

Васильев В.Г.

**Комплексная технология  
автоматической  
классификации текстов**

ИПИ РАН

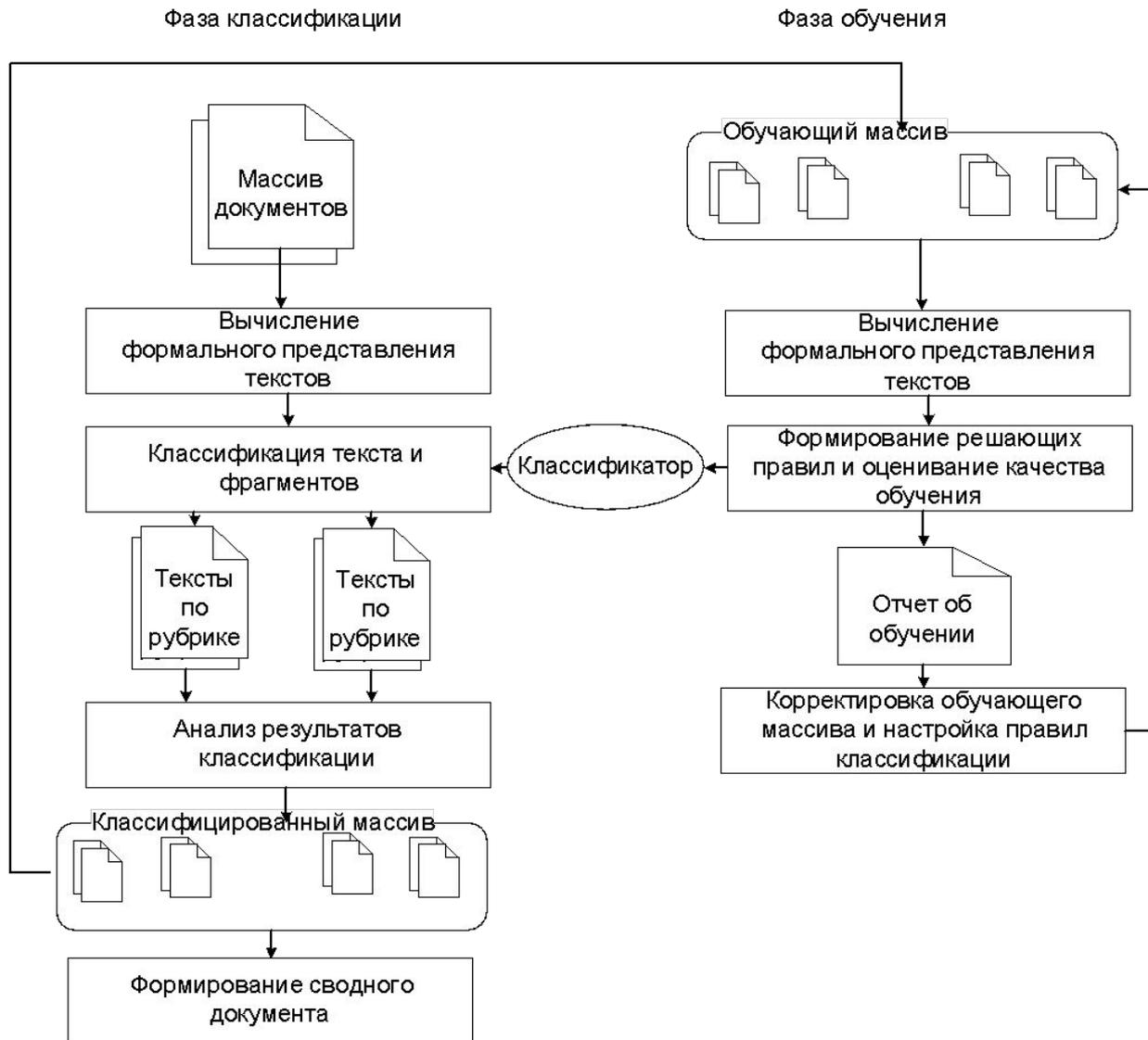
# Особенности реальных массивов текстов

- Недостаточное количество обучающих примеров
- Наличие ошибок в эталонной классификации
- Несоответствие обучающих и обрабатываемых данных
- Совместное использование нескольких принципов деления на классы
- Политематический и зашумленный характер текстов
- Сложность интерпретации результатов классификации
- Наличие повторяющейся и дублирующей информации

# Организационные проблемы

- Ограниченный доступ разработчиков систем автоматической классификации к исходным данным и массивам текстов
- Выполнение настройки и использования средств классификации пользователями, которые не являются специалистами в области автоматической обработки текстов

# Комплексная технология классификации ТЕКСТОВ



# Недостаточное количество обучающих примеров

## Прикладные проблемы:

- невозможность построения правил классификации для большинства методов, основанных на обучении по примерам;
- низкая надежность оценки качества обучения.

**Решение:** поддержка совместного использования трех типов решающих правил для рубрик:

- статистических (обучаемых на примерах документов),
- логических (задаются экспертами на специальном информационно-поисковом языке),
- шаблонных (задаются экспертами в виде регулярных выражений).

# Наличие ошибок в эталонной классификации

## Прикладные проблемы:

- формирование ошибочных правил классификации;
- результаты оценки качества обучения оказываются некорректными.

## Решение:

- выполнение при обучении оценки качества классификации и ошибок в эталонном множестве документов;
- учет степени тематической близости рубрик друг к другу;
- реализация интерактивной процедуры обучения классификатора.

# Пример оценки эталонного множества документов

Отчет о результатах обучения классификатора - Windows Internet Explorer

C:\Users\Vital\Programs\Projects\TextLearn\reuters\_docset\reuters\_report.full.html

Поиск "Live Search"

Отчет о результатах ... x Отчет о результатах обу... Домой Веб-каналы (1) Печать Страница Сервис

## Отчет о результатах обучения классификатора

### Общие показатели

Показатель	Значение (дов. 95%)
Ошибка	1% (1%, 1%)
Точность	97% (97%, 97%)
Полнота	97% (97%, 98%)
F-мера	97%

### Показатели отдельных рубрик

Номер	Название рубрики (Число документов)	Правил.	Пропущ.	Добавл.	Точность	Полнота
1	<a href="#">acq</a> (2261)	2176	33	52	98% (97%, 99%)	99% (98%, 99%)
2	<a href="#">alum</a> (59)	52	6	1	99% (90%, 100%)	90% (79%, 97%)
3	<a href="#">austdlr</a> (5)	2	2	1	67% (10%, 100%)	50% (7%, 94%)
4	<a href="#">austral</a> (0)	0	0	0	0% (0%, 98%)	0% (0%, 98%)
5	<a href="#">barley</a> (49)	46	2	1	98% (89%, 100%)	96% (86%, 100%)
6	<a href="#">bfr</a> (1)	0	1	0	0% (0%, 98%)	0% (0%, 98%)
7	<a href="#">bop</a> (102)	100	1	1	100% (95%, 100%)	100% (95%, 100%)
8	<a href="#">can</a> (3)	2	1	0	100% (16%, 100%)	67% (10%, 100%)
9	<a href="#">carcass</a> (77)	71	4	2	98% (91%, 100%)	95% (87%, 99%)
10	<a href="#">castor-meal</a> (0)	0	0	0	0% (0%, 98%)	0% (0%, 98%)
11	<a href="#">castor-oil</a> (2)	0	2	0	0% (0%, 98%)	0% (0%, 85%)
12	<a href="#">castorseed</a> (1)	0	1	0	0% (0%, 98%)	0% (0%, 98%)

Мой компьютер 100%

# Несоответствие обучающих и обрабатываемых данных

## Прикладные проблемы:

- результаты классификации текстов могут быть неопределенными;
- результаты оценки качества обучения являются завышенными.

## Решение:

- выполнение оценки качества классификации в процессе обучения;
- обеспечение переобучения в процессе обработки новой информации;
- использование дополнительных словарей квазисинонимов для повышения полноты классификации.

# Иерархический характер и использование нескольких принципов деления на классы

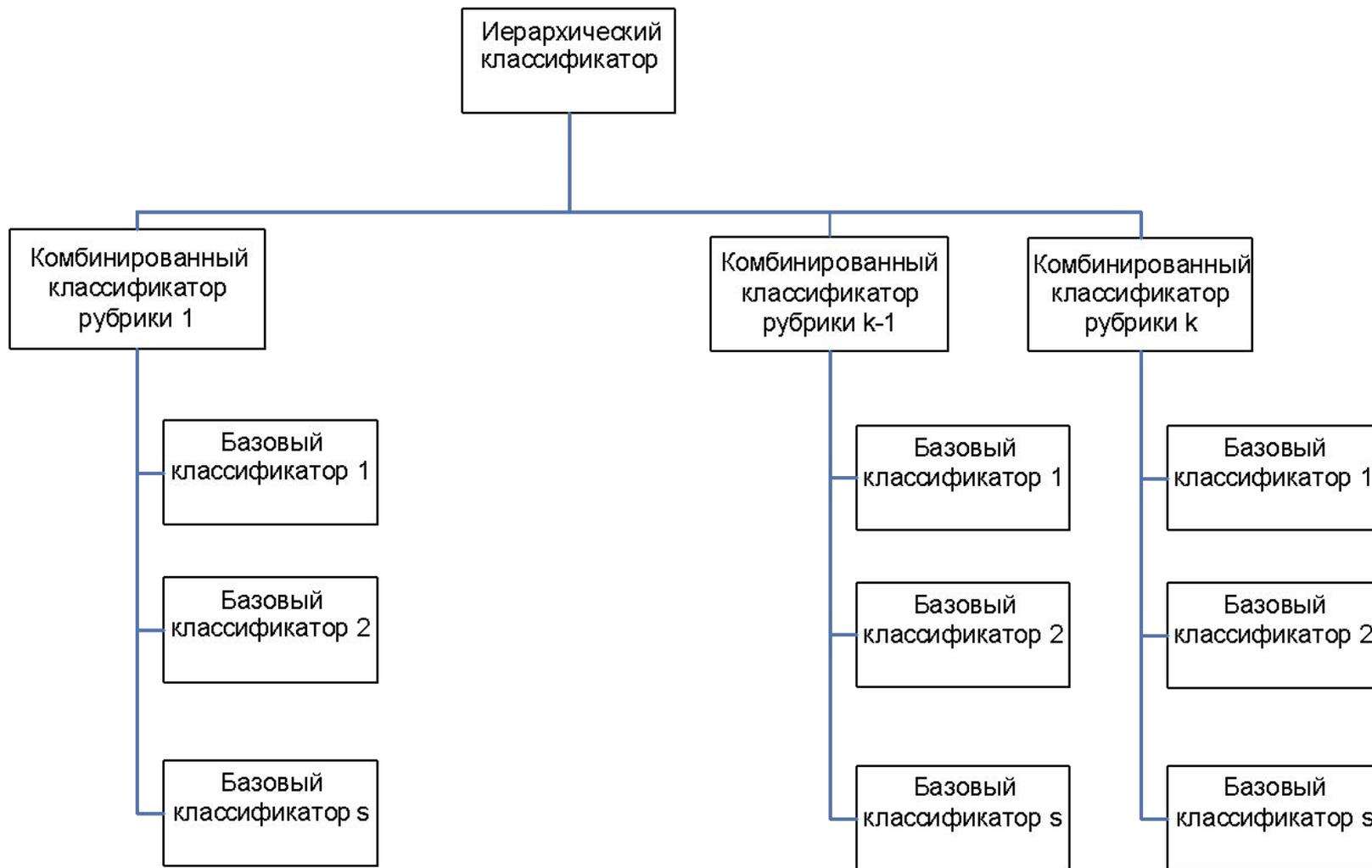
## **Прикладные проблемы:**

- сложность построения эффективных процедур классификации, основанных на использовании одной модели или метода для всех рубрик и уровней классификатора.

## **Решение:**

- поддержка нескольких типов признаков (лексических, грамматических, синтаксических);
- комбинирование различных методов классификации;
- поддержка режима фасетной классификации.

# Комбинированный иерархический метод классификации



# Базовые методы классификации

## Методы байесовской классификации

- Смесь факторных анализаторов
- Смесь полиномиальных распределений
- Смесь многомерных распределений Бернулли
- Смесь распределений фон Мизеса-Фишера

## Методы классификации на основе расстояний

- $k$ -ближайших соседей
- SVM
- Роччио

□

## Методы классификации на основе правил

- Деревья решений
- Логические правила

# Пример реализации базовых методов

Метод	К-ближайших соседей	Смесь факторных моделей
Размер обуч. множества	10 – 50000	10 – 50000
Веса признаков	TF-IDF [1]	TF-IDF
Снижение размерности	селекция по частоте документов	селекция по частоте документов, последовательный метод LSI
Оценка параметров	число соседей – 5, максимальное число эталонов – 250, оригинальный алгоритм отбора эталонов на основе кластерного анализа	оригинальный робастный алгоритм [11], размерность пространства факторов – 5
Решающее правило	байесовское правило с откл. вер. уровня 60% [2]	байесовское правило с откл. вер. уровня 60%

# Комбинированные классификаторы рубрик

## Методы построения комбинированных классификаторов

- построение фиксированного решающего правила;
- оценка качества работы базовых классификаторов;
- использование статистического моделирования (например, boosting и bagging).

□

## Метод наилучшего классификатора

- для каждой рубрики требуется хранить в памяти только одно решающее правило;
- результаты оценки качества могут использоваться для корректировки состава обучающих примеров.

# Интегральная оценка качества работы для массива «Reuters-21578-6»

Метод классификации	Точность (дов. инт.)	Полнота (дов. инт.)	F-мера	Процент обученных рубрик
<b>TREE</b>	81% (80%, 82%)	87% (86%, 87%)	84%	50%
<b>PPCA</b>	94% (93%, 94%)	95% (95%, 95%)	94%	50%
<b>MNS</b>	66% (65%, 67%)	60% (60%, 61%)	63%	65%
<b>KNN</b>	76% (76%, 77%)	73% (72%, 74%)	75%	56%
<b>BERN</b>	81% (80%, 82%)	87% (86%, 87%)	84%	70%
<b>VMF</b>	92% (91%, 92%)	93% (92%, 93%)	92%	56%
<b>ROC</b>	41% (40%, 42%)	36% (35%, 36%)	38%	70%
<b>SVM</b>	95% (95%, 95%)	96% (95%, 96%)	95%	56%
<b>Комбинированный без правил</b>	97% (98%, 99%)	97% (98%, 99%)	97%	70%
<b>Комбинированный с правилами</b>	98% (98%, 99%)	99% (98%, 99%)	99%	100%

# Оценка качества работы базовых методов для рубрик «Reuters-21578»

Рубрика (размер)	TREE	PPCA	MNS	KNN	BERN	VMF	ROC	SVM
acq (2261)	85%	95%	40%	54%	92%	95%	3%	<b>98%</b>
alum (59)	91%	85%	59%	83%	73%	89%	54%	<b>94%</b>
dmk (15)	76%	88%	88%	92%	64%	<b>97%</b>	92%	88%
housing (18)	85%	84%	84%	<b>94%</b>	81%	91%	88%	88%
l-cattle (10)	-	-	36%	<b>95%</b>	67%	90%	88%	29%
meal-feed (51)	<b>97%</b>	93%	63%	80%	81%	94%	77%	94%
palm-oil (42)	85%	<b>98%</b>	85%	91%	82%	94%	91%	94%
propane (6)	-	-	-	-	55%	-	<b>80%</b>	-
rapeseed (35)	72%	86%	76%	90%	<b>94%</b>	91%	75%	90%
sfr (3)	-	-	<b>100%</b>	-	80%	-	80%	-
soy-oil (26)	22%	36%	7%	20%	<b>51%</b>	40%	33%	26%
strategic- metal (39)	75%	<b>80%</b>	32%	61%	60%	77%	51%	73%
zinc (48)	74%	85%	60%	70%	74%	<b>91%</b>	78%	81%

# Политематический и зашумленный характер текстов

## Прикладные проблемы:

- сложность формирования решающих правил из-за негативного влияния посторонней информации и наложение рубрик друг на друга;
- неопределенность расположения в тексте информации, релевантной рубрике.

## Решение:

- идентификация форматов, языков и кодировок документов;
- очистка текста документов от элементов оформления;
- исключение из текстов вспомогательной информации;
- использование робастных алгоритмов оценивания параметров;
- выделение значимых фрагментов в текстах.

# Выделение значимых фрагментов

## Модель представления текста

Текст  $X$  представляется в виде множества векторов

$$F = \left\{ \bigotimes_{i=l_1}^{l_2} X_i \mid 1 \leq l_1 \leq l_2 \leq n \right\},$$

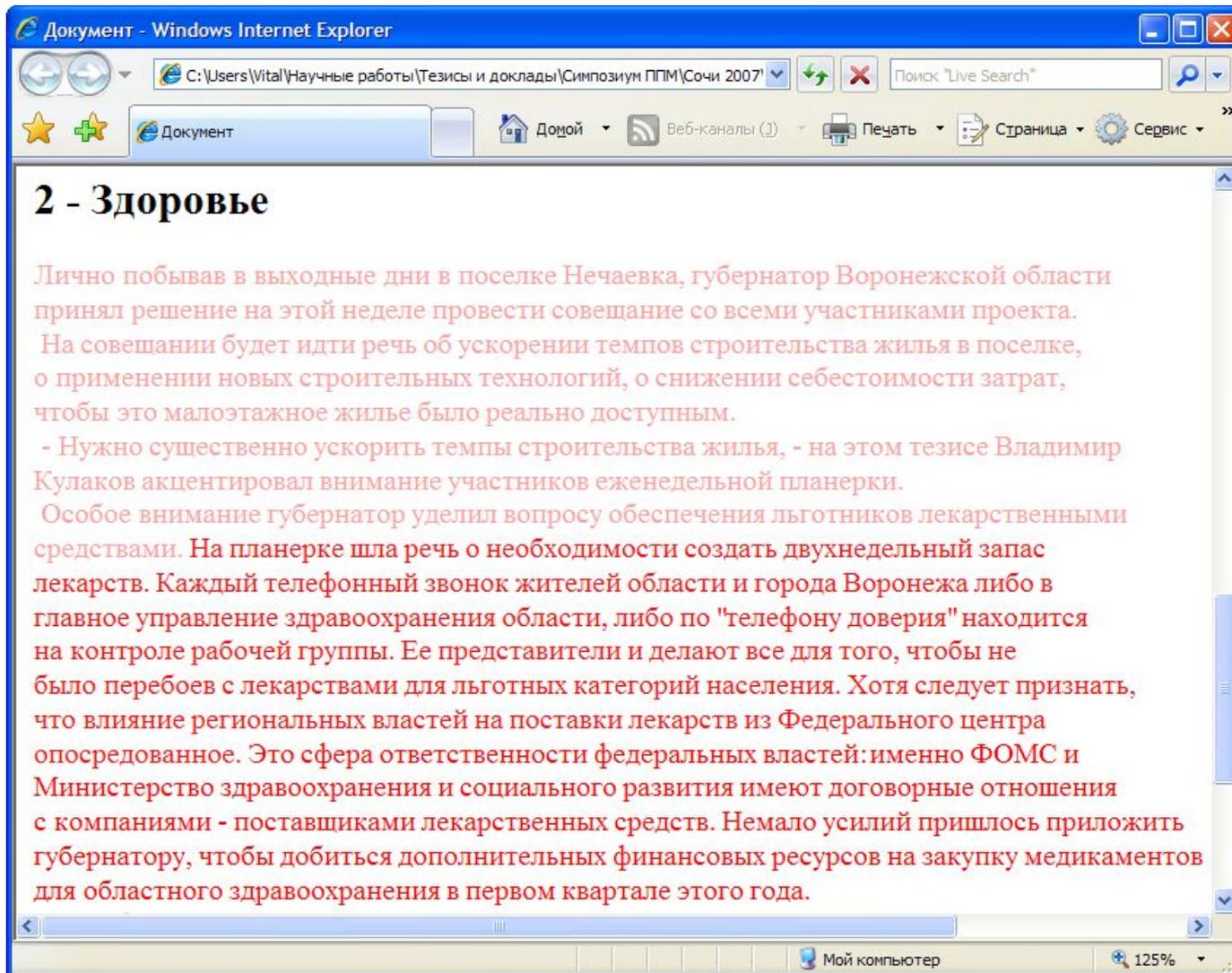
соответствующего множеству всех фрагментов текста (непрерывных последовательностей предложений).

## Оценка степени соответствия текста рубрике

$$w_{ij} = \max_{Y \in F, X_i \in Y} g_j(Y), \quad i = 1, \dots, n, \quad j = 1, \dots, k,$$

где  $g_j(Y)$  - функция оценки принадлежности текста  $Y$  к рубрике  $\omega_j$ .

# Пример разметки текста с помощью иерархического покрытия



# Наличие повторяющейся и дублирующей информации во входном потоке текстов

## **Прикладные проблемы:**

- сложность просмотра и анализа результатов классификации.

## **Решение:**

- упорядочение документов в рубриках с учетом их тематической близости друг к другу;
- выявление "почти дубликатов" документов;
- выявление основных тем документов в рубриках;
- автоматическое формирование сводных документов.

# Пример выявления основных тем в рубрике при классификации

Описание рубрики - Windows Internet Explorer

T:\Тезисы и доклады\Симпозиум ППМ\Сочи 2007\hp\_news\_docset\cls\_report.xml\_rubrs\rubr\_2.html

Поиск "Live Search"

Описание рубрики

## Здоровье

Число документов	76
Число документов с подрубриками	76
Связанные рубрики	2 <a href="#">Здоровье</a> 76
	1 <a href="#">Доступное жилье</a> 3
	3 <a href="#">Образование</a> 1
	4 <a href="#">Развитие АПК</a> 1

### Основные темы

Номер	Название темы	Размер
1.	<a href="#">АДМИНИСТРАЦИЯ ОБЛАСТЬ, КРАСНОЯРСКИЙ КРАЙ, ГЛАВНЫЙ ВРАЧ, ТЕРРИТОРИАЛЬНЫЙ ОРГАН, ДЕЙСТВОВАТЬ ЗАКОНОДАТЕЛЬСТВО</a> <b>Прокуратура выявила нарушения в реализации нацпроекта "Здоровье"</b> Прокуратурой области выявлены значительные нарушения, допускаемые администрацией городского округа город Воронеж при реализации национального проекта "Здоровье". В целях укрепления системы первичной медицинской и санитарной помощи, повышения доступности высокотехнологичных медицинских услуг, улучше	24
2.	<a href="#">ДЕНЕЖНЫЙ СРЕДСТВО, УГОЛОВНЫЙ ДЕЛО, ЛЕГАЛИЗАЦИЯ ДЕНЕЖНЫЙ СРЕДСТВО, ЛЕГАЛИЗАЦИЯ СРЕДСТВО, ДЕНЕЖНЫЙ ЛЕГАЛИЗАЦИЯ</a> <b>Смертельный наркоз от Софэкс .</b> Оригинал этого материала Вслух.Ру , 02.04.2007 Смертельный наркоз от Софэкс Московская торгово-закупочная фирма незаконно торговала медицинским эфиром, который впоследствии использовался для производства синтетических наркотиков Игорь Ковалев, Анна Исаева, Олег Градов Московская торгово-закупоч	29
3.	<a href="#">НАУЧНО-ПРОИЗВОДСТВЕННЫЙ КОМПЛЕКС, КАРДИОЛОГИЧЕСКИЙ НАУЧНО-ПРОИЗВОДСТВЕННЫЙ КОМПЛЕКС, КАРДИОЛОГИЧЕСКИЙ КОМПЛЕКС, БЛАГОТВОРИТЕЛЬНЫЙ АКЦИЯ, РОСТОВСКИЙ ОБЛАСТЬ</a> <b>Весточки</b> <b>ОБВИНЯЕМЫЙ</b> в попытке получения взятки в крупном размере заместитель директора департамента развития городского хозяйства воронежской мэрии Анатолий Батихев арестован 12 мая по решению Ленинского районного суда. БОЛЕЕ 100 тысяч верующих приложились к мошам святителя Спиридона Тримифунтс	8
4.	<a href="#">ВЫСОКОТЕХНОЛОГИЧНЫЙ МЕДИЦИНСКИЙ ПОМОЩЬ, ВЫСОКОТЕХНОЛОГИЧНЫЙ ПОМОЩЬ, ФЕДЕРАЛЬНЫЙ БЮДЖЕТ, ФЕДЕРАЛЬНЫЙ АГЕНТСТВО, ПРИОРИТЕТНЫЙ НАЦИОНАЛЬНЫЙ ПРОЕКТ</a> <b>Региональные и муниципальные клиники будут оказывать высокотехнологичную медицинскую помощь за счет федерального бюджета.</b> 14 мая 2007 года С 2007 года региональные и муниципальные учреждения здравоохранения смогут получать федеральное финансирование на оказание высокотехнологичной медицинской помощи (ВМП). Правила формирования государственного задания на этот вид медицинской помощи утверждены соответствующим постановл	5

Готово

Мой компьютер

100%