Natural Language Processing

Methods in Authorship Attribution

Authorship attribution is the process of determining the writer of a document.

Attribution is the assignment text to a genre, style, period of time, area, social class, author’s gender etc.

Application:

- author of spam, threat detection;
- text attribution in criminalistics;
- plagiarism detection etc.

Writer invariant, also called **authorial invariant** or **author's invariant**, is a property of a text which is invariant of its author, that is, it will be similar in all texts of a given author and different in texts of different authors.

Authorship Attribution

Problems:

- **Profiling problem**: There is no candidate set at all. In this case, the challenge is to provide as much demographic or psychological information as possible about the author.
- **Verification problem**: There is no closed candidate set but there is one suspect. In this case, the challenge is to determine if the suspect is or is not the author.

Authorship Attribution

Methods:

- **the unitary invariant approach**: a single numeric function of a text is sought to discriminate between authors;
- **the multivariate analysis approach**: statistical multivariate discriminant analysis is applied to word frequencies and related numerical features;
- **the machine learning approach**: modern machine learning methods are applied to sets of training documents to construct classifiers that can be applied to new anonymous documents

Unitary Invariant Approach

The early twentieth century: the search for invariant properties of textual statistics. The existence of such invariants suggested the possibility that some related feature might be found that was at least invariant for any given author, though possibly varying among different authors.

Multivariate Analysis Approach

A basic intuition is that finding the most probable attribution can be viewed as taking documents as points in some space, and assigning a questioned document to the author whose documents are 'closest' to it, according to an appropriate distance measure.

Результаты:

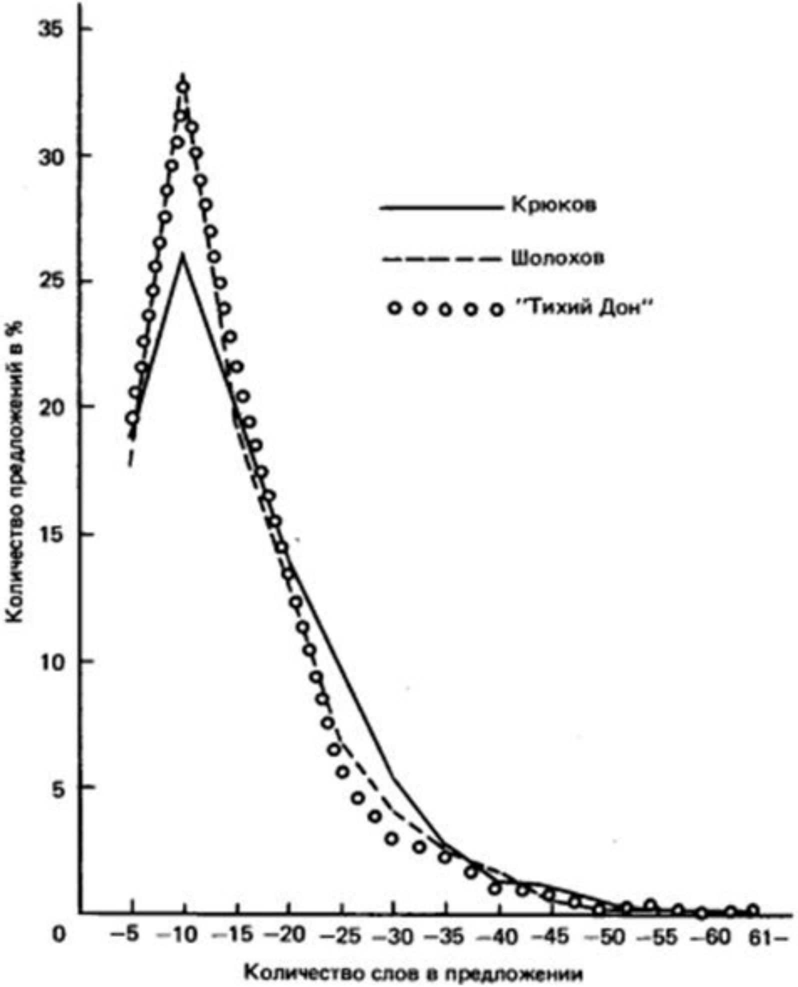
- средняя длина предложений:

Крюков – 13,9 слова

Шолохов – 12,9 слова

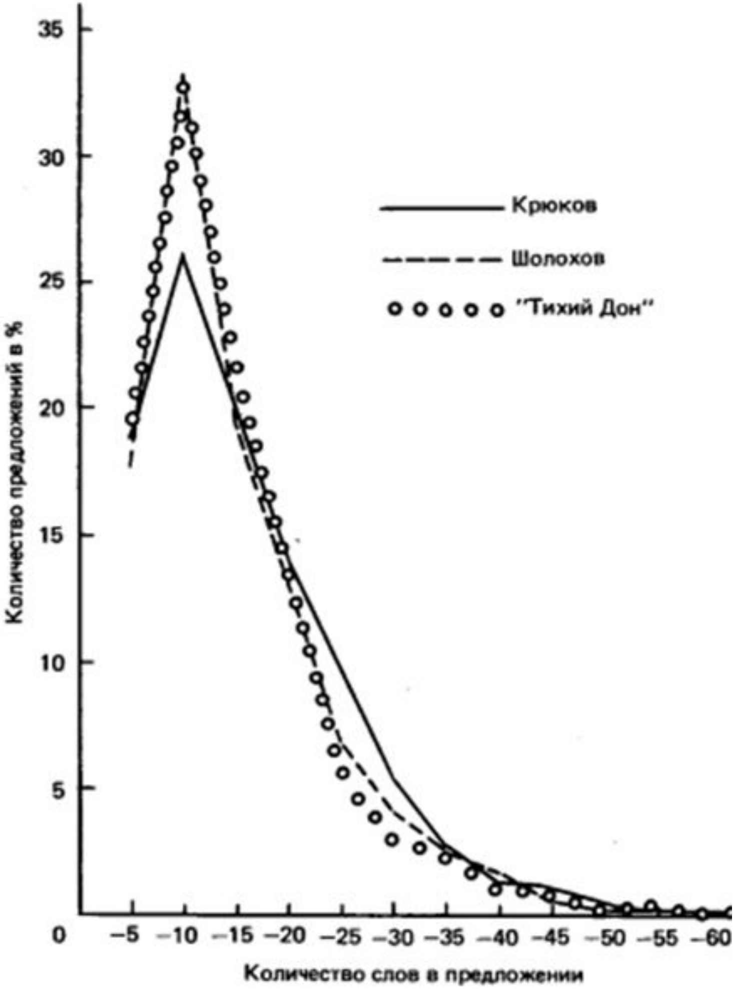
«Тихий Дон» – 12,4 слова

- распределение длины предложений по количеству слов



Average sentence length:

- Krukov - 13,9 words;
- Sholohov - 12,9 words;
- ‘Quietly Flows the Don’ - 12,4 words



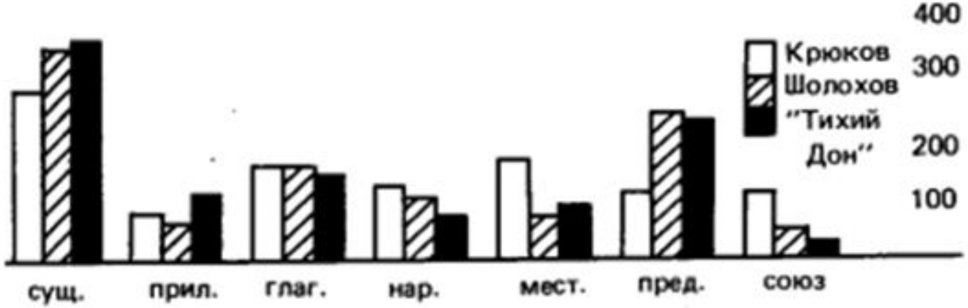
Дистрибуция частей речи. Средние значения

Части речи	Крюков	Шолохов	«Тихий Дон»
существительное	147,85	183,60	184,50
прилагательное	64,20	63,75	72,25
глагол	86,75	91,70	91,05
наречие	51,75	38,60	34, 30
местоимение	47,50	24,00	25,85
предлог	60,80	71,70	69,30
союз	41,15	26,65	22,75
Итого	500,00	500,00	500,00

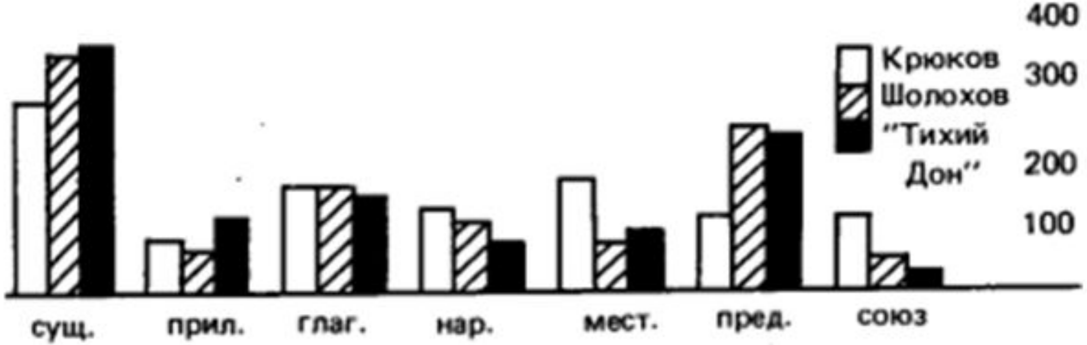
Distribution of Parts of Speech

Part of Speech	Krukov	Sholohov	‘Quietly Flows the Don’
Noun	147,8	183,60	184,50
Adjective	64,20	63,75	72,25
Verb	86,75	91,70	91,05
Adverb	51,75	38,60	34,30
Pronoun	47,50	24,00	25,85
Preposition	60,80	71,70	69,30
Conjunction	41,15	26,65	22,75

- 1-я позиция:



- 2-я позиция:



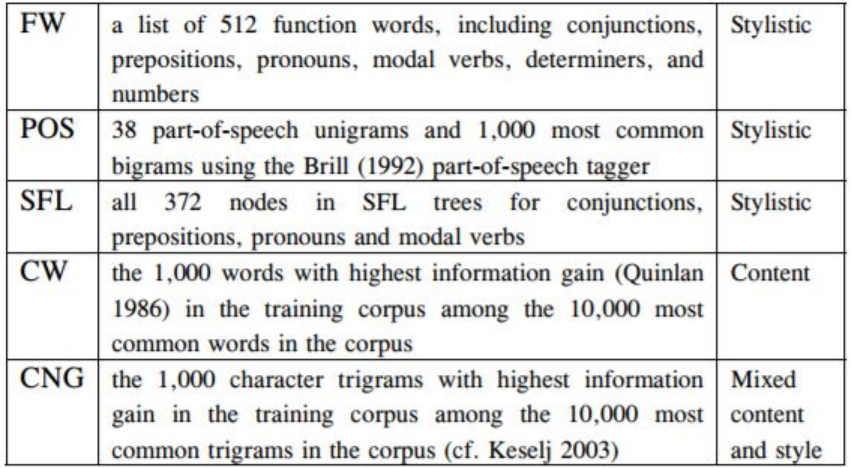
Machine Learning Approach

Training texts are represented as labeled numerical vectors and learning methods are used to find boundaries between classes (authors) that minimize some classification loss function.

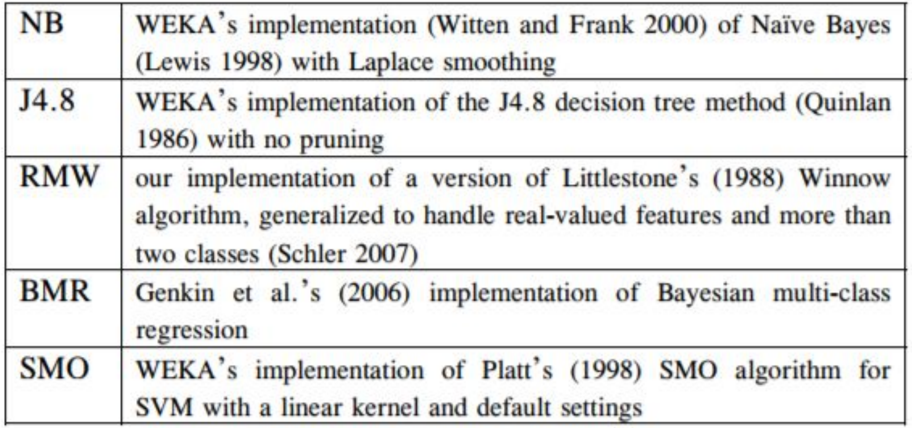
Features for Authorship Attribution

- Complexity measures: average word length, average number of words in sentence, type-token ratio, the number of words appearing with given frequency in a text etc.;
- Function words;
- Syntax and parts-of-speech;
- Functional lexical taxonomies;
- Content words;
- Character n-grams;
- Other specialized features: frequency of distinctive punctuation habits or orthographic/syntactic errors, formatting and other structural features

Comparison of ML methods

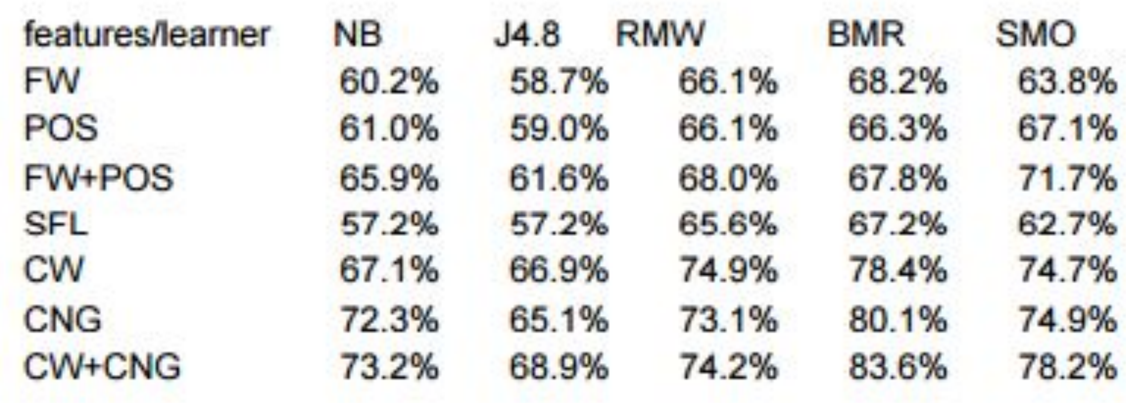


FW	a list of 512 function words, including conjunctions, prepositions, pronouns, modal verbs, determiners, and numbers	Stylistic
POS	38 part-of-speech unigrams and 1,000 most common bigrams using the Brill (1992) part-of-speech tagger	Stylistic
SFL	all 372 nodes in SFL trees for conjunctions, prepositions, pronouns and modal verbs	Stylistic
CW	the 1,000 words with highest information gain (Quinlan 1986) in the training corpus among the 10,000 most common words in the corpus	Content
CNG	the 1,000 character trigrams with highest information gain in the training corpus among the 10,000 most common trigrams in the corpus (cf. Keselj 2003)	Mixed content and style



NB	WEKA's implementation (Witten and Frank 2000) of Naïve Bayes (Lewis 1998) with Laplace smoothing
J4.8	WEKA's implementation of the J4.8 decision tree method (Quinlan 1986) with no pruning
RMW	our implementation of a version of Littlestone's (1988) Winnow algorithm, generalized to handle real-valued features and more than two classes (Schler 2007)
BMR	Genkin et al.'s (2006) implementation of Bayesian multi-class regression
SMO	WEKA's implementation of Platt's (1998) SMO algorithm for SVM with a linear kernel and default settings

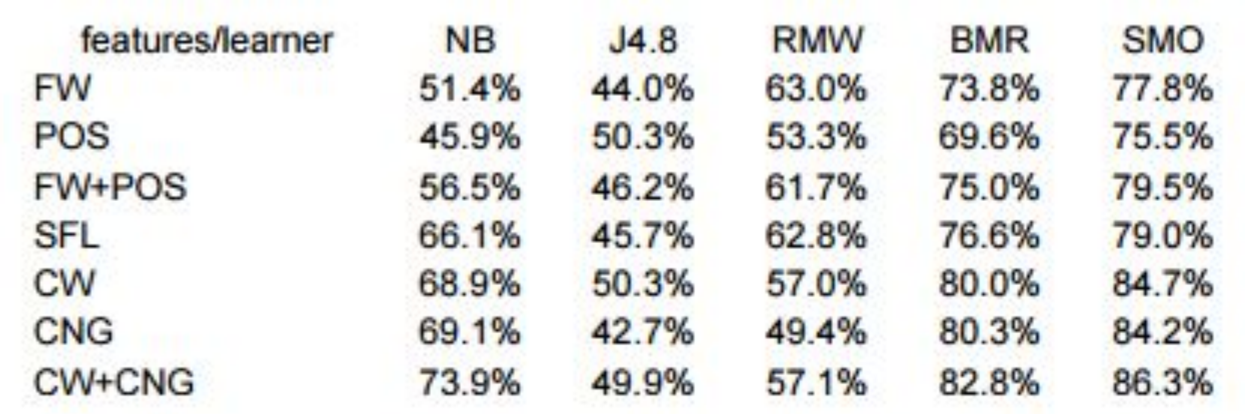
Comparison of ML methods



features/learner	NB	J4.8	RMW	BMR	SMO
FW	60.2%	58.7%	66.1%	68.2%	63.8%
POS	61.0%	59.0%	66.1%	66.3%	67.1%
FW+POS	65.9%	61.6%	68.0%	67.8%	71.7%
SFL	57.2%	57.2%	65.6%	67.2%	62.7%
CW	67.1%	66.9%	74.9%	78.4%	74.7%
CNG	72.3%	65.1%	73.1%	80.1%	74.9%
CW+CNG	73.2%	68.9%	74.2%	83.6%	78.2%

Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the email corpus

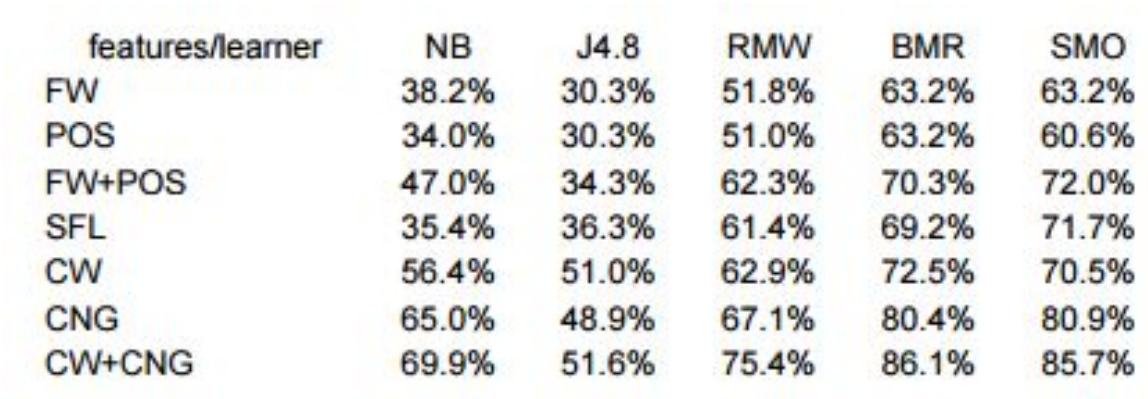
Comparison of ML methods



features/learner	NB	J4.8	RMW	BMR	SMO
FW	51.4%	44.0%	63.0%	73.8%	77.8%
POS	45.9%	50.3%	53.3%	69.6%	75.5%
FW+POS	56.5%	46.2%	61.7%	75.0%	79.5%
SFL	66.1%	45.7%	62.8%	76.6%	79.0%
CW	68.9%	50.3%	57.0%	80.0%	84.7%
CNG	69.1%	42.7%	49.4%	80.3%	84.2%
CW+CNG	73.9%	49.9%	57.1%	82.8%	86.3%

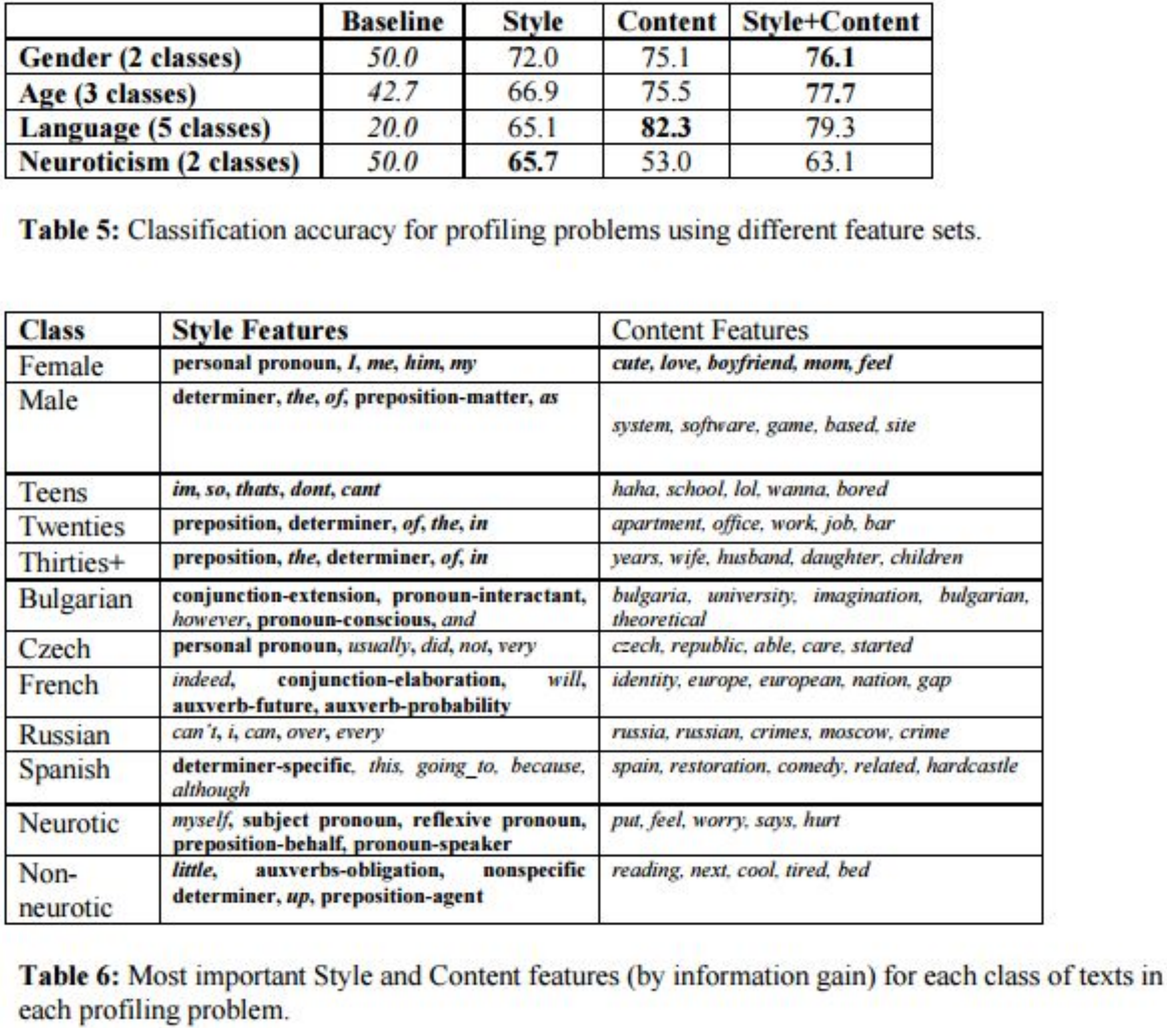
Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the literature corpus

Comparison of ML methods



features/learner	NB	J4.8	RMW	BMR	SMO
FW	38.2%	30.3%	51.8%	63.2%	63.2%
POS	34.0%	30.3%	51.0%	63.2%	60.6%
FW+POS	47.0%	34.3%	62.3%	70.3%	72.0%
SFL	35.4%	36.3%	61.4%	69.2%	71.7%
CW	56.4%	51.0%	62.9%	72.5%	70.5%
CNG	65.0%	48.9%	67.1%	80.4%	80.9%
CW+CNG	69.9%	51.6%	75.4%	86.1%	85.7%

Accuracy test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the blog corpus



	Baseline	Style	Content	Style+Content
Gender (2 classes)	*50.0*	72.0	75.1	**76.1**
Age (3 classes)	*42.7*	66.9	75.5	**77.7**
Language (5 classes)	*20.0*	65.1	**82.3**	79.3
Neuroticism (2 classes)	*50.0*	**65.7**	53.0	63.1

Table 5: Classification accuracy for profiling problems using different feature sets.

Class	**Style Features**	Content Features
Female	**personal pronoun, *I, me, him, my***	***cute, love, boyfriend, mom, feel***
Male	**determiner, *the, of,* preposition-matter, *as***	*system, software, game, based, site*
Teens	***im, so, thats, dont, cant***	*haha, school, lol, wanna, bored*
Twenties	**preposition, determiner, *of, the, in***	*apartment, office, work, job, bar*
Thirties+	**preposition, *the,* determiner, *of, in***	*years, wife, husband, daughter, children*
Bulgarian	**conjunction-extension, pronoun-interactant,** *however,* **pronoun-conscious,** *and*	*bulgaria, university, imagination, bulgarian, theoretical*
Czech	**personal pronoun,** *usually, did, not, very*	*czech, republic, able, care, started*
French	*indeed,* **conjunction-elaboration,** *will,* **auxverb-future, auxverb-probability**	*identity, europe, european, nation, gap*
Russian	*can't, i, can, over, every*	*russia, russian, crimes, moscow, crime*
Spanish	**determiner-specific**, *this, going_to, because, although*	*spain, restoration, comedy, related, hardcastle*
Neurotic	*myself,* **subject pronoun, reflexive pronoun, preposition-behalf, pronoun-speaker**	*put, feel, worry, says, hurt*
Non-neurotic	***little,*** **auxverbs-obligation, nonspecific determiner, *up,* preposition-agent**	*reading, next, cool, tired, bed*

Table 6: Most important Style and Content features (by information gain) for each class of texts in each profiling problem.

Method of Entropy Encoding

Data compression involves encoding information using fewer bits than the original representation.

Entropy encoding: replacement frequent data with short code words and less frequent with long code words.

Lossless compression reduces bits by identifying and eliminating statistical redundancy.

Lossy compression reduces bits by removing unnecessary or less important information.

Method of Entropy Encoding

Coefficient of compression:

$$k = S_o \,/\, S_c, \text{ where}$$

k - coefficient of compression;

S_o - the volume of initial data;

S_c - the volume of compressed data

The higher coefficient is, the more effective algorithm is.

Method of Entropy Encoding

The average length of the code word:

$$\overline{n} = \sum_{i} n_i P(X_i)$$

n_i - length of code word X_i;

$P(X_i)$ - probability of the code word X_i.

Method of Entropy Encoding

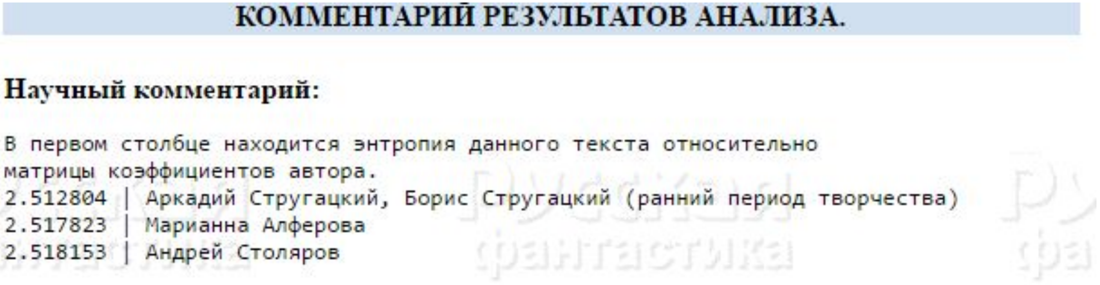
Shannon's source coding theorem:

$$\frac{H(X)}{\log_2 a} \leq ES < \frac{H(X)}{\log_2 a} + 1$$

- *H*(*X*) - source's entropy;
- *ES* - optimal length of the code word;
- *a* - the number of letters in the coding alphabet

Method of Entropy Encoding

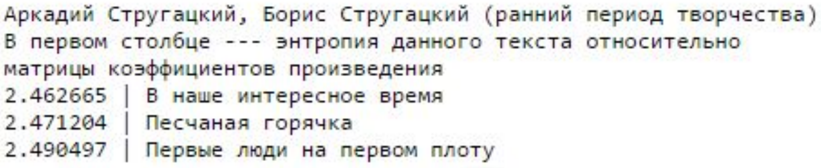
http://www.rusf.ru/books/analysis/index.htm - “Лингвоанализатор”



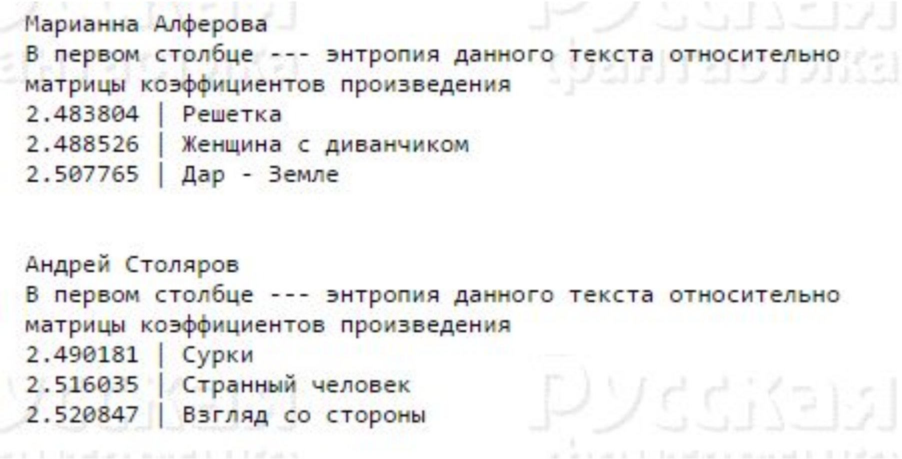
КОММЕНТАРИЙ РЕЗУЛЬТАТОВ АНАЛИЗА.

Научный комментарий:

```
В первом столбце находится энтропия данного текста относительно
матрицы коэффициентов автора.
2.512804 | Аркадий Стругацкий, Борис Стругацкий (ранний период творчества)
2.517823 | Марианна Алферова
2.518153 | Андрей Столяров
```

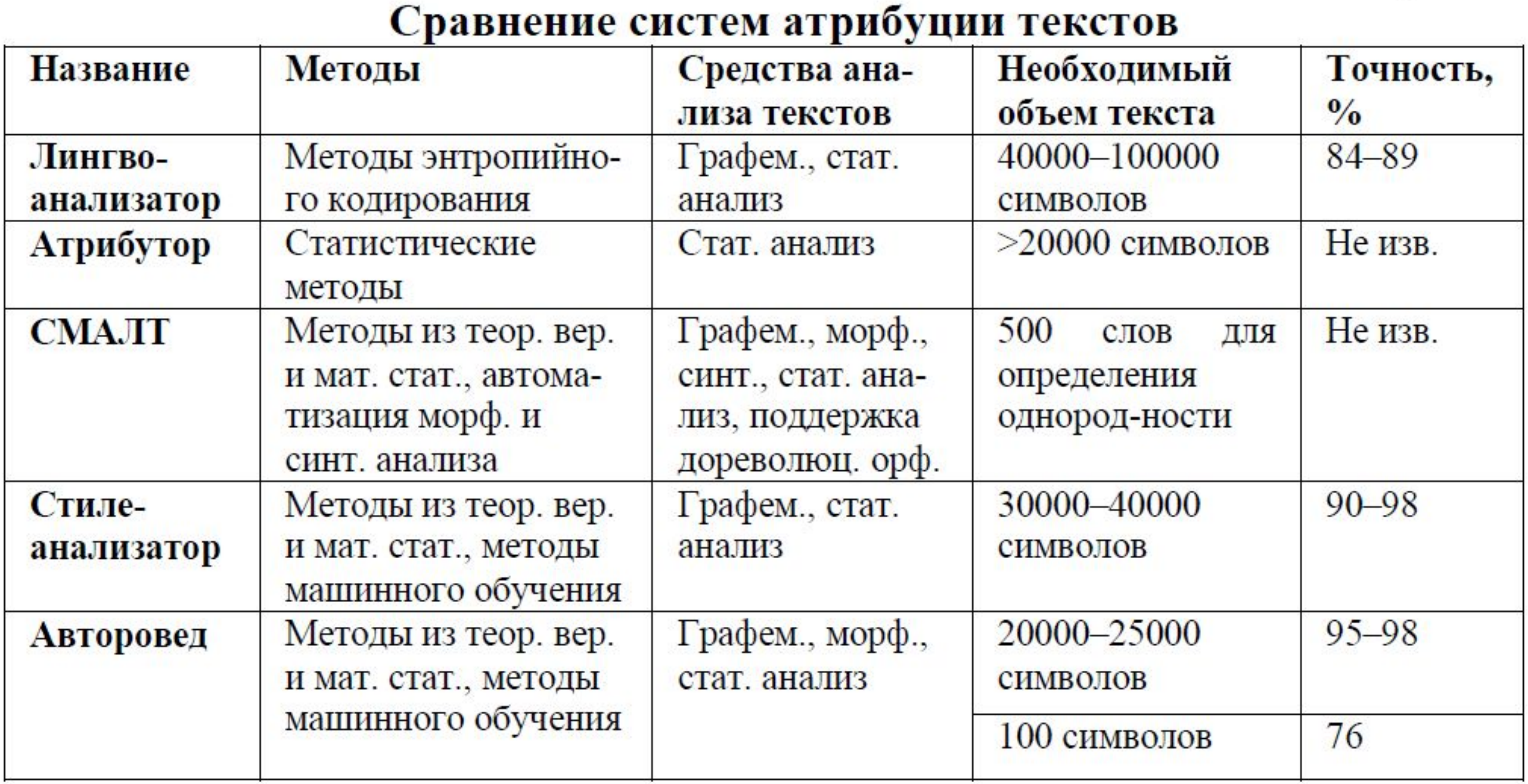


```
Аркадий Стругацкий, Борис Стругацкий (ранний период творчества)
В первом столбце --- энтропия данного текста относительно
матрицы коэффициентов произведения
2.462665 | В наше интересное время
2.471204 | Песчаная горячка
2.490497 | Первые люди на первом плоту
```



```
Марианна Алферова
В первом столбце --- энтропия данного текста относительно
матрицы коэффициентов произведения
2.483804 | Решетка
2.488526 | Женщина с диванчиком
2.507765 | Дар - Земле

Андрей Столяров
В первом столбце --- энтропия данного текста относительно
матрицы коэффициентов произведения
2.490181 | Сурки
2.516035 | Странный человек
2.520847 | Взгляд со стороны
```



Сравнение систем атрибуции текстов

Название	Методы	Средства анализа текстов	Необходимый объем текста	Точность, %
Лингво-анализатор	Методы энтропийного кодирования	Графем., стат. анализ	40000–100000 символов	84–89
Атрибутор	Статистические методы	Стат. анализ	>20000 символов	Не изв.
СМАЛТ	Методы из теор. вер. и мат. стат., автоматизация морф. и синт. анализа	Графем., морф., синт., стат. анализ, поддержка дореволюц. орф.	500 слов для определения однород-ности	Не изв.
Стиле-анализатор	Методы из теор. вер. и мат. стат., методы машинного обучения	Графем., стат. анализ	30000–40000 символов	90–98
Авторовед	Методы из теор. вер. и мат. стат., методы машинного обучения	Графем., морф., стат. анализ	20000–25000 символов	95–98
			100 символов	76

Comparison of text attribution systems

	Methods	**NLP tools**	**Required text volume**	**Precision, %**
Lingvoanalyzator	Methods of Entropic Coding	Grapheme, stat. analysis	40000-100000 chars	84-89
Atributor	Statistical methods	Stat. analysis	>20000 chars	?
SMALT	Statistical methods	Grapheme, morph., syntax analysis, pre-revolutionary orph.	500 words	?
Stile-analyzator	Statistical methods, machine learning	Grapheme, stat. analysis	30000-40000 chars	90-98
Avtoroved	Statistical methods, machine learning	Grapheme, morph., stat. analysis	20000-25000 chars	95-98
			100 chars	76