Автоматизация подготовки исходных текстовых данных из сети интернет для дальнейшего анализа

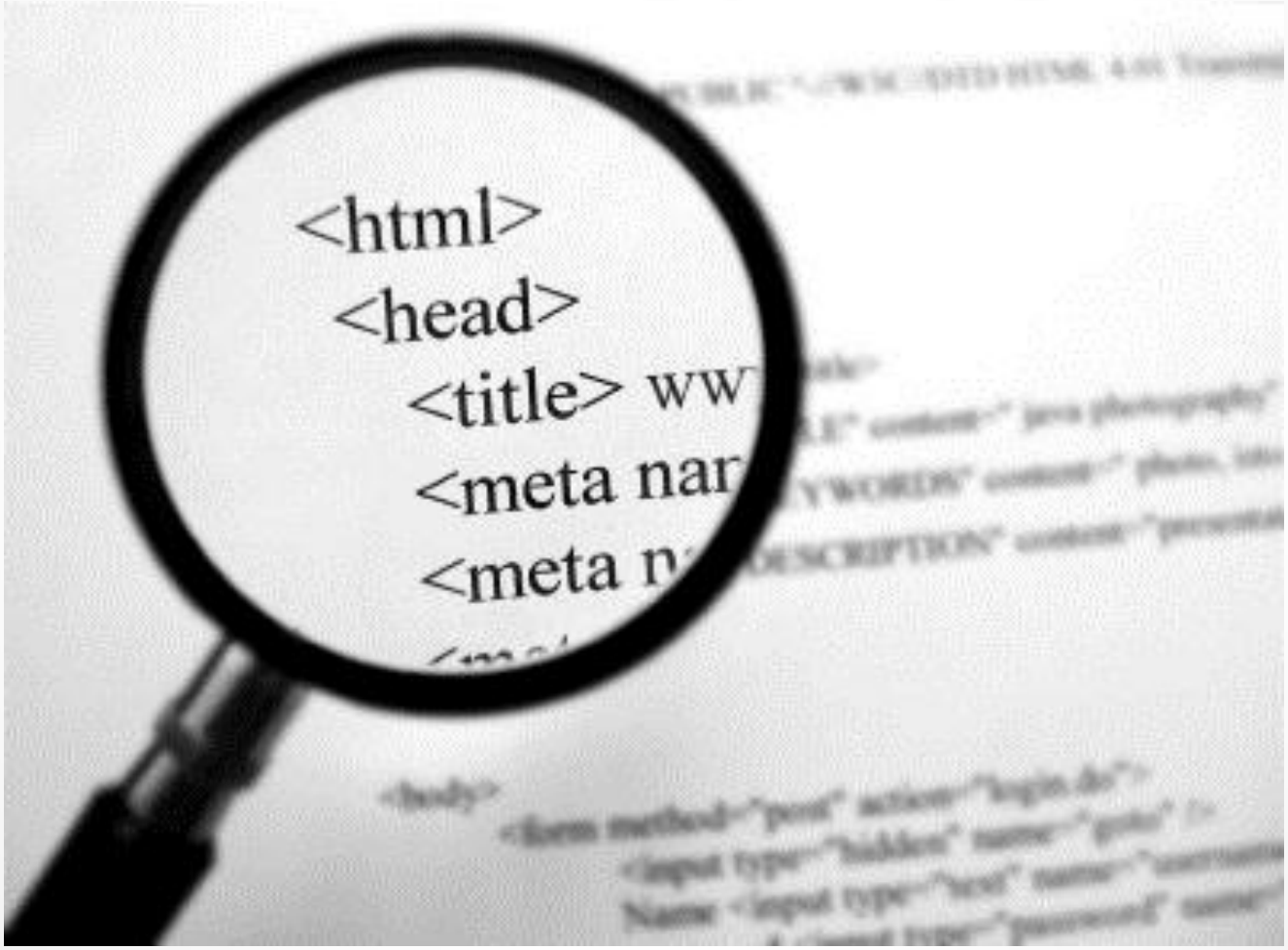
Найденов Никита Анатольевич

naidyonov@gmail.com

Вычислительный центр им. А.А. Дородницына РАН

2012

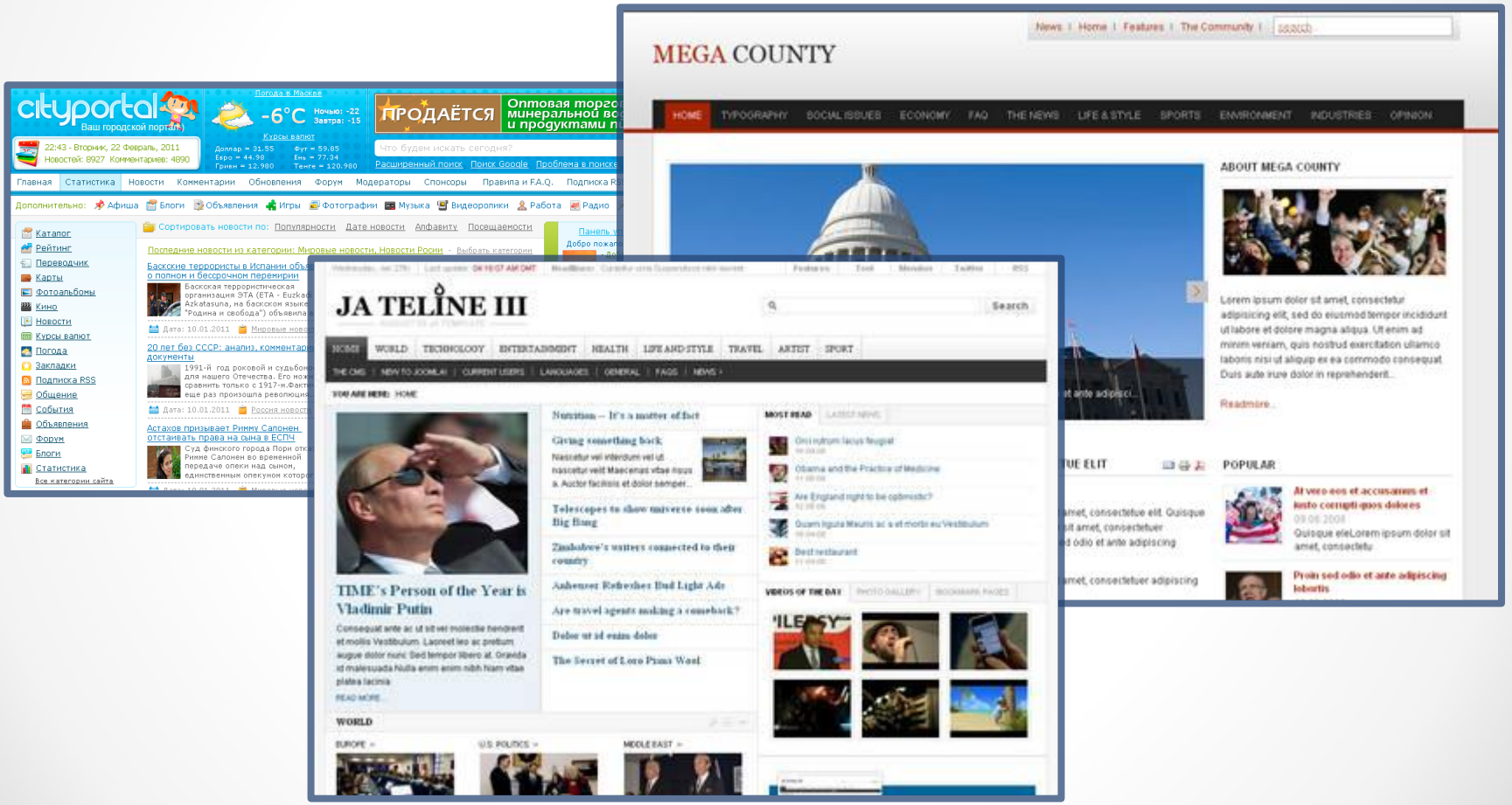
WEB структура



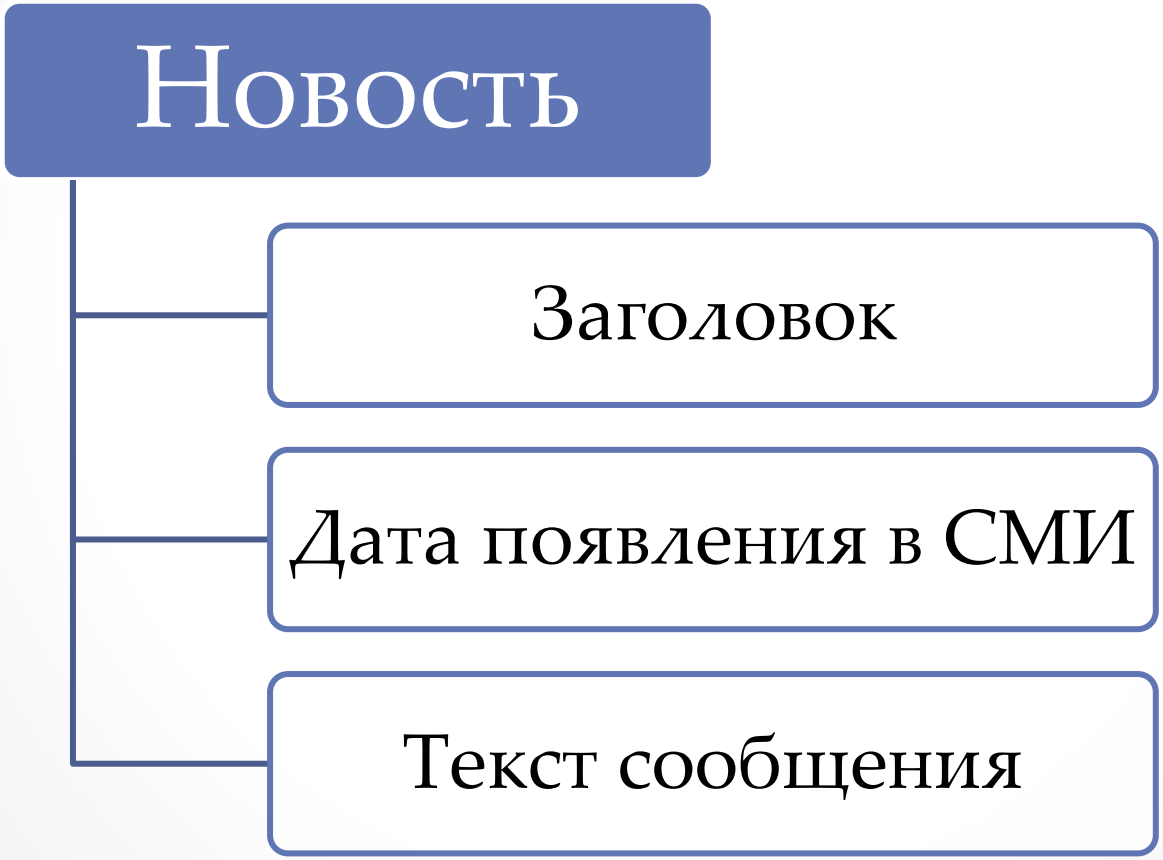
Предварительная обработка данных

- Консолидация
  - это комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование к единому формату, в котором они могут быть загружены в хранилище данных или аналитическую систему
- Трансформация
  - комплекс методов и алгоритмов, направленных на оптимизацию представления и форматов данных с точки зрения решаемых задач и целей анализа
- Очистка
  - процесс выявления и исключения различных факторов, мешающих корректному анализу данных: аномальных и фиктивных значений, пропусков, дубликатов и противоречий, шумов и т.д.

Описание данных



Описание данных



Постановка задачи

Требуется собрать новостные материалы из открытых Интернет источников за определенный период времени.

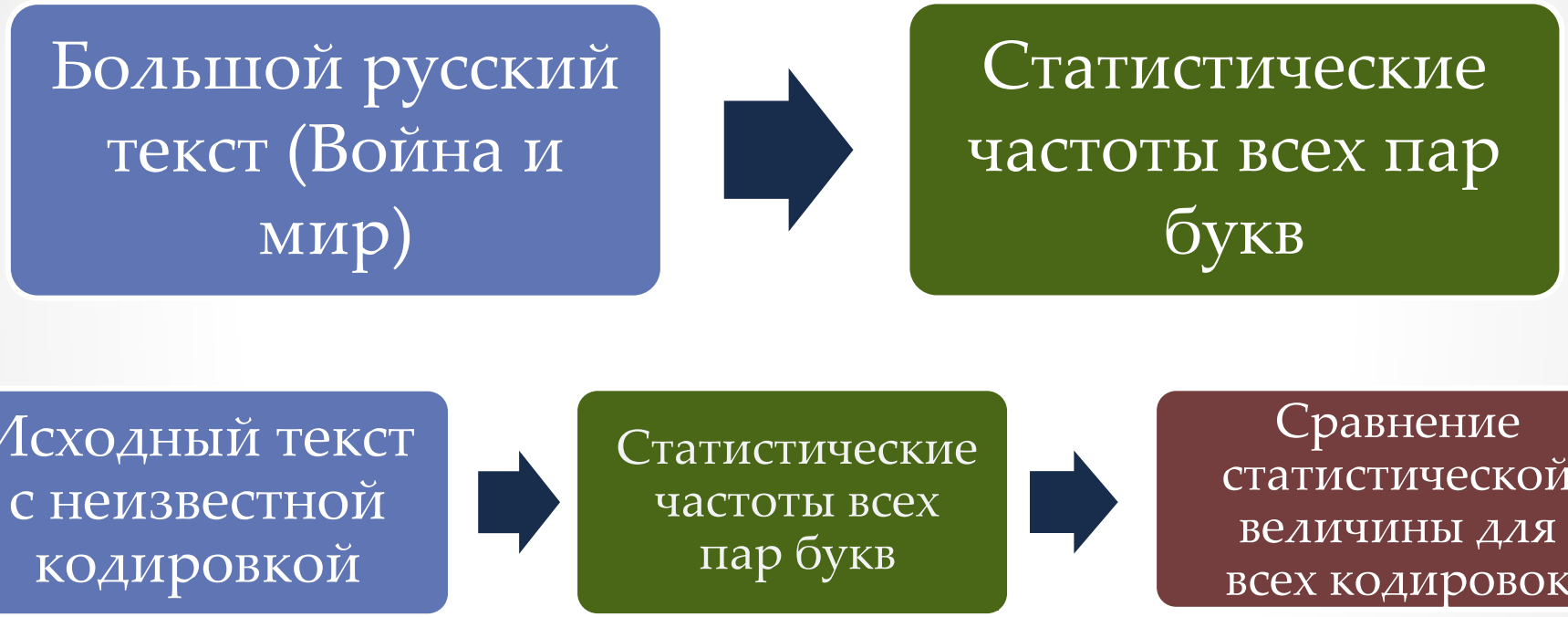
Источники данных – новостные ресурсы (список ссылок на сайты)

Результат - список новостей за определенный период

Проблемы

- Разнородность ресурсов
  - Кодировка
  - Ограничение доступа
  - Структура
- Большой объем данных

Кодировка

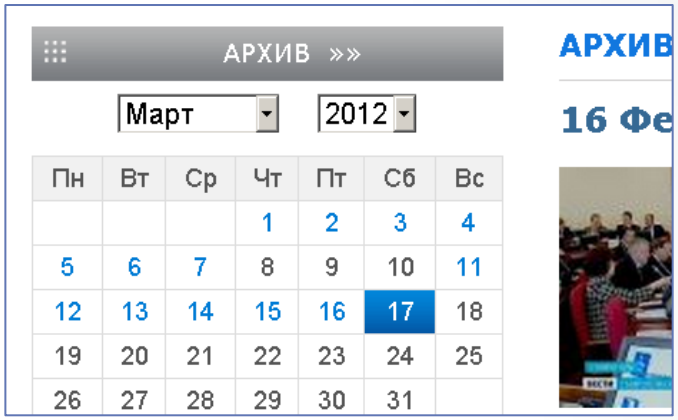

Большой русский текст (Война и мир)
Статистические частоты всех пар букв
Исходный текст с неизвестной кодировкой
Статистические частоты всех пар букв
Сравнение статистической величины для всех кодировок

Ограничение доступа

- Скорость ответа на запрос от ресурса
- Ограничение в частоте выполнения запросов

Форматы источников

1. Календарь



2. Список новостей

Формат источника с календарем

1. Формат ссылок на даты

   http://tula.rfn.ru/archive/index.html?date=**14-03-2012**

2. Формат ссылок на новости

   …/archive/index.html?id=**162784**&date=**14-03-2012**

3. CSS Selectors на определение полей
   - Дата – “span.data”
   - Заголовок – “span.mainhead4”
   - Текст – “span.text”

4. CSS Selectors “мусора”

Формат источника со списком новостей

1. Формат ссылок на даты

   http://vesti-ural.ru/page/**2**/

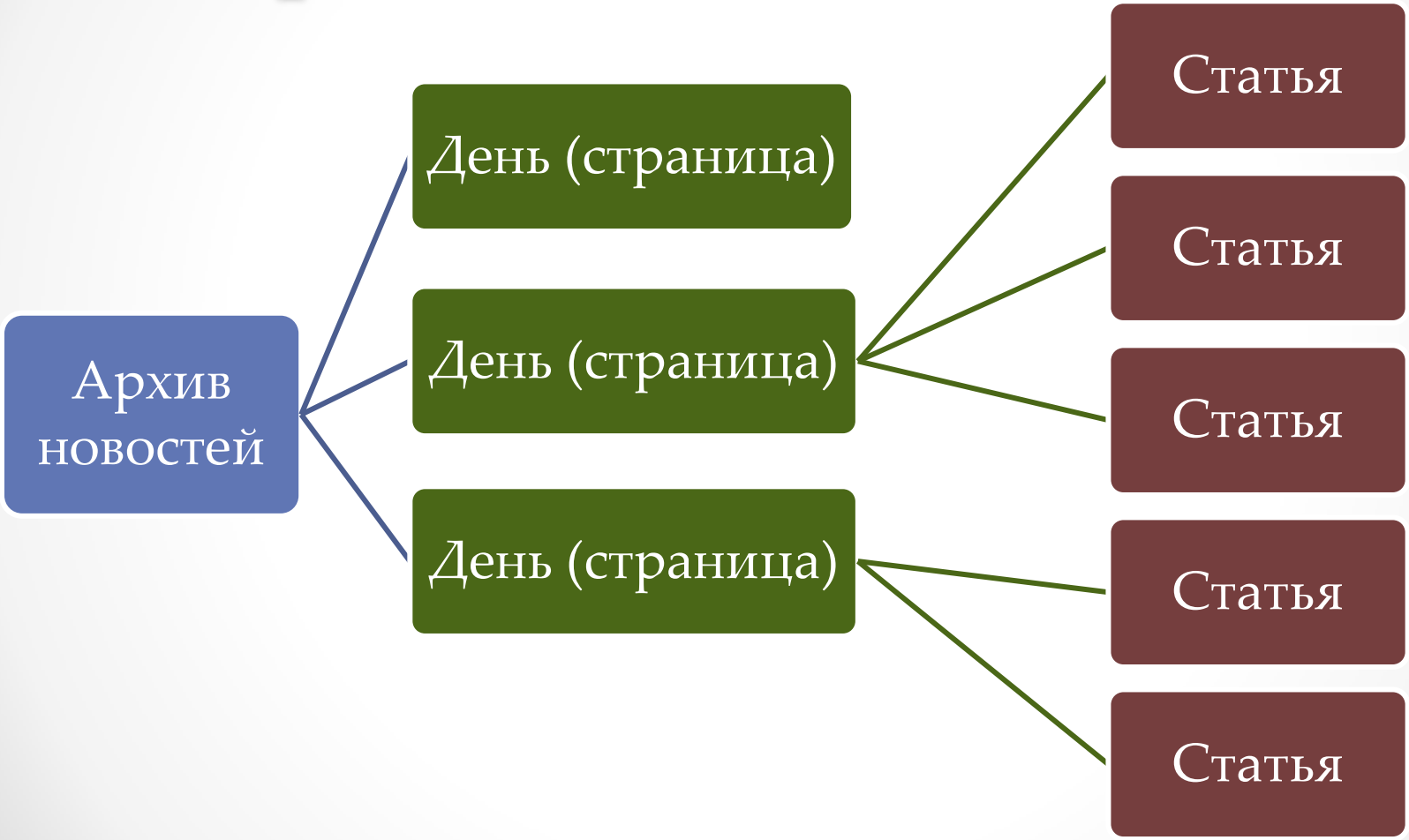
2. Формат ссылок на новости

   http://vesti-ural.ru/**7732-v-kurga...tyre-passazhira**.html

3. CSS Selectors на определение полей
   - Дата – ".block_full_news > div.data"
   - Заголовок – ".block_full_news > h1"
   - Текст – ".block_full_news > div#newsid_7732"

4. CSS Selectors "мусора"

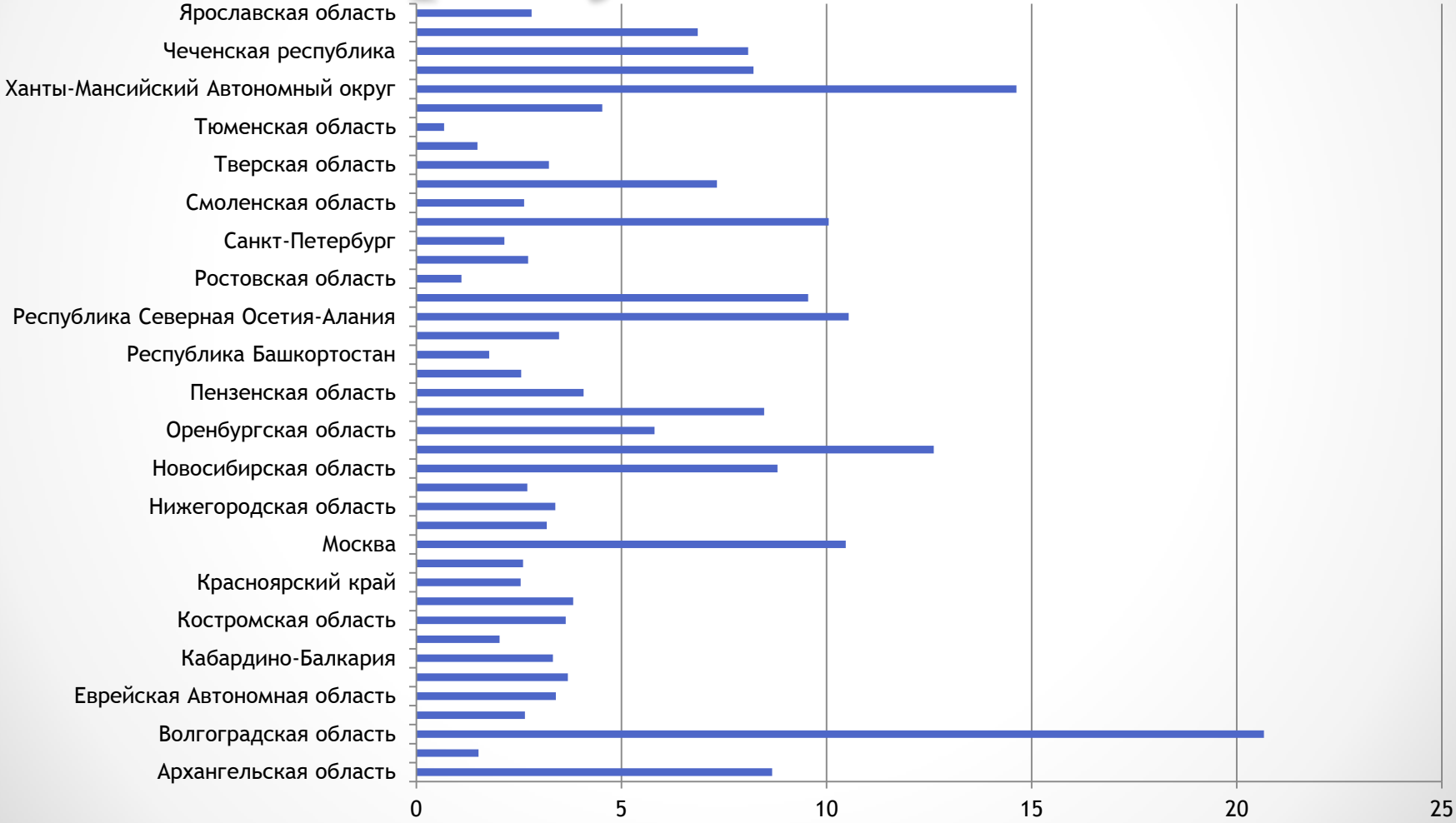
Параллельная обработка источника



Реальные данные

- **43 региональных новостных ресурса** (большинство относится к Государственной телерадиокомпании)

Описание результатов



Описание результатов

- Параллельная обработка ресурсов уменьшает время сбора данных ~ в 10 раз
- Общий объём данных – **230 Мб**

Автоматизация системы

1. Поиск календаря/списка на сайте
2. Формирование шаблона ссылок на страницы со списком новостей
3. Формирование шаблона ссылок на новости
4. Извлечение даты, заголовка и текста

Автоматизация системы

- Обработка атрибута **href** и содержания у тегов **<a> **
- Extracting useful text from HTML

Спасибо за внимание!