



Specyfikacja NewsFeed

Wstęp

Poniższy dokument opisuje sposób zadawania zapytań oraz odbierania wyników wyszukiwania NewsPoint w formacie XML za pomocą protokołu HTTP.

1 Zapytanie do NewsPoint

Zapytanie HTTP składa się z URL'a bazowego:

<http://xml.newspoint.pl>

oraz parametrów definiujących kryteria wyszukiwania i warunki prezentacji wyników, doklejanych do powyższego adresu.

1.1 Parametry

1.1.1 Parametry wymagane

– **username** {string}

Nazwa użytkownika (login) konta NewsPoint

– **password** {string}

Zahaszowana wersja hasła do konta NewsPoint

1.1.2 Parametry opcjonalne

Każde zapytanie może zawierać jeden i tylko jeden z poniższych parametrów:

• Aby otrzymać uprzednio zdefiniowaną listę wyszukiwania:

– **list_id** {int}

ID Profilu lub listy Wyszukiwania Inteligentnego

• Aby otrzymać artykuły Archiwum:

– **archive_id** {int}

ID Archiwum

• Aby wysłać natychmiastowe zapytanie złożone ze słów kluczowych:

– **word_rule**[] {and*,or,near}

Parametr word_rule określa logiczne relacje między słowami kluczowymi w zapytaniu, podawanymi w parametrze search_srt[] (patrz poniżej)

– **search_str**[] {string}

Łańcuch wyszukiwania zawierający słowa kluczowe. W razie pominięcia tej opcji NewsFeed zwróci najbardziej aktualne, nie filtrowane wiadomości. Aby wysłać zapytanie składające się z kilku wyrażen, należy podać parametr kilka razy, oddzielając kolejne wyrażenia parametrem word_rule.

– **search_scope** {text*,summary,header}

Parametr wskazuje, w których elementach artykułu należy szukać: w tytule, we wprowadzeniu bądź treści artykułu:

text - szukaj we wszystkich elementach artykułu

summary - szukaj we wprowadzeniu i tytule

header - szukaj tylko w tytule

Wyniki wyszukiwania uzyskane poprzez podanie jako kryterium wyszukiwania ID profilu bądź listy Wyszukiwania Inteligentnego czy też opcji wyszukiwania podanych powyżej można dalej zawęzić podając jedno lub więcej spośród wyrażen:

- **limit_word_rule**[] {and,or,near* }
- **limit_search_str**[] {string }

Parametry warunkujące sposób opisu wyników wyszukiwania w dostarczonym XML:

- **identical** {true*,false }

Wydawcy często publikują aktualności pochodzące z tego samego źródła, wówczas identyczne artykuły pojawiają się na wielu różnych witrynach. Jeśli wybierzemy opcję „false” NewsFeed dostarczy wyłącznie unikalne wiadomości oraz listę Witryn, na których się one pojawiły.

Domyślnie NewsFeed ma ustawiony parametr taki sam jak w Profilu. Jeśli NewsFeed nie dotyczy profilu domyślna wartość to „true”.

- **summary** {true*,false }

Zwraca wprowadzenie każdego artykułu

- **body** {true,false* }

Zwraca treść każdego artykułu

- **matches** {true,false* }

Zwraca listę znalezionych w artykule słów z zapytania

- **quotes** {true,false* }

Zwraca kontekst znalezionych słów

- **num_art** {int }

Liczba artykułów zwracanych na każdej stronie

- **maxage** {sec }

Maksymalny wiek zwracanych wiadomości

- **from** {unix_timestamp }

NewsPoint przyporządkowuje każdej wiadomości datę jej znalezienia przez system. Możliwe jest wybranie artykułów znalezionych w określonym przedziale czasu. Pole from określa początkowy czas przedziału w formacie uniksowego znacznika czasu (liczba sekund od 1 stycznia 1970).

- **to** {unix_timestamp }

Data końcowa przedziału czasu

- **context** {string }

Zapytanie do NewsPoint może czasem zwracać bardzo dużą liczbę wyników. Aby zmniejszyć zwracaną liczbę wyników, NewsPoint domyślnie zwraca tylko 10 wyników (wartość tą można zmienić za pomocą parametru `num_art` opisanego powyżej) oraz informacje pozwalające na pobranie dalszych wyników. Taka funkcjonalność jest pomocna przy implementacji stronicowania wyników (Przyciski typu „następna strona”, „poprzednia strona”).

– **header_length** {int}

Maksymalna długość zwracanego tytułu

– **summary_length** {int}

Maksymalna długość zwracanego wprowadzenia

– **source_length** {int}

Maksymalna długość zwracanej nazwy Wydawcy

1.2 Przykład

Poniższe zapytanie zwróci wszystkie artykuły zawierające słowa kluczowe „Bush” i „Iraq”:

http://xml.newspoint.pl/?username=X&password=Y&search_str=Bush%20Iraq

2 Wyniki wyszukiwania

Po otrzymaniu zapytania NewsPoint znajduje wszystkie artykuły zgodne z kryteriami wyszukiwania i zwraca je w formacie XML. Znaczniki XML zawierają szczegółowe informacje na temat każdego artykułu oraz dodatkowe informacje opisujące powód zwrócenia danego artykułu w wynikach konkretnego zapytania.

Poniżej znajduje się opis najważniejszych znaczników XML. Znaczniki nie opisane służą jedynie do użytku wewnętrznego lub są przestarzałe i mogą zostać wkrótce zlikwidowane.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
```

```
<searchresult documents="..." ...>
```

```
<search_matches>
```

```
<search_match type="..." color="...">...</search_match>
```

```
[...]
```

```
</search_matches>
```

```
<document id_site="..." id_article="...">
```

```
<unix_timestamp>...</unix_timestamp>
```

```
<local_time>...</local_time>
```

```
<url>...</url>
```

```
<orig_url>...</orig_url>
```

```
<language encoding="iso-639">...</language>
```

```
<first_source id="...">
```

```
<name>...</name>
```

```
<sitename>...</sitename>
```

```
<url>...</url>
```

```
<siteurl>...</siteurl>
</first_source>
<header matches="...">...</header>
<summary matches="...">...</summary>
```

```
<matches>
<match type="..." color="...">...</match>
[...]
</matches>
```

```
<quotes>
<quote matches="...">...</quote>
[...]
</quotes>
```

```
<colorbar ...>
<item length="..." color="...">[<quote matches="...">[...]</quote>]</item>
[...]
</colorbar>
```

```
<identical_documents>
[<document ...>...</document>]
</identical_documents>
```

```
</document>
```

```
[<document ...>...</document>]
</searchresult>
```

Wyjaśnienie znaczników:

<searchresult>

Główny element dokumentu. Zawiera systemowe informacje nt. wyszukiwania. Atrybut `documents` podaje liczbę zwróconych dokumentów.

<search_matches>

Element zawiera wszystkie dopasowania w wynikach wyszukiwania. Każde dopasowanie zdefiniowane jest w znaczniku `<search_match>`. Dopasowaniami są nie tylko słowa kluczowe ale również języki, kategorie tematyczne i filtry. Zwykle hasła wyszukiwania mają atrybut `type` ustawiony na wartość "word". Każdemu słowu jest przypisany numer koloru (a atrybucie `color`). Numery kolorów słów zaczynają się od 4 (niższe numery są zarezerwowane do innych celów).

<document>

Element reprezentujący artykuł. Atrybuty `id_site` i `id_article` wspólnie jednoznacznie identyfikują artykuł.

<unix_timestamp>

Każdy artykuł otrzymuje znacznik czasu odpowiadający momentowi jego odkrycia przez NewsPoint. Wartość reprezentuje liczbę sekund od 1 stycznia 1970 (zobacz <http://pl.wikipedia.org/wiki/EPOCH>).

<local_time>

Znacznik czasu sformatowany jako data czytelna dla człowieka.

<url>

URL Widoku Strony prezentującego artykuł oraz opcjonalnie podświetlającego w treści poszukiwane słowa.

<orig_url>

Oryginalny URL artykułu w Internecie

<language>

Język artykułu

<first_id_source>

Część artykułów jest zbierana przez system wiele razy, ponieważ linki do nich znajdują się w różnych sekcjach witryny. Element określa sekcję, w której artykuł został zauważony po raz pierwszy.

<name> Nazwa sekcji

<sitename> Witryna, do której należy sekcja

<url> URL sekcji

<siteurl> URL witryny

<header>

Ten element zawiera tytuł artykułu. Jeśli atrybut matches jest ustawiony na true, poszukiwane słowa są w treści umieszczone wewnątrz znaczników <match>. Pozostały tekst znajduje się w znaczniku <text>.

<summary>

Streszczenie artykułu. Poszukiwane słowa w streszczeniu są zaznaczone o ile atrybut matches ma wartość true.

<matches>

Wewnątrz tego elementu są umieszczane wszystkie odnalezione hasła wyszukiwania opisywane za pomocą atrybutu <match>. Dla słów wyszukiwania, atrybut type ma wartość "word".

<quotes>

System zwraca pewien tekst będący kontekstem wystąpienia w artykule słów kluczowych. Każdy kontekst jest zawarty w znaczniku <quote>. Dopasowane słowo znajduje się w znaczniku <match>, zaś reszta tekstu w elementach <text>.

<identical_documents>

Jeśli w zapytaniu podano odpowiednie opcje, artykuły identyczne do danego są zwracane wewnątrz tego znacznika. Każdy identyczny dokument jest zawarty w znaczniku <document> opisanym powyżej.

3 Obsługa wyjątków i błędów

Jeżeli podczas przetwarzania zapytania wystąpił błąd, zwracany jest dokument XML zawierający wyjaśnienie błędu w następującej składni:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
```

```
<EXCEPTION_COLLECTION>
```

```
<EXCEPTION code='x'>error description</EXCEPTION>
```

```
[<EXCEPTION>...</EXCEPTION>]
```

```
</EXCEPTION_COLLECTION>
```