Фактографическое аннотирование новостных сюжетов

Лев Гершензон, Александр Головко

16.04.2007

План

- Что такое Яндекс.Новости?
  - Автоматическая кластеризация сообщений в сюжеты
  - Ранжирование новостных сюжетов
  - Автоматическое аннотирование сюжета: выбор заголовка, текста, картинки
- Выделение объектов из текста
- Аннотирование кластера документов (сюжета)
  - Выбор наиболее релевантных объектов и фактов
  - Выбор предложений для аннотации

Яндекс.Новости

Автоматическая кластеризация 80 000 новостных сообщений в сюжеты – новости об одном событии.

- Определение ключевых слов документа
- Поиск для каждого документа по его ключевым словам близких документов
- Многопроходная кластеризация документов по специально построенным из документа и по пользовательским запросам

Ранжирование сюжетов

- количество сообщений
- новизна
- пользовательский интерес
  «новостные» запросы
  «кликабельность» сюжетов

Яндекс.Новости. Страница рубрики

Политика

выпуск: Россия | Украина

Предприниматель Борис Березовский готовит революцию в России **(67)** 190 мнений
Генпрокуратура России несколько раз уже пыталась вернуть Березовского на родину.

На "Марш несогласных" в Москву прислан региональный ОМОН **(82)** 91 мнение карта
Прокуратура официально предостерегла Михаила Касьянова и Эдуарда Лимонова от участия в "Марше несогласных", намеченного на 14-ое апреля в Москве.

Депутаты обеспокоены засильем рекламы "секса по телефону" в СМИ **(61)** 24 мнения
"В эфире отечественных телеканалов появилось много ориентированных преимущественно на подростков программ явно сексуального характера", - сказал Катренко, отметив, что "поставщики услуг сексуального характера" резко активизировали свою деятельность.

Чужих Кремлей мы не хотим ни пяди **(37)** 34 мнения
"В ближайшем обозримом будущем президентом Белоруссии будет Александр Лукашенко", - заявил белорусский лидер.

Юрий Лужков пообещал, что в этом году в Москве родится не менее 100 тысяч детей **(42)** 23 мнения
Главная интрига состоявшейся 12 апреля встречи московского мэра с президентом России заключалась в том, поставит ли Юрий Лужков перед Владимиром Путиным вопрос о доверии.

Киргизская оппозиция объявила конкурс патриотической песни **(166)** 100 мнений

СБ ООН намерен принять резолюцию о ситуации в зоне грузино-абхазского конфликта **(218)** 24 мнения

Россель поделился с Абрамовичем воспоминаниями **(86)** 32 мнения

Лидер "Справедливой России" предложил создать Социнтерн **(47)** 43 мнения

Взрывы офисов не помешают "Процветающей Армении" продолжать политическую борьбу **(44)**

Цхинвали пообещал уничтожить грузинских полицейских **(80)** 30 мнений

Кадыров утвердил трех вице-премьеров **(176)**

Я

Яндекс.Новости. Страница сюжета

Заголовок

- Соответствие лексическому ядру
- «Красота»: длина, синтаксическая полнота
- Новизна

Картинка

Аннотация

Сюжет в лицах, Карта к сюжету

Список сообщений, составляющих сюжет

- Отсортирован по времени
- Релевантные, не дублирующиеся сообщения

Яндекс.Новости. Страница сюжета

"Другая Россия", вопреки запретам столичных властей, намерена пройти по российской столице

Фото: Вслух.Ру

По данным очевидцев, на улицах Москвы видны машины с бойцами ОМОНа из Удмуртии, Мордовии, Пензенской и Нижегородской областей, на Пушкинской площади уже заготовлены десятки металлических ограждений. // REGNUM 14:39

Представители *"Другой России"* 30 марта подали заявку в *мэрию Москвы* на проведение 14 апреля в центре города - на Пушкинской площади и Тверской улице "марша несогласных". // Правда.ru 08:25


Фото: Московский комсомолец

Прокуратура официально предостерегла Михаила Касьянова и Эдуарда Лимонова от участия в "Марше несогласных", намеченного на 14-ое апреля в Москве. // Эхо Москвы 12.04.07 22:58

"Согласным", с той же радостью, предоставляют площадки, в которых отказывают "несогласным", как это произошло с Пушкинской площадью в Москве. // Радио Свобода 12.04.07 12:00


Фото: Российская газета

Уже почти месяц многие знали, что время и место сбора - в 12 часов на Пушкинской площади. // Всероссийский гражданский конгресс 12.04.07 12:54

«...» Как рассказала советник политика Елена Дикун, в четверг Касьянов получил соответствующее письмо. // Радио Свобода 12.04.07 18:44

Всего в сюжете: сообщений: **117**, фото: **29**
Первое сообщение: Эхо Москвы, 11.04.07 17:08

20:11 Нескучная суббота-2 Lenta.ru

14:39 Бойцы Тульского ОМОНа собираются на "Марш несогласных" REGNUM

13:01 В Москве милиция задерживает прибывающих из регионов участников "Марша несогласных" Радио Свобода

Карта к сюжету

Посмотреть на большой карте

Сюжет в лицах

Эдуард Лимонов
Михаил Касьянов
Гарри Каспаров
Виктор Бирюков
Сергей Удальцов

Подписка на новости

Новости на вашем сайте

Извлечение фактов из текстов сюжета

Извлекаемые объекты и факты

- ФИО
- названия организаций
- географические объекты
- даты и числа
- цитаты

Справочная информация

адрес – ссылка на карту

фио – ссылка на пресс-портрет

новостной источник – ссылка на сайт/статью

Извлечение фактов из текста

12 марта этого года задержан заместитель главного бухгалтера финансово-экономического управления *УВД* Хабаровского края **Владимир Дуничев**, похитивший более 10 миллионов рублей.

Отбор предложений для аннотации

- отождествление объектов одного типа из разных документов сюжета
- приписывание объектам весов по упоминаемости и по типу
- выбор всех предложений из всех документов, содержащих ключевые слова сюжета
- взвешивание предложений по входящим в них ключевым словам и входящим в них фактам

Отбор предложений для аннотации

•просев полученных предложений:

- по шинглам – удаление лексических дублей

4 апреля гособвинение потребовало приговорить Ульмана и Перелевского к 23 годам тюрьмы, а Воеводина и Калаганского - к 18 годам.

Гособвинение требует приговорить Эдуарда Ульмана и Алексея Перелевского к 23 годам лишения свободы каждого, Александра Калаганского - к 18 годам.

- по объектам – удаление содержательных дублей

На процессе в Северо-Кавказском военном суде объявлен перерыв до 13 апреля из-за неявки троих обвиняемых Эдуарда Ульмана, Александра Калаганского и Владимира Воеводина.

Подсудимые по делу о расстреле чеченских жителей Эдуард Ульман, Александр Калаганский и Владимир Воеводин не явились в четверг на заседание Северо-Кавказского военного суда.

- выбор из дублирующихся самого раннего

•выбор N самых весомых предложений

Пути развития

- Учет сценария события для определения необходимых составляющих аннотации
  - *Футбольный матч*
  - *Пожар*
  - *Принятие нового закона*
- Улучшение связности текста аннотации

Спасибо!