80	3,18	0,00	23,30	16,60	44,10	0,00	68,63	57,56	16,45	8,46
7	2,88	6,71	7,96	51,90	32,30	19,60	8,42	2,46	21,80	14,30	41,10	0,35	68,13	52,01	6,16	3,84
8	2,76	6,63	7,88	53,70	33,00	15,10	6,32	2,46	26,70	19,00	52,50	0,00	70,85	54,24	8,50	5,22
9	2,88	6,46	7,68	52,30	33,20	21,20	13,10	7,07	29,00	23,60	67,80	0,35	60,77	47,31	3,99	2,53
10	3,24	7,58	8,75	64,70	47,70	27,00	7,88	2,49	20,70	21,90	70,90	0,00	78,35	59,74	8,21	6,05
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
70	2,91	6,90	8,10	60,10	35,10	15,30	7,84	3,36	22,00	17,90	52,00	1,87	61,19	48,13	7,67	4,48
71	2,64	6,39	7,60	52,30	33,60	15,20	4,24	1,77	29,70	18,90	49,70	1,77	69,29	51,43	12,33	7,92
72	2,93	6,87	8,12	57,90	39,10	23,00	8,81	1,53	26,80	32,60	95,50	1,15	67,69	55,00	6,57	4,44
73	2,63	6,20	7,43	45,60	30,90	15,10	6,38	1,68	26,20	21,30	55,90	1,68	65,77	52,35	7,15	4,84
74	2,71	6,48	7,78	52,80	31,50	15,70	4,12	0,75	26,60	15,70	42,60	0,75	60,30	48,69	12,82	7,65
75	2,63	6,85	8,05	49,10	30,30	18,00	4,87	1,50	22,10	24,30	63,90	0,00	74,34	66,79	10,08	6,22
76	2,86	6,77	8,01	57,90	33,80	20,30	10,50	2,26	24,40	26,60	76,20	2,63	60,38	49,81	5,51	3,22
77	2,97	6,88	8,19	57,80	36,10	20,20	7,22	1,52	22,10	18,80	55,80	0,00	63,88	48,67	8,01	5,00
78	2,40	5,83	6,99	44,20	24,90	9,63	1,00	0,00	29,20	20,10	48,20	1,33	58,14	44,85	44,33	24,97
79	2,69	6,54	7,84	50,60	28,00	12,90	4,80	1,11	21,40	27,10	72,90	0,00	62,36	54,24	10,54	5,83

уровень восприятия информации. Вторая группа – объекты, имеющие базовый уровень восприятия информации и выше.

Таким образом, по результатам проведенного разделения, 63 из 79 объектов были отнесены к первой группе объектов, 16 – ко второй.

По результатам выполненных этапов работы была составлена таблица (таблица 2), данные которой послужили основой для проведения основного этапа исследования – дискриминантного анализа.

Значения итоговой таблицы были обработаны в программе StatGraphics 5.1 методом дискриминантного анализа.

Дискриминантная функция для первой группы объектов выглядит следующим образом:

$$\begin{aligned}
 & -4217,54 + 131,775 * \text{длина\_предложения\_в\_словах} - \\
 & 46,67 * \text{длина\_предложения\_в\_слогах} \\
 & + 104,972 * \text{длина\_слов\_в\_буквах} + 196,488 * \text{длина\_слов\_по\_Деверу} + 2454,24 * \text{длина\_слов\_в\_слогах} - \\
 & 1,08959 * \text{процент\_опорных\_слов\_самостоятельно} + \\
 & 4,72454 * \text{процент\_опорных\_слов\_программное\_средство} + \\
 & +45,910 * \text{процент\_односложных\_слов} - 4,72738 * \text{процент\_слов\_в\_3\_слога\_и\_более} - 43,5519 * \text{процент\_слов\_в\_4\_слога\_и\_более} - 18,401 * \text{процент\_слов\_в\_5\_слогов\_и\_бо-}
 \end{aligned}$$

$$\begin{aligned}
 & \text{лее} + 2,43421 * \text{процент\_слов\_в\_6\_слогов\_и\_более} - \\
 & 61,2428 * \text{процент\_слов\_в\_7\_слогов\_и\_более} + 50,5953 * \text{процент\_чисел\_от\_общего\_количества\_слов} - 79,0391 * (3/6) + \\
 & 141,653 * (4/6).
 \end{aligned}$$

Дискриминантная функция для второй группы объектов:

$$\begin{aligned}
 & -4190,26 + 132,716 * \text{длина\_предложения\_в\_словах} - \\
 & 47,0458 * \text{длина\_предложения\_в\_слогах} \\
 & + 81,2358 * \text{длина\_слов\_в\_буквах} + 226,745 * \text{длина\_слов\_по\_Деверу} + 2420,8 * \text{длина\_слов\_в\_слогах} - 1,36107 * \text{процент\_опорных\_слов\_самостоятельно} + 4,39123 * \text{процент\_опорных\_слов\_программное\_средство} + 44,7041 * \text{процент\_односложных\_слов} - 4,10418 * \text{процент\_слов\_в\_3\_слога\_и\_более} - 42,9318 * \text{процент\_слов\_в\_4\_слога\_и\_более} - 19,3299 * \text{процент\_слов\_в\_5\_слогов\_и\_более} + 3,5031 * \text{процент\_слов\_в\_6\_слогов\_и\_более} - 60,7879 * \text{процент\_слов\_в\_7\_слогов\_и\_более} + 49,6441 * \text{процент\_чисел\_от\_общего\_количества\_слов} - 79,0679 * (3/6) + 142,504 * (4/6).
 \end{aligned}$$

Основные показатели проведенного дискриминантного анализа:

Eigenvalue (собственное значение) – 1,74515, Relative Percentage (процентное содержание) – 100, Canonical Correlation (каноническая корреляция) – 0, Wilks Lambda

Таблица 2 – Фрагмент таблицы данных для дискриминантного анализа

№	Ср. дл. слов в слогах	Ср. дл. слов в буквах	Ср. дл. слов по Деверу	% слов в 3 сл. и более	% слов в 4 сл. и более	% слов в 5 сл. и более	% слов в 6 сл. и более	% слов в 7 сл. и более	% односложных слов	Ср. дл. предл. в словах	Ср. дл. предл. в слогах	% чисел от общ. кол. слов	% опорных слов (самостоятельно)	% опорных слов (программное средство)	3/6	4/6	фактор
1	2,38	5,95	7,38	43,10	23,20	11,10	3,03	0,34	29,30	21,20	50,50	1,01	79,18	51,54	14,22	7,66	1
2	2,77	6,48	7,77	56,70	33,80	13,50	2,55	0,73	24,40	13,10	36,30	0,00	71,94	58,63	22,24	13,25	2
3	2,65	6,17	7,55	48,80	29,90	16,30	3,65	0,66	26,90	13,10	34,70	0,00	58,80	43,19	13,37	8,19	2
4	2,84	6,53	7,65	54,70	37,60	16,40	6,57	0,37	25,20	24,90	70,80	0,37	67,18	46,95	8,33	5,72	1
5	2,70	6,38	7,53	51,90	25,40	14,40	5,15	2,06	25,10	22,40	60,50	0,00	66,79	51,29	10,08	4,93	1
6	2,65	6,41	7,64	52,30	26,90	13,80	3,18	0,00	23,30	16,60	44,10	0,00	68,63	57,56	16,45	8,46	1
7	2,88	6,71	7,96	51,90	32,30	19,60	8,42	2,46	21,80	14,30	41,10	0,35	68,13	52,01	6,16	3,84	1
8	2,76	6,63	7,88	53,70	33,00	15,10	6,32	2,46	26,70	19,00	52,50	0,00	70,85	54,24	8,50	5,22	1
9	2,88	6,46	7,68	52,30	33,20	21,20	13,10	7,07	29,00	23,60	67,80	0,35	60,77	47,31	3,99	2,53	1
10	3,24	7,58	8,75	64,70	47,70	27,00	7,88	2,49	20,70	21,90	70,90	0,00	78,35	59,74	8,21	6,05	1

(коэффициент Шапиро-Уилкса) – 0,364278, Chi-Square (хи-квадрат) – 69,6787, DF (количество степеней свободы) – 16, P-Value (P-значение) – 0. Результаты классификации объектов по группам можно представить в виде таблицы (таблица 3).

Таблица 3 – Результаты дискриминантного анализа

Фактор	Размер группы	1	2
1	63	62 (98,41 %)	1 (1,59 %)
2	16	1 (6,25 %)	15 (93,75 %)
<b>Всего</b>	79	97,47 %	

Из полученных данных следует, что количество правильно классифицированных объектов составило – 77 из 79, в том числе по первой группе 62 объекта из 63 (98,41 %), по второй – 15 из 16 (93,75 %). Таким образом, было установлено, что процент правильно классифицированных данных составляет 97,47 %, то есть в данном случае можно утверждать, что сформулированная дискриминантная функция может быть использована для анализа текстовых фрагментов учебных изданий объемом 1800–2000 символов с целью определения оценки качества учебных текстов.

Полученные в ходе анализа результаты были проверены также с помощью программного пакета AnswerTree, при этом для анализа был использован метод CR&T. Результат работы программы представлен на рисунке 1 и таблице 4.

Таблица 4 – Результаты распознавания для метода дерева решений

	1	2	Сумма
1	58 (92,06 %)	1	59
2	5	15 (93,75 %)	20
<b>Всего</b>	63	16	79

Основные показатели дерева решений: Risk Estimate (оценка риска) – 0,0709325, SE of Risk Estimate (средне-квадратическая ошибка) – 0,0347199. Nodes (количество узлов) – 13, Number of Levels (количество уровней) – 4, terminal nodes (терминальные узлы) – 7. Очевидно, что метод дерева решений позволяет определить с высоким показателем точности группу объекта, к которой будет отнесен тот или иной признак.

На следующем этапе работы была поставлена задача проверки полученных решающих правил на конкретном примере. Для его реализации были отобраны 15 дополнительных текстовых фрагментов. С каждым из них были выполнены все этапы работы, указанные в первой части исследования. Полученные данные были проверены при помощи дискриминантных функций и дерева решений на истинность значений. Показатели основных статистических показателей представлены в таблице 5.

После получения значений основных статистических показателей, были рассчитаны соответствующие дискриминантные функции, значения которых представлены в таблице 6.

Согласно результатам дискриминантного анализа, издания под номерами 4, 8 и 15 имеют уровень восприятия информации «базовый и выше», издания под номерами 1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 13, 14 ниже базового. Для подтверждения данного вывода был проведен дополнительно опрос среди студентов 3–4 курсов специальности «Издательское дело». Результаты сгруппированы в таблице 7.

### Заключение

Очевидно, что результат, полученный по методу дерева решений, практически полностью совпадает с результатами опроса, а по методу дискриминантного анализа процент некорректно установленных объектов составляет 26,6 %.



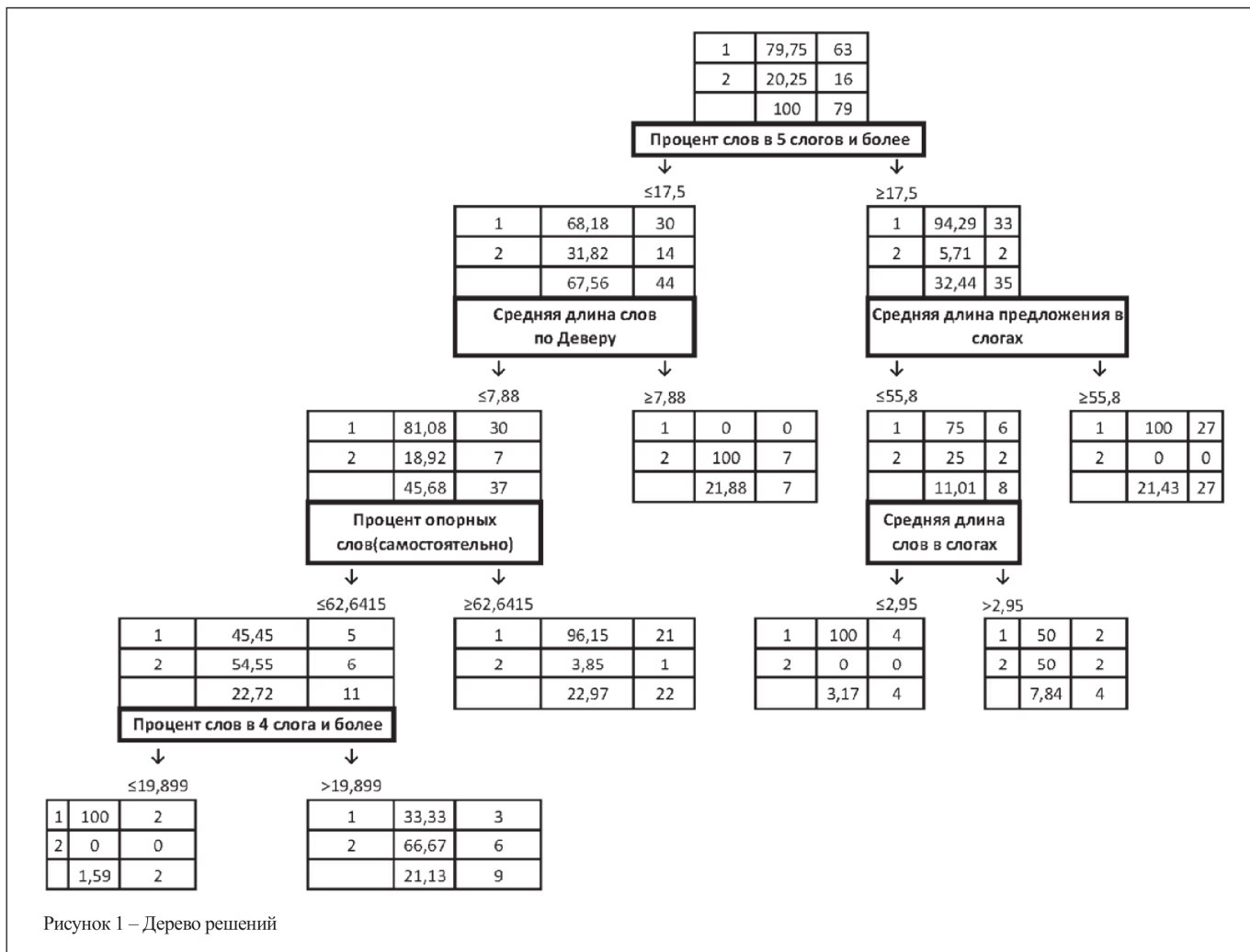


Таблица 5 – Статистические показатели анализируемых текстов

№	Средняя длина слов по Деверу	% слов в 3 слога и более	% слов в 4 слога и более	% слов в 5 слога и более	% слов в 6 слога и более	% слов в 7 слога и более	% односложных слов	Средняя длина предложения в слогах	Средняя длина предложения в слогах	% чисел от общего количества слов	% опорных слов (самостоятельно)	% опорных слов (программное средство)	3/6	4/6
1	8,58	60,3	35,6	25,8	10,50	3,37	21,7	19,1	58,5	0,38	71,16	58,80	5,74	3,39
2	8,88	57,5	38,2	28,0	16,10	9,45	20,9	31,8	104,0	0	70,87	55,12	3,57	2,37
3	7,90	51,9	33,2	17,6	7,61	1,73	20,1	22,2	64,2	0,35	64,58	48,61	6,82	4,36
4	8,27	57,2	39,1	18,8	9,06	2,90	25,0	19,7	58,4	0	64,60	55,11	6,31	4,32
5	7,28	48,1	26,6	13,8	6,25	2,19	27,5	20,0	52,4	0	55,94	40,00	7,70	4,26
6	7,93	56,6	36,1	19,1	7,64	1,04	21,5	16,0	46,2	0,70	64,34	54,20	7,41	4,73
7	8,11	59,2	35,5	19,1	7,09	1,77	19,9	16,6	49,2	0,71	68,68	60,50	8,35	5,01
8	7,04	45,2	23,7	9,6	2,69	0,90	25,4	13,9	31,6	8,98	77,75	44,21	16,80	8,81
9	7,88	55,9	29,7	18,3	5,17	0,69	26,9	18,1	48,4	3,10	71,28	56,75	10,81	5,74
10	8,03	56,1	38,6	20	7,02	2,11	23,2	23,8	70,0	0,35	66,90	55,28	7,99	5,50
11	7,23	43,6	23,4	12,8	4,05	0,94	25,9	16,1	39,6	2,80	63,09	47,95	10,77	5,78
12	8,40	55,9	37,4	23,3	10,40	5,93	25,6	18,0	56,5	0	60,60	58,36	5,38	3,60
13	8,20	59,4	35,9	20,3	8,33	1,81	21,7	25,1	71,8	3,26	78,07	54,65	7,13	4,31
14	7,74	49,7	33,4	20,6	12,8	3,04	28,4	17,4	48,3	2,70	59,12	43,92	3,88	2,61
15	7,57	46,2	25,1	13,9	3,30	1,32	22,8	12,6	30,6	0,33	67,99	47,86	14,00	7,61

Таблица 6 – Значения дискриминантных функций анализируемых текстов

№	Значение $f_1$		Значение $f_2$	Принадлежность к группе объектов
1	4307,67	>	4300,09	1
2	3919,03	>	3909,63	1
3	4135,96	>	4134,07	1
4	4250,39	<	4250,69	2
5	4094,52	>	4093,02	1
6	4144,49	>	4143,97	1
7	4192,12	>	4190,41	1
8	3879,80	<	3880,75	2
9	4326,28	>	4316,47	1
10	4160,68	>	4158,61	1
11	4119,39	>	4112,11	1
12	4435,86	>	4424,38	1
13	4166,50	>	4164,03	1
14	4242,02	>	4237,72	1
15	3685,01	<	3687,71	2

Таблица 7 – Результаты опроса респондентов

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Средний балл	5,32	5,58	5,47	5,47	5,51	6,11	4,64	5,19	5,00	4,89	5,32	5,66	5,34	5,32	4,81

Таким образом, использование статистических параметров текста позволяет установить группу по удобочитаемости, к которой будет отнесен объект исследования.

**Литература:**

1. Айвазян, С.А. Классификация многомерных наблюдений [Текст] / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – М. : Статистика, 1987. – С. 134.
2. Зильберглейт, М.А. Статистический анализ текстов учебных изданий по издательскому делу / М.А. Зильберглейт, А.С. Малюкевич // Электроника инфо: науч.-практ. журнал для специалистов. – 2013. – № 1. – С. 25–29.
3. Анализатор текстов [Электронный ресурс]. – Апрель, 2012. – Режим доступа : <http://shipbottle.ru/ir.ru>.

**Abstract**

The article describes the use of pattern recognition methods to analyze the use of texts in training of editorial and publishing profile. The main methods of detection were used: the method of discriminant analysis, the method of decision tree.

*Поступила в редакцию 16.05.2013 г.*

**ТРЕБОВАНИЯ К НАУЧНЫМ СТАТЬЯМ, ПУБЛИКУЕМЫМ В РАЗДЕЛЕ «РЕЦЕНЗИРУЕМЫЕ СТАТЬИ»**

1. Научная статья – законченное и логически цельное произведение по раскрываемой теме – должна соответствовать одному из следующих научных направлений: информационные технологии и системы, оптоэлектроника, микро- и наноэлектроника, приборостроение.

2. Объем научной статьи не должен превышать 0,35 авторского листа (14 тысяч печатных знаков, включая пробелы между словами, знаки препинания, цифры и другие), что соответствует 8 страницам текста, напечатанного через 2 интервала между строками (5,5 страницы в случае печати через 1,5 интервала).

3. Статьи в редакцию представляются в двух экземплярах на бумаге формата А4 (220015, г. Минск, пр. Пушкина, 29Б), а также в электронном виде (e-mail: [sadov@bsu.by](mailto:sadov@bsu.by)). К статье прилагаются сопроводительное письмо организации за подписью руководителя и акт экспертизы. Статья должна быть подписана всеми авторами.

Статьи принимаются в формате doc, rtf, pdf, набранные в текстовом редакторе word, включая символы латинского и греческого алфавитов вместе с индексами. Каждая иллюстрация (фотографии, рисунки, графики, таблицы и др.) должна быть представлена отдельным файлом и названа таким образом, чтобы была понятна последовательность ее размещения. Фотографии принимаются в форматах tif или jpg (300 dpi). Рисунки, графики, диаграммы принимаются в форматах tif, cdr, eps или jpg (300 dpi, текст в кривых). Таблицы принимаются в форматах doc, rtf или Excel.

4. Научные статьи должны включать следующие элементы: аннотацию; фамилию и инициалы автора (авторов) статьи, ее название; введение; основную часть, включающую графики и другой иллюстративный материал (при их наличии); заключение; список цитированных источников; индекс УДК; аннотацию на английском языке.

5. Название статьи должно отражать основную идею выполненного исследования, быть по возможности кратким, содержать ключевые слова, позволяющие индексировать данную статью.

6. Аннотация (100–150 слов) должна ясно излагать содержание статьи и быть пригодной для опубликования в аннотациях к журналам отдельно от статьи.

В разделе «Введение» должен быть дан краткий обзор литературы по данной проблеме, указаны не решенные ранее вопросы, сформулирована и обоснована цель работы.

Основная часть статьи должна содержать описание методики, аппаратуры, объектов исследования и подробно освещать содержание исследований. Полученные результаты должны быть обсуждены с точки зрения их научной новизны и сопоставлены с соответствующими известными данными. Основная часть статьи может делиться на подразделы (с разъяснительными заголовками).

Иллюстрации, формулы, уравнения и сноски, встречающиеся в статье, должны быть пронумерованы в соответствии с порядком цитирования в тексте.

В разделе «Заключение» должны быть в сжатом виде сформулированы основные полученные результаты с указанием их новизны, преимуществ и возможностей применения.

Список цитированных источников располагается в конце текста, ссылки нумеруются согласно порядку цитирования в тексте. Порядковые номера ссылок должны быть написаны внутри квадратных скобок (например: [1], [2]).

В соответствии с рекомендациями ВАК Республики Беларусь от 29.12.2007г. №29/13/15 научные статьи аспирантов последнего года обучения публикуются вне очереди при условии их полного соответствия требованиям, предъявляемым к рецензируемым научным публикациям.