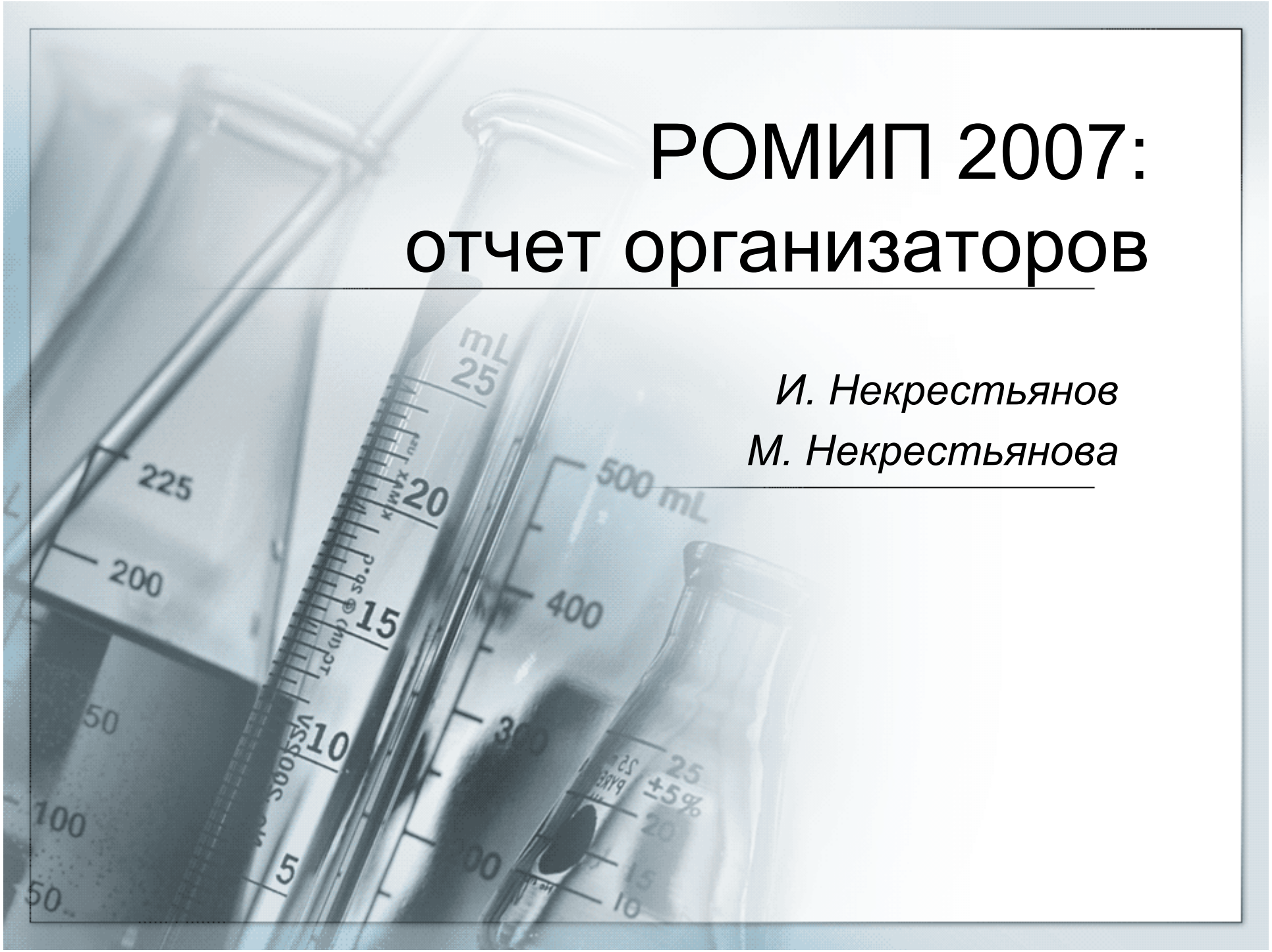


РОМИП 2007: отчет организаторов

И. Некрестьянов

М. Некрестьянова

Что такое РОМИП?

Российский семинар по оценке методов информационного поиска

- Русскоязычные задания
- Крупные коллекции
- Использование апробированных подходов к оценке
- Равноправие и анонимность участников
- Использование независимых экспертов для оценки результатов поиска
- Возможность повторного использования результатов

ПЯТЫЙ ЦИКЛ!

Что требуется от участника?

- Подать заявку
- Участвовать в формирование правил проведения дорожек
- Выполнить полученные задания и сдать ответы в оргкомитет
- Проанализировать результаты оценки и подготовить статью
- Компенсировать часть затрат на проведение оценки и организацию семинара
- Сделать доклад на очном семинаре

Принять участие не так сложно!

Как производится оценка?

- Задания для оценки отбираются оргкомитетом ПОСЛЕ сбора ответов участников
- Оценка на основе сравнения с «эталоном»
- «Эталон» обычно строится с помощью асессоров, которые оценивают ответы на правильность
- При повторении дорожки часть заданий оценивается повторно
- По результатам вычисляются некоторые стандартные метрики

Построение «эталона»

- Асессор НЕ знает какой системы это ответ
- Один асессор оценивает ответы разных систем
- Порядок выдачи заданий для оценки случаен
- Обычно минимум два асессора оценивает одно и тоже задание
- Многозначная шкала (4 значения)
- Оцениваются *уникальные* ответы (экономия)
- Использование расширенных описаний (Цель – упразднить неоднозначность трактовки, понять точку зрения асессора)

Коллекции РОМИП

	Состав	Размер	Предоставлена
Narod.Ru	Веб-сайты из домена narod.ru	728 000 док. 22 000 сайтов.	Яндекс
Legal	Законодательство РФ	60 000 док.	Кодекс
DMOZ	Веб-сайты из русской части DMOZ	300 000 док. 2087 сайтов	Рамблер
News	Все новости за три периода из 17 источников	31 500 док. 75 Мб.	Яндекс

Новые коллекции РОМИП 2007

	Состав	Размер	Предоставлена
By.Web	страницы домена .by из индекса Яндекс (май 2007) (на глубину 3 ссылки от стартовой.) процент ссылок внутрь коллекции - ~25%	1 524 676 док 8 Гб	Яндекс
KM.RU	~90% от объема www.km.ru на май 2007 (57 сайтов)	3 010 455 док 13.7 Гб	КМ Онлайн
Legal2007	Законодательство РФ, Москвы и Санкт-Петербурга (декабрь 2006)	300 000 док. 1.7 Гб	Кодекс

Дорожки по поиску

	Заданий	Оценивалось	Объем оценки	Согласие асессоров
Поиск по Веб	19628	60 (20+20+20)	BY: 6944 KM: 5801 ALL:13046	BY: 0.90 KM: 0.88 ALL: 0.89
Поиск по норм. Коллекции	14797	50	8168	0.72
Поиск по смешанной коллекции	34425	110	-	0.82
Поиск по образцу	13089	**Нет участников**	-	-

Поиск по Веб коллекциям

- Единый набор запросов для обоих коллекций
  - 20 от KM, 20 от BY, 20 старых
- Ряд участников сдал результаты выполнения заданий только по одной коллекции
  - 5 прогона по BY, 3 по KM, 3 по BY+KM
- При оценке отбирались запросы с котлами содержащими документы из обоих коллекций
  - Соотношение числа документов в котлах (BY/KM):
    - Km-запросы: 2228/2298
    - By-запросы: 2472/2033
- «Шум» в коллекции BY
  - Разные кодировки и документы на белорусском языке
    - 65 неоцененых документов для BY.Web (0 для KM.RU)
    - Асессоры меняли кодировку вручную, но участники могли об этом не задумываться

Поиск по Веб коллекциям

- Совместная оценка для разных дорожек
- Соотношение релевантных ответов (из коллекции by/из коллекции km)
  - для Km-запросов: 57/74 (AND) и 268/226 (OR)
  - для By-Запросов: 63/386 (AND) и 212/697 (OR)
  - для остальных: 118/200 (AND) и 496/406 (OR)
- Соотношение релевантных документов по происхождению (OR/AND)
  - KM: 1329/660 BY: 976/238 ALL: 2305/898
- Проблема: как сравнивать прогоны по отдельной коллекции и их объединению?
  - Наш подход:
    - пусть решают участники
    - рассчитан ряд мер в разных условиях
  - Bpref

Поиск по нормативной коллекции

- Не было возможности использовать «экспертов-юристов»
- Подход:
  - асессоры без юридического образования (2 оценки)
  - описания запросов построенные экспертами, но “понятные” обычным асессорам
    - ~ 35 запросов с описаниями на основе описаний от ранее оценивавшихся запросов
    - ~15 новых запросов отобранных профессиональным юристом, для которых он составил расширенные описания
- Очень много релевантных документов
  - >40% слаборелевантных, >20% сильнорелевантных
  - Было: в 2006 году - 16.5%; в 2005 – 29.5%
- Довольно низкое согласие асессоров (0.72)

Тематическая классификация

	Категорий	Обучающее множество	Оцени-валось	Объем оценки	Согласие асессоров
Веб сайтов	247	2116 сайтов	26 (19+7)	1420 (1248+ 172)	0.81
Веб докум.	247	2116 сайтов	19	3814	0.79
Нормат. докум.	727	36300 док.	В процессе	0	-

- Обучение на DMOZ для Веб-классификации
- Есть дубликаты в обучающем множестве по legal
- ~50 обучающих примеров для каждой legal категории

Дорожки классификации (2)

- Попытка сделать более “сбалансированные” котлы
  - Проблема “Больших котлов” =>
    - Сайты: отбор не очень больших котлов (<400).
      - Но не менее 3 категорий из общего крупного раздела
    - Страницы: оценка выборок по 200 документов на категорию
  - Как примерно одинаково оценить разные прогоны?
    - Важно пересечение ответов (релевантный кусок?)
    - Важны и особенности конкретного прогона (помогает аппроксимировать полноту)
- Проблема: один участник вернул ранжированный список ответов (но только первую 1000 документов)
  - Выбор в котел не случаен, а из начала списка
  - Его доля релевантных больше:
    - AND: 33% против 15% (+100), OR: 50% против 32.5% (+250)

Организация новостей

Задача: структурировать поток новостей в набор событийных сюжетов, связанных между собой.

Событие: отражение события из реального мира в потоке новостных сообщений
(«смерть папы римского»)

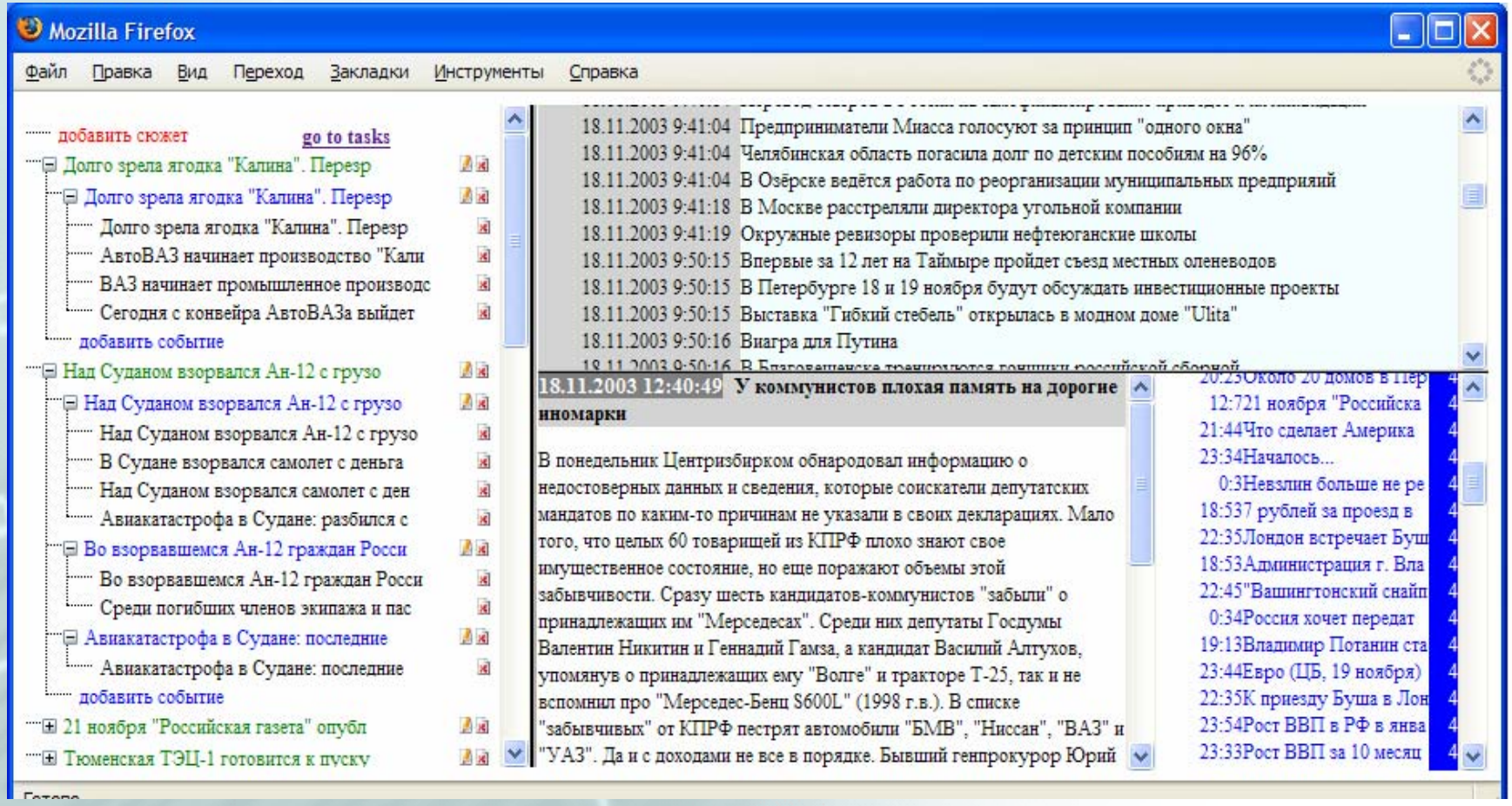
Сюжеты – события, связанные между собой нетематическими связями
(«перенос футбольного тура в связи с кончиной папы»)

- 2006: 2 участника, 4 прогона
- 2007: 2 (новых) участника, 3 прогона

Подходы к оценке

- Позиция “редактора”
  - Полная разметка новостного потока вручную
  - Хорошо позволяет оценить полноту (и точность)
  - Легко повторно применяемо
  - Но проверяется ответ для “куда бы я бы положил это на месте редактора”
- Позиция “читателя”
  - Проверяем насколько хороши результаты системы
    - Нет ли мусора в выделенном событии/сюжете?
  - Трудно оценить полноту и переиспользовать
  - Ближе к реальному конечному пользователю
- Гибридные подходы

Инструмент асессора 2006



Проблемы оценки 2006

- Что оценивалось
  - все сообщения за первые два дня для каждой недели
  - ~3500 в каждой неделе
- Проблема:
  - Очень сложная задача
- Следствия:
  - мало выделенных кластеров в первых результатах
  - распределение кластеров не совпадает с ожиданиями
  - сильные отличия у разных асессоров
  - внутренняя структура сюжетов часто недоработана (очевидные дубли не склеены)

Почему сложная?

- Выбор между сотнями готовых кластеров
  - Назначение в кластер реализовано как DnD операция!
    (перетащить файл в одну из нескольких сотен существующих папок)
- Невозможно быстро понять о чем этот сюжет не просмотрев сообщения еще раз
- Контекст быстро забывается, если асессор отвлекся на что-то другое

Чтобы стать хорошим редактором
надо пройти обучение!

Подход 2007 года

- Отталкиваться от результатов систем
- Идея 1: Перепроверка ответа системы (пометить ошибки)
  - Нет оценки полноты
- Идея 2: Объединить ответы разных систем
  - Позволяет как-то оценить полноту
  - Как объединять?
    - Хочется получить ответ с относительно полными выделенными сюжетами
- Почему это проще?
  - Уменьшается анализируемое множество
  - Названия более описательные

Оценка в деталях

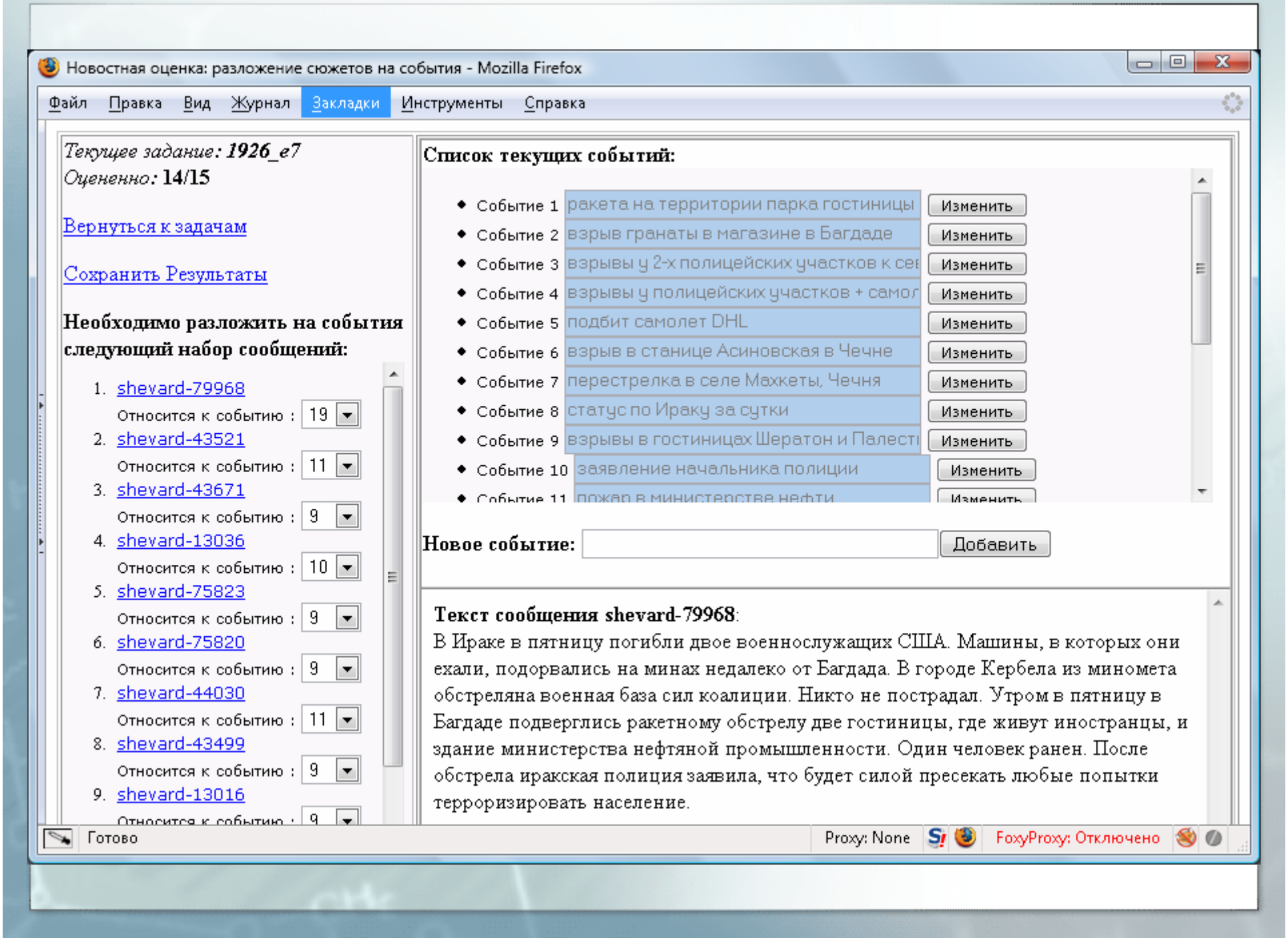
- Для каждого сообщения строим
  - Событийный котел
    - Все сообщения который хоть кто-то отнес к тому же событию, что и данное
  - Сюжетный котел
    - Аналогично, но для сюжетов
- Вычисляем размер “ядра” для каждого котла
  - Число сообщений из котла, котлы которых совпадают с данным (то есть все системы отнесли их к одному событию/сюжету)

Оценка в деталях (2)

- Для каждого сюжетного котла строим покрытие событийными котлами
  - Выбираем сообщение, которое еще не покрыто и имеет максимальный размер ядра
  - Добавляем его событийный котел к покрытию
  - Повторяем до получения полного покрытия
- Что оценивать:
  - Отбор вручную сюжетных котлов, принципы:
    - Относительная “внятность” и “интересность” котла
    - Наличие “схожих” других котлов

Задача асессора

- Этап 1
  - Дан событийный пул
  - Разбить сообщения на группы, задав каждой группе имя события
    - В одной группе должны быть одинаковые новости про одно событие (эхо события из реального мира)
- Этап 2
  - Даны списки названий групп (можно посмотреть состав)
  - Объединить те, что представляют одно событие
- Этап 3
  - Даны списки названий групп
  - Сгруппировать их в сюжеты


Новостная оценка: разложение сюжетов на события - Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
Текущее задание: 1926_e7
Оценено: 14/15
Вернуться к задачам
Сохранить Результаты
Необходимо разложить на события следующий набор сообщений:
1. shevard-79968
Относится к событию : 19
2. shevard-43521
Относится к событию : 11
3. shevard-43671
Относится к событию : 9
4. shevard-13036
Относится к событию : 10
5. shevard-75823
Относится к событию : 9
6. shevard-75820
Относится к событию : 9
7. shevard-44030
Относится к событию : 11
8. shevard-43499
Относится к событию : 9
9. shevard-13016
Список текущих событий:
Событие 1 ракета на территории парка гостиницы Изменить
Событие 2 взрыв гранаты в магазине в Багдаде Изменить
Событие 3 взрывы у 2-х полицейских участков к се Изменить
Событие 4 взрывы у полицейских участков + самол Изменить
Событие 5 подбит самолет DHL Изменить
Событие 6 взрыв в станице Асиновская в Чечне Изменить
Событие 7 перестрелка в селе Махкеты, Чечня Изменить
Событие 8 статус по Ираку за сутки Изменить
Событие 9 взрывы в гостиницах Шератон и Палести Изменить
Событие 10 заявление начальника полиции Изменить
Новое событие: Добавить
Текст сообщения shevard-79968:
В Ираке в пятницу погибли двое военнослужащих США. Машины, в которых они ехали, подорвались на минах недалеко от Багдада. В городе Кербела из миномета обстреляна военная база сил коалиции. Никто не пострадал. Утром в пятницу в Багдаде подверглись ракетному обстрелу две гостиницы, где живут иностранцы, и здание министерства нефтяной промышленности. Один человек ранен. После обстрела иракская полиция заявила, что будет силой пресекать любые попытки терроризировать население.
Готово
Proxy: None
FoxyProxy: Отключено

Текущий статус

- Построены котлы по всем доступным прогонам
- Отобрано 17 сюжетных котлов с числом сообщений ~20, 60, 200
  - 1105 сообщений, 2257 покрытие
  - размер ядер 1-17
  - 1 пул из 080404 (образец), 3 пула из shevard, 13 из vybory
- Оценено примерно половина событийных пулов (этап 1 и 2)
- Метрики?
  - Сравнение кластеризации с эталоном (например, энтропия)

Ошибки на уровне событий

- Временной разброс
  - Курсы валют за разные даты
- Географический разброс
  - Теракт в Ираке/взрыв в Чечне
- Несвязанные события
  - Курс USD/курс EUR
- “Обзорные” сообщения (сообщение – эхо многих событий)
  - Сообщение с обзором всех инцидентов в Ираке за день вызывает отнесение напрямую несвязанных сообщений про отдельные инциденты к одному событию

Особенности РОМИП'2007

- 3 новых (больших) коллекции
- Большая новая таксономия для нормативных документов
- 5 запросов на коллекции/результаты для использования в исследованиях (за последний год)
- Несколько участников сошло с дистанции
  - нет участников в дорожках: QA, поиск по образцу и только 1 участник в дорожке аннотирования
- Большое число опозданий и проблем с результатами
  - "Расслоение" сроков сдачи – невозможность запустить оценку ни по одной дорожке
- Накладки с расписанием у оргкомитета
- Труды НЕ ГОТОВЫ к очному семинару

А будут ли опубликованы труды?

- На сайте по мере поступления статей (начиная со следующей недели)
- В печатном виде по завершении формирования сборника
- Печатные сборники можно будет бесплатно получить в Петербурге и Москве

Вопросы?