

POLITECHNIKA WROCŁAWSKA

PROJEKT

PROJEKTOWANIE SYSTEMÓW Z DOSTĘPEM W JĘZYKU NATURALNYM

**Opracowanie aplikacji wspomagającej
tworzenie korpusu z tekstami w języku
polskim.**

Authors:

Rafał PIENIAŻEK
Jakub POMYKAŁA

Supervisor:

Dr inż. Dariusz BANASIAK

18 grudnia 2017

Spis treści

1	Dodatek do opisu parametrów	2
1.1	Tryb TXT	2
1.2	Tryb JSON	2
2	Dodatek do implementacji parserów	3
2.1	Ekstrakcja tekstu z artykułu	3
2.2	Ekstrakcja autora	5
2.3	Ekstrakcja adresów URL	6
3	Dodatek do wniosków	7

1 Dodatek do opisu parametrów

1.1 Tryb TXT

Tryb TXT zawiera jedynie treść samego artykułu bez dodatkowych metadanych.

1.2 Tryb JSON

W trybie JSON zostanie zapisana cała klasa *Article* przedstawiona poniżej.

```
public class Article {  
    private String source;  
    private String title;  
    private String body;  
    private HashMap<String, String> metadata;  
    //...  
}
```

Przykład wynikowego pliku *json*.

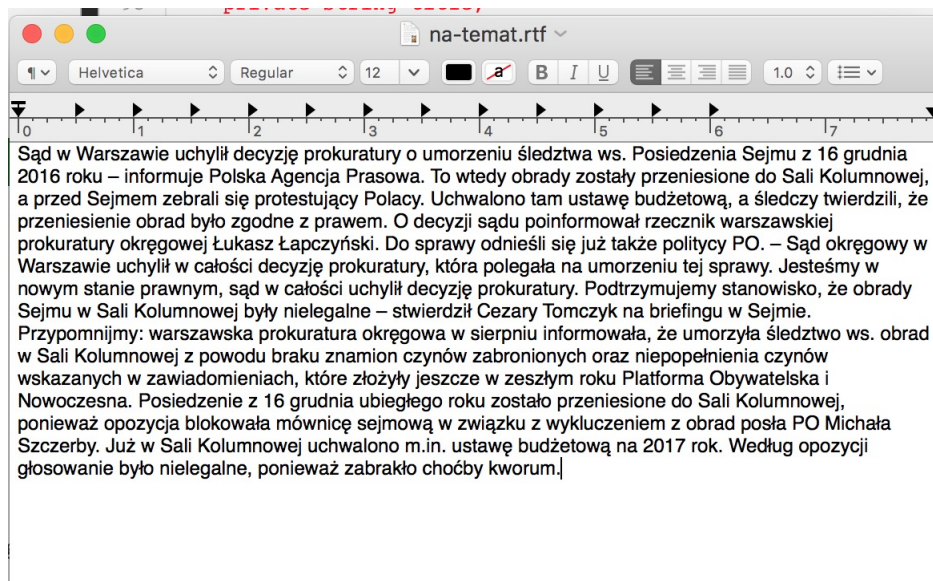
```
{  
  "source" : "https://niebezpiecznik.pl/post/zatrzymania-w-sprawie-wycieku  
-z-bazy-pesel-komornik-naduzywal-uprawnien-na-rzecz-windykatora/",  
  "title" : "Komornik z windykatorem naduzywali dostepu do bazy PESEL.  
Prokuratura i ABW zamknely kramik",  
  "body" : "Cztery osoby zatrzymane (...)",  
  "metadata" : {  
    "date" : "7:58 14/12/2017",  
    "author" : "Marcin Maj"  
  }  
}
```

2 Dodatek do implementacji parserów

2.1 Ekstrakcja tekstu z artykułu

```
<div class="art_header_photo"></div>
<div class="visible-xs clearfix"></div>
<span class="art_hfs">5</span>
"ąd w Warszawie uchylił decyzję prokuratury o umorzeniu śledztwa ws. Posiedzenia Sejmu z 16 grudnia 2016 roku – informuje Polska Agencja Prasowa. To wtedy obrady zostały przeniesione do Sali Kolumnowej, a przed Sejmem zebrał się protestujący Polacy. Uchwalono tam ustawę budżetową, a śledczy twierdzili, że przeniesienie obrad było zgodne z prawem."
<br>
<div class="ebNative"></div>
<!-- adslot: ("description":"Artykuł0142y natemat.pl after_lead") :adslot -->
<div class="hidden-xs adform-slot" data-mid="583954" data-lazy-ignore="1">
<script type="text/javascript"></script>
<script type="text/javascript" data-adfscript="adx.adform.net/adx/7mid=583954&rnd=<random_number>"></script>
</div>
"O decyzji sądu poinformował rzecznik warszawskiej prokuratury okręgowej Łukasz Łapczyński. Do sprawy odnieśli się już także politycy PO. – Sąd okręgowy w Warszawie uchylił w całości decyzję prokuratury, która polegała na umorzeniu tej sprawy. Jesteśmy w nowym stanie prawnym, sąd w całości uchylił decyzję prokuratury. Podtrzymujemy stanowisko, że obrady Sejmu w Sali Kolumnowej były nielegalne – stwierdził Cezary Tomczyk na briefingu w Sejmie."
<a href="http://natemat.pl/t/801,sejm">Sejmie</a>
"
<blockquote class="block_pos_inside"></blockquote>
<a href="http://natemat.pl/214409,pls-owska-prokuratura-kochana-michal-szczerba-kpl-z-umorzenia-sledztwa-w-sprawie-uchwalenia-ustawy-budzetowej">Przypomnijmy</a>
" warszawska prokuratura okręgowa w sierpniu informowała, że umorzyła śledztwo ws. obrad w Sali Kolumnowej z powodu braku znamion czynów zabronionych oraz niepopelnienia czynów wskazanych w zawiadomieniach, które zostały jeszcze w zeszłym roku Platforma Obywatelska i Nowoczesna."
<blockquote class="block_pos_inside"></blockquote>
"Posiedzenie z 16 grudnia ubiegłego roku zostało przeniesione do Sali Kolumnowej, ponieważ opozycja blokowała mównicę sejmową w związku z wykluczeniem z obrad posła PO Michała Szczerby."
<a href="http://natemat.pl/199229,co-byl-najwiekszy-blug-wyjasnilo-sie-dlaczego-kuchcinski-wykluczyl-szczerbe-i-od-czego-zaczal-sie-kryzys-w-sejmie">z wykluczeniem z obrad posła PO Michała Szczerby</a>
" Już w Sali Kolumnowej uchwalono m.in. ustawę budżetową na 2017 rok. Według opozycji głosowanie było nielegalne, ponieważ zabrakło choćby kworum."
<br>
<br>
<script type="text/javascript"></script>
<script type="text/javascript" src="http://bbcdn-bbnaut.ibillboard.com/library/bbnaut-lib-1.7.5.min.js"></script>
<div class="onnetwork-video"></div>
<script type="text/javascript" src="https://mrla.exs.pl/playermin.php"></script>
<script type="text/javascript"></script>
<br>
<em></em>
<div class="post-like-us"></div>
</div>
</section>
<section class="art_right-column"></section>
</div> == $0
```

Wszystkie akapity tekstu zostały połączone w jeden tekst



W przypadku strony <https://niebezpiecznik.pl> metoda *parseLink* przyjmuje URL artykułu i na jego podstawie dokonuje ekstrakcji poszczególnych metadanych.

```
private Article parseLink(String articleUrl) {
    try {
        Article article = new Article(articleUrl);
        Document doc = JsoupConnector.connectThrowable(articleUrl, SLEEP_TIME);
        article.setTitle(doc.select(".postcontent").select("h1").text());
        article.getMetadata().put("author", getAuthor(doc));
        article.getMetadata().put("keywords", getKeywords(doc));
        article.getMetadata().put("date", getDate(doc));
        article.setBody(getBody(doc));

        return article;
    } catch (Exception e) {
        return null;
    }
}
```

Usuwane są wszystkie teksty które znajdują się w tagach *blockquote* wskazujące na obecność cytatów w tekście.

```
private String getBody(Document doc) {
    doc.select("blockquote").remove();
    return doc.select(".postcontent").select("p").text();
}
```

getDate przy pomocy wyrażenia regularnego wyluskuje datę napisania artykułu.

```
private String getDate(Document doc) {
    String postmeta = doc.select(".dolna-ramka").select("span").text();
    Pattern pattern = Pattern.compile("(.)+ w kategorii");
    Matcher matcher = pattern.matcher(postmeta);
    if (matcher.find()) {
        return matcher.group(1);
    }
    return null;
}
```

W przypadku strony <http://niebezpiecznik.pl> autor jest podany wprost w tagu *a* z ustawionym atrybutem *rel=author*.

```
private String getAuthor(Document doc) {
    return doc.select("a[rel=author]").text();
}
```

2.2 Ekstrakcja autora

W kodzie HTML poniżej znajduje się informacja o imieniu i nazwisku autora artykułu ze strony natemat.pl.

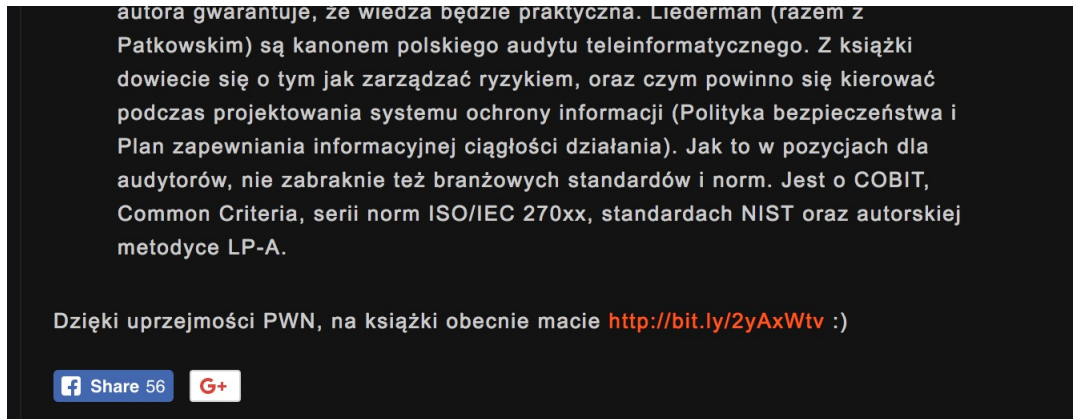
```
▼<div class="art__author__container clearfix">
  ▼<div class="art__author__avatar">
    ▶<a href="http://natemat.pl/u/3009,piotr-rodzik">...</a>
  </div>
  ▼<div class="art__author__description">
    ▼<div class="art__author__name">
      ▼<a href="http://natemat.pl/u/3009,piotr-rodzik" itemprop="url">
        <span itemprop="name">Piotr Rodzik</span> == $0
      </a>
    </div>
  </div>
  ▼<div class="art__date">
    <span class="date" title="2017-12-18T13:01:55+01:00">18 grudnia 2017</span>
  </div>
  ▼<div class="post-left-ad">...</div>
</div>
</section>
</div>
<span class="art__hfs">S</span>
```

Poniższy kod z pobranego dokumentu strony wyodrębnia imię i nazwisko autora.

```
private String getAuthor(Document doc) {
    Element authorElement = doc.select(".art__author__name").first();
    String author = authorElement.text().trim();
    if (author.contains("Partnerem")) {
        return null;
    }
    authorElement.remove();
    return author;
}
```

2.3 Ekstrakcja adresów URL

Na poniższym rysunku znajduje się przykład adresu URL, który zostanie usunięty.



Poniżej znajduje się fragmentu kodu źródłowego strony.

```
<div class="entry">
  <p>...</p>
  <div class="info">...</div>
  <p>A teraz wracamy do nowości w PWN:</p>
  <ul>...</ul>
  <p>
    "Dzięki uprzejmości PWN, na książki obecnie macie "
    <a href="http://bit.ly/2yAxWtv">http://bit.ly/2yAxWtv</a> == $0
    " :)"
  </p>
  <div class="fb-share-button fb_iframe_widget" data-href="https://niebezpiecznik.pl/post/2-ciekawe-ksiazki-z-oferty-pwn/" data-layout=
  "button_count" style="float: left; margin-right: 10px;" fb-xfbml-state="rendered" fb-iframe-plugin-query=
  "app_id=&container_width=0&href=https%3A%2F%2Fniebezpiecznik.pl%2Fpost%2F2-ciekawe-ksiazki-z-oferty-
  pwn%2F&layout=button_count&locale=en_US&sdk=joey">...</div>
  <p>...</p>
  <h4>Przeczytaj także:</h4>
  <ul class="similar-posts">...</ul>
  <p></p>
```

Funkcja *removeLinks* przyjmuje jako parametr tekst z którego będą usuwane linki URL.

```
private String removeLinks(String text) {
    if (text == null) {
        throw new IllegalArgumentException("Article body cannot be null");
    }
    return text.toLowerCase().replace("http.?:/\\/\\S+", "").trim();
}
```

Zmienna *String page* zawiera cały tekst artykułu z którego zostaną usunięte wszystkie adresy url, które zaczynają się od *http://* oraz *https://*.

Wszystkie wykonywane operacje na pobranych tekstach to:

- Usuwanie linków
- Uswanie cytatów
- Usuwanie kodów źródłowych
- Usuwanie linków do obrazków

3 Dodatek do wniosków

Lista stron z których zostały pobrane informacje:

- <https://niebezpiecznik.pl> - ponad 1 tysiąc artykułów
- <https://zaufanatrzeciastrona.pl> ponad 2 tysiące artykułów
- <https://natemat.pl> - ponad 16 tysięcy artykułów
- <https://www.purepc.pl> ponad 1 tysiąc artykułów