Парсинг - получение контента с web-страниц на PHP

Запросы к web-страницам

- CURL

Выбор текстовых данных из структуры HTML-документа

- Simple HTML DOM Parser,
- PHPQuery

1.Библиотека CURL (client urls)

библиотека функций, позволяет взаимодействовать с множеством различных серверов по многим различным протоколам.

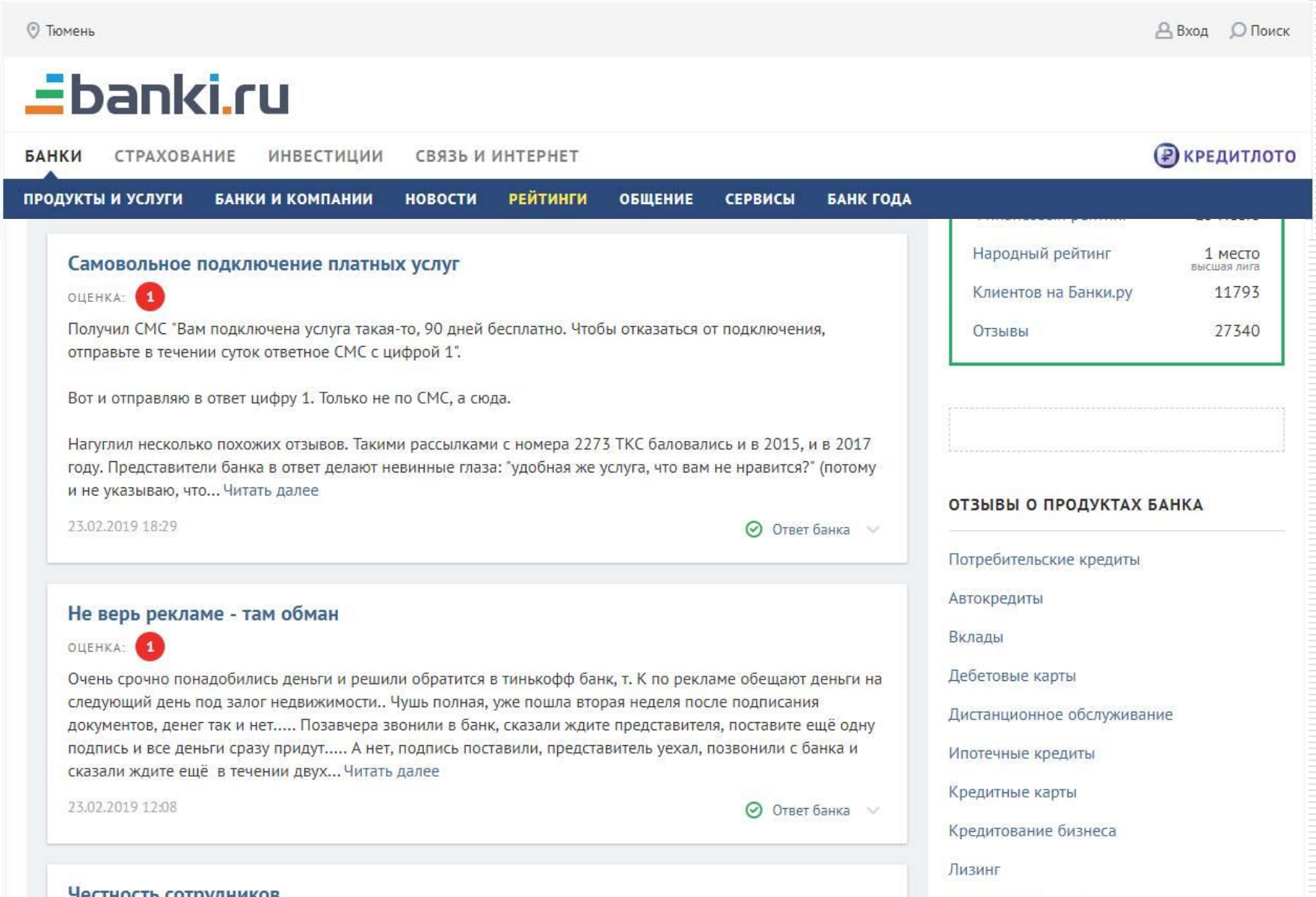
```
<?php
$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, "http://localhost/index1.php");
$fp=fopen("example_homepage.txt","w");
curl_setopt($ch,CURLOPT_RETURNTRANSFER,1);
curl_setopt($ch, CURLOPT_FILE,$fp);
curl_setopt($ch, CURLOPT_HEADER, 0);
curl_exec($ch);
curl_close($ch);
fclose($fp);
?>
```

```
curl_multi_init();
curl_multi_exec($mh, $run)
curl_multi_info_read($mh);
curl_multi_getcontent($ch)
curl_getinfo($ch,CURLINFO_EFFECTIVE_URL)
```

Многопоточные запросы

```
<?php
   $aURLs = array("http://www.php.net","http://www.w3cschools.com");
   $mh = curl_multi_init();
   $aCurlHandles = array();
      foreach ($aURLs as $id=>$url)              {
       $ch = curl_init();
      curl_setopt($ch, CURLOPT_URL, $url);
      curl_setopt($ch, CURLOPT_RETURNTRANSFER,1);
      curl_setopt($ch, CURLOPT_HEADER, 0);
      $aCurlHandles[$url] = $ch;
      curl_multi_add_handle($mh,$ch);    }
   $active = null;
   do {       $mrc = curl_multi_exec($mh, $active);    }
   while ($mrc == CURLM_CALL_MULTI_PERFORM);
   while ($active && $mrc == CURLM_OK) {
      if (curl_multi_select($mh) != -1) {
         do {     $mrc = curl_multi_exec($mh, $active);  }
            while ($mrc == CURLM_CALL_MULTI_PERFORM);       }}
   foreach ($aCurlHandles as $url=>$ch) {
      $html = curl_multi_getcontent($ch);
            …
      curl_multi_remove_handle($mh, $ch);     }
   curl_multi_close($mh);
?>
```

2. Simple HTML DOM Parser
Пример: отзывы о банках



Код страницы

```
<article class="responses__item" data-state="collapsed" data-id="10222008" data-test="responses-item">
        <div>
            <a class="header-h3" data-test="responses-header" href="/services/responses/bank/response/10222008/">
                        Обман в информации о доходе по вкладу в ИБ
            </a>
        </div>

        <div class="responses__item__message" data-preview data-test="responses-message">
                        Банк стал слишком безответственным за свои косяки в сервисах дистанционного
                        обслуживания. Для дистанционного банка это недопустимо.<br><br>В интернет банке в
                        информации на вкладки "События" по действующему вкладу в полях "Прогноз дохода" и
                        "Прогноз баланса" завышены цифры для моего вклада на 6 мес. Там отображаются прогнозы,
                        которые сошлись бы, если бы вклад был открыт на 7 мес.<br><br>На вкладке прогноз
                        дохода такая же ошибка. Там же на графике...
                        <span data-click="expand" class="pseudo-link">Читать далее</span>
        </div>
        <div class="responses__item__message markup-inside-small markup-inside-small--bullet" data-full>
                        Банк стал слишком безответственным за свои косяки в сервисах дистанционного
                        обслуживания. Для дистанционного банка это недопустимо.<br/><br/>В интернет банке в
                        информации на вкладки "События" по действующему вкладу в полях "Прогноз дохода" и
                        "Прогноз баланса" завышены цифры для моего вклада на 6 мес. Там отображаются прогнозы,
                        которые сошлись бы, если бы вклад был открыт на 7 мес.<br/><br/>На вкладке прогноз
                        дохода такая же ошибка. Там же на графике при наведении на "Дек" во всплывающем окне
                        "Баланс на 23 декабря" не соответствует реальному балансу вклада.<br/><br/>Вклад очень
                        простой, открыт в одной валюте, никаких пополнений. Вероятно банк экономит при найме
                        профессионалов для разработки своих сервисов.<br/><br/>Оценку сниму только, если банк
                        на дату окончания вклада выплатит именно ту сумму, которая отображается в ИБ.
        </div>
        <footer class="clearfix">
                <time class="display-inline-block" data-test="responses-datetime">23.12.2018 15:01</time>
        </footer>
</article>
<article class="responses__item" data-state="collapsed" data-id="10222009" data-test="responses-item">
        <div>
            <a class="header-h3" data-test="responses-header" href="/services/responses/bank/response/10222009/">
                        Не верь рекламе - там обман
            </a>
        </div>

        <div class="responses__item__message" data-preview data-test="responses-message">
                        Очень срочно понадобились деньги и решили обратится в тинькофф банк, т. К по рекламе
                        обещают деньги на следующий день под залог недвижимости.. Чушь полная, уже пошла
```

Значимые элементы кода

```
<article class="responses__item" data-state="collapsed"...>
<a class="header-h3" data-test="responses-header"
href="/services/responses/bank/response/10239314/">
ЗАГОЛОВОК
</a>

<div class="responses__item__message" data-preview="" data-
test="responses-message">
ТЕКСТ ОТЗЫВА
</div>

<div class="responses__item__message markup-inside-small
markup-inside-small--bullet" data-full="">
ПОЛНЫЙ ТЕКСТ ОТЗЫВА
</div>
</article>
```

Код парсера

```
<?php
require_once 'library/simple_html_dom.php';
$html = new simple_html_dom();
$html = file_get_html('https://www.banki.ru/?rate=1');
foreach($html->find('article.responses__item') as $article) {

   $header = $article->find('a.header-h3', 0)->plaintext;
   $item = $article->find('div.markup-inside-small', 0)->plaintext;

    print("<h3>".$header."</h3>");
    print("<p>".$item);
}
$html->clear();
unset($html);
?>
```

Документация:

http://simplehtmldom.sourceforge.net/manual.htm

Вывод в браузер

Не связывайтесь с Тинькофф - опозоритесь перед своими клиентами

Хочу предупредить всех кто занимается открытием р/с для своих клиентов сотрудничая с Тинькоф: на встречи приезжают безграмотные студенты которые не могут провести грамотное подписание документов по открытию р/с, приходится повторно дергать людей переподписывая документы, не привозят заказанные(!) к р/с корпоративные карты в результате чего люди не могут пользоваться счетом и вносить деньги. В службе поддержки ФИО сотрудников не дают! Т.е. за их безграмотную работу ответственности никто не несет! Не теряйте время,деньги и доверие людей изза таких банков.

Самовольное подключение платных услуг

Получил СМС "Вам подключена услуга такая-то, 90 дней бесплатно. Чтобы отказаться от подключения, отправьте в течении суток ответное СМС с цифрой 1". Вот и отправляю в ответ цифру 1. Только не по СМС, а сюда. Нагуглил несколько похожих отзывов. Такими рассылками с номера 2273 ТКС баловались и в 2015, и в 2017 году. Представители банка в ответ делают невинные глаза: "удобная же услуга, что вам не нравится?" (потому и не указываю, что за услуга - это неважно). И народ в каментах пишет "отключить можно за 30 секунд - сложно, что ли?". Услуга, может, и удобная. Отключить, может, и несложно. Возмущает-то совсем другое: подключение без спроса. Я не хочу что-то делать для отключения. Я хочу, чтобы меня СНАЧАЛА спросили, надо оно мне или нет. Это низкопробный обман в расчёте на то, что я забуду отправить СМС для отключения. Или просто не стану отвечать, т. к. СМС очень похоже на мошенническое. Все признаки, которые описаны в тинькофф-журнале про мошеннические СМС. Да, так делают почти все - и ОпСоСы, и банки. Но это не оправдание.

Не верь рекламе - там обман

Очень срочно понадобились деньги и решили обратится в тинькофф банк, т. К по рекламе обещают деньги на следующий день под залог недвижимости.. Чушь полная, уже пошла вторая неделя после подписания документов, денег так и нет..... Позавчера звонили в банк, сказали ждите представителя, поставите ещё одну подпись и все деньги сразу придут..... А нет, подпись поставили, представитель уехал, позвонили с банка и сказали ждите ещё в течении двух рабочих дней придут... Сегодня суббота, т. Е во вторник как бы наверно придут и это будет ТРИ НЕДЕЛИ... Зачем вы врете в рекламе и вводите людей в заблуждение...

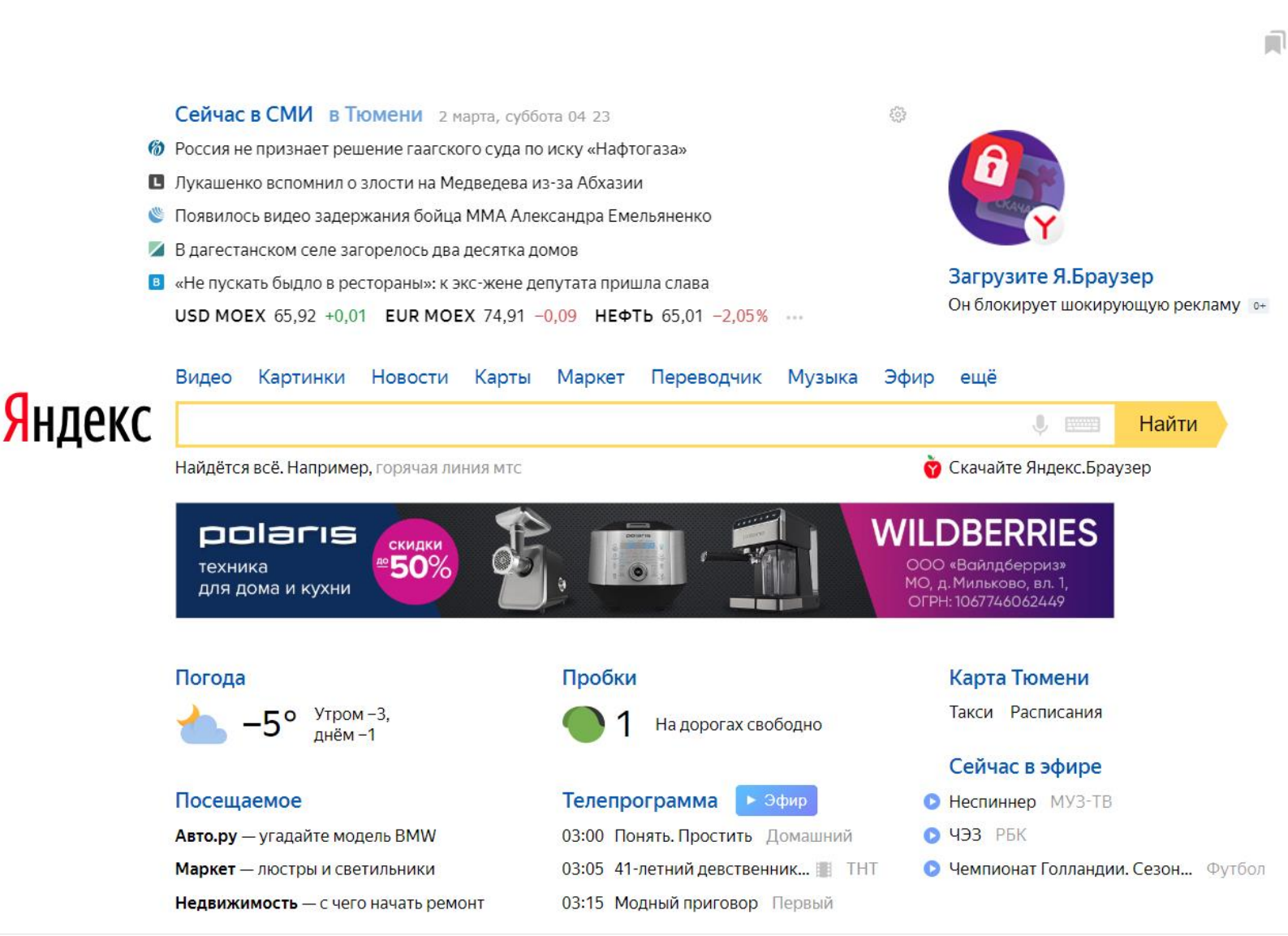
Честность сотрудников

В декабре 2018 обратился в кц тинькофф для отключение смс уведомления по карте, специалист убедил меня что можно подключить бесплатную версию на три месяца, я согласился. Проходит месяц смотрю выписку и вижу списание за смс информацию которой даже не получал, позвонил, сказали что это мол за прошлый месяц думаю ладно может быть со следующего, но нет я вновь увидел это списание то есть из этого исходит что специалисты наврали мне два раза. И второе, при оформлении карты мне выдали договор в котором ни слова о штраф за овердрафт со стороны банка и о чудо при списании комиссии за пользование кредитными деньгами возник овердрафт на сумму около ста рублей и штраф за него 390р. Уважаемый тинькофф банк, я считал вас надёжным банком с известным именем, но такими манипуляциями вы сами себя окунаете в грязь так что вынужден попрощаться с вами.

Тинькофф-Бокс и первое впечатление

Хочу поделится своими впечатлениями работы с Банком Тинькофф! Обслуживаюсь в банке с начала 2016 года как физ. лицо. С середины февраля 2019 года как ИП, открыл расчетный счет (радует то, что открыли за 1 час). Повелся на акцию Тинькофф-Бокс. При открытии счета а это 16.02.2019 Бокс не привезли. Неделю ждал пока рассмотрят

2. PHPQuery
Пример: новости



Код страницы

```
<div class="content-tabs__items content-tabs__items_active_true" role="tabpanel" id=
"tabnews_newsc" aria-labelledby="tabnews_news" aria-hidden="false">
<ol class="list news__list">
<li class="list__item  list__item_icon">
<a class="home-link list__item-content home-link_black_yes" aria-label="Россия не признает
решение гаагского суда по иску «Нафтогаза»" href=
"https://news.yandex.ru/story/Rossiya_ne_priznaet_reshenie_gaagskogo_suda_po_isku_Naftogaza
data-statlog="news.news.links.1" data-statlog-showed="1">
<div class="news__agency-icon news__agency-icon-kommersant news__agency-icon_desktop" title=
"Коммерсантъ"></div>
Россия не признает решение гаагского суда по иску «Нафтогаза»</a></li>

<li class="list__item  list__item_icon"><a class="home-link list__item-content
home-link_black_yes" aria-label="Лукашенко вспомнил о злости на Медведева из-за Абхазии" href=
"https://news.yandex.ru/story/Lukashenko_vspomnil_o_zlosti_na_Medvedeva_iz-za_Abkhazii"
data-statlog="news.news.links.2" data-statlog-showed="1"><div class="news__agency-icon
news__agency-icon-lenta news__agency-icon_desktop" title="Lenta.ru"></div>Лукашенко вспомнил о
злости на Медведева из-за Абхазии</a></li>

<li class="list__item  list__item_icon"><a class="home-link list__item-content
home-link_black_yes" aria-label="Появилось видео задержания бойца ММА Александра Емельяненко" href
="https://news.yandex.ru/story/Poyavilos_video_zaderzhaniya_bojca_MMA_Aleksandra_Emelyanenko"
data-statlog="news.news.links.3" data-statlog-showed="1">

<div class="news__agency-icon news__agency-icon-ria news__agency-icon_desktop" title="РИА Новости"
></div>Появилось видео задержания бойца ММА Александра Емельяненко</a></li>
```

Значимые элементы кода

<ol class="list news__list">

<li class="list__item list__item_icon">

<a

class="home-link list__item-content home-link_black_yes" aria-label="Россия не признает решение гаагского суда по иску «Нафтогаза»"

href="https://news.yandex.ru/story/Rossiya_ne_priznaet data-statlog="news.news.links.1" data-statlog-showed="1">

Россия не признает решение гаагского суда по иску «Нафтогаза»

...

https://code.google.com/archive/p/phpquery/downloads

```
<?php
require ('library/phpQuery.php');
$html = file_get_contents("https://yandex.ru/");
phpQuery::newDocument($html);

$links = pq("#tabnews_newsc")->find("a");
$tmp = array();
foreach($links as $link){
$link = pq($link);
$tmp[] = array(
        "text" => $link->text(),
        "url"  => $link->attr("href")
);       }
phpQuery::unloadDocuments(); ?>

<ul><?php foreach($tmp as $value): ?>
<li> <a href="<?php    echo($value["url"]); ?>" target="_blank">
            <?php   echo($value["text"]); ?>
      </a>
</li>
<?php endforeach; ?></ul>
```

- Россия не признает решение гаагского суда по иску «Нафтогаза»
- Лукашенко вспомнил о злости на Медведева из-за Абхазии
- Появилось видео задержания бойца ММА Александра Емельяненко
- В дагестанском селе загорелось два десятка домов
- Представитель Полины Юмашевой подтвердил ее развод с Олегом Дерипаской

Результаты парсинга: сохранение

1. Выгрузка в базу данных
2. Сохранение в файл в формате:
   - CSV
   - JSON
   - XML

```php
<?php
$list = array (
  array('aaa', 'bbb', 'ccc', 'ddd'),
  array('123', '456', '789'),
    );
$fp = fopen('file.csv', 'w');
foreach ($list as $fields) {
    fputcsv($fp, $fields);
}
fclose($fp);
?>
```

publish_date	headline_text
20030220	в Москве планируется создать нейросеть для считывания показаний счётчиков
20030220	за переводы в Facebook теперь полностью отвечает искусственный интеллект
20030220	нейронные сети Google упростят создание приложений с поддержкой распознавания объектов
20030220	носимое устройство на базе ИИ отличает человека от машины по голосу
20030220	oracle внедрила искусственный интеллект в свои облачные сервисы
20030220	найдены области мозга которые распознают летящий в лицо предмет
20030220	для распознавания лиц импульсная нейронная сеть лучше чем свёрточная
20030220	медики обнаружили новое полезное свойство чтения
20030220	знаменитый ученый стивен хокинг рассказал в интервью о будущем человечества на марсе
20030220	колонизация марса произойдет в течение ближайших ста лет
20030220	стивен хокинг уверен что марс будет колонизирован людьми в ближайшее столетие
20030220	марс наиболее похожая на землю планета солнечной системы
20030220	полет человека на марс с помощью пилотируемого космического корабля
20030220	уже в ближайшее столетие люди заселят марс
20030220	стивен хокинг назвал микроскопические черные дыры источником неограниченной энергии
20030220	популярный британский ученый рассказал что будет с человечеством
20030221	стивен хокинг высказался против выхода великобритании из ес
20030221	роскосмос продолжит работу над рядом ключевых проектов лунной программы
20030221	специалисты разработали специального робота-строителя
20030221	учёные сша сделали первый шаг на пути к колонизации марса и луны
20030221	американцы создали робота-строителя для колонизации марса и луны.
20030221	астроном владимир сурдин усомнился в возможной колонизации марса в ближайшие годы
20030221	первый опыт колонизации марса закончился условной гибелью людей
20030221	в Москве планируется создать нейросеть для считывания показаний счётчиков
20030221	за переводы в Facebook теперь полностью отвечает искусственный интеллект
20030221	нейронные сети Google упростят создание приложений с поддержкой распознавания объектов
20030221	носимое устройство на базе ИИ отличает человека от машины по голосу
20030221	oracle внедрила искусственный интеллект в свои облачные сервисы
20030221	найдены области мозга которые распознают летящий в лицо предмет
20030221	для распознавания лиц импульсная нейронная сеть лучше чем свёрточная
20030221	медики обнаружили новое полезное свойство чтения
20030221	знаменитый ученый стивен хокинг рассказал в интервью о будущем человечества на марсе