LALS

Lexical Analysis Lemmatization Service

Lematizačný servis pre slovenský jazyk

Príručka pre integráciu do SOLR
(SOLR Integration Guide)

Verzia 1.0

Obsah

📄 LALS – Solr Integration Guide

---

1 Prehľad

LALS poskytuje vlastný tokenizer pre Apache Solr, ktorý umožňuje:

- lematizáciu slovenského textu
- spracovanie HTML dokumentov
- prácu s textom bez diakritiky
- využitie pozícií slov pre highlighting

Tokenizer je distribuovaný ako JAR plugin, ktorý sa integruje priamo do Solr.

👉 Tokenizer komunikuje so službou LALS prostredníctvom REST API.

---

2 Požiadavky

- Apache Solr kompatibilný s použitou verziou pluginu
- bežiaca služba LALS (REST API)

👉 Tento plugin využíva službu LALS na spracovanie textu.
👉 Bez bežiacej služby LALS nie je možné plugin používať.

---

3 Inštalácia pluginu

Plugin je distribuovaný ako JAR súbor, napr.:

solr-tokenizers-9.11.1.jar

👉 verzia JAR musí zodpovedať verzii Lucene/Solr (napr. 9.11.1)

🔹 **Krok 1: Kopírovanie súboru**

Skopírujte JAR súbor do adresára:

solr/server/solr-webapp/webapp/WEB-INF/lib

🔹 **Príklad**

Ak je Solr nainštalovaný v `/opt`, výsledná cesta bude:

/opt/solr/server/solr-webapp/webapp/WEB-INF/lib/solr-tokenizers-9.11.1.jar

---

🔹 **Krok 2: Reštart Solr**

Po pridaní pluginu je potrebné reštartovať Solr:

bin/solr restart

---

4 Konfigurácia schémy

Tokenizer sa používa priamo v definícii `fieldType`.

4.1 Základný text (s diakritikou)

```
<fieldType name="text_sk_la" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
            url="http://localhost:60998/text/"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
            url="http://localhost:60998/text/"/>
  </analyzer>
</fieldType>
```

4.2 Text bez diakritiky

```
<fieldType name="text_sk_unla" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
            url="http://localhost:60998/text/"
            postUnaccent="true"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
            url="http://localhost:60998/text/"
            unaccent="true"/>
  </analyzer>
</fieldType>
```

4.3 HTML dokumenty

```
<fieldType name="html_sk_la" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
            url="http://localhost:60998/html/"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
            url="http://localhost:60998/text/"/>
  </analyzer>
</fieldType>
```

4.4 HTML dokumenty bez diakritiky

```
<fieldType name="html_sk_unla" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
          url="http://localhost:60998/html/"
          postUnaccent="true"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="org.la.solr.tokenizer.LsSolrTokenizerFactory"
          url="http://localhost:60998/text/"
          unaccent="true"/>
  </analyzer>
</fieldType>
```

👉 Táto konfigurácia umožňuje:

- indexovať HTML dokumenty
- zachovať formátovanie
- zároveň vyhľadávať bez ohľadu na diakritiku

4.5 Vysvetlenie parametrov

Parameter	Popis
url	endpoint služby LALS (`/text/` alebo `/html/`)
unaccent	spracovanie vstupu bez diakritiky (typicky pri vyhľadávaní)
postUnaccent	odstránenie diakritiky po lematizácii (typicky pri indexovaní)

---

5 Princíp fungovania

5.1 Indexovanie

- text dokumentu sa spracuje pomocou LALS
- do indexu sa ukladajú:
  - tokeny (slová)
  - lemma tvary
  - pozície slov

5.2 Vyhľadávanie

- dotaz používateľa sa spracuje rovnakým tokenizerom
- zabezpečí sa konzistentné spracovanie

👉 **Výsledok:**

- lepšie pokrytie rôznych tvarov slov

- vyššia úspešnosť vyhľadávania

---

6 HTML a highlighting

Pri použití `/html/` endpointu:

- HTML značky sú ignorované
- zachovávajú sa pozície slov

👉 **To umožňuje:**

- zvýrazňovanie výsledkov priamo v HTML dokumente
- zachovanie formátovania (odstavce, štýly)

👉 **Výsledok:**

- používateľ vidí zvýraznený formátovaný dokument
- nie len plain text

---

7 Príklad použitia (end-to-end scenár)

Táto kapitola demonštruje kompletný priebeh spracovania dokumentu a vyhľadávania pomocou LALS tokenizeru v Apache Solr.

7.1 Indexovanie dokumentu

7.1.1 Vstupný dokument

Do Solr je uložený dokument s HTML obsahom:

id: doc-001

title: Výročná správa

content_html: Výročná správa o činnosti Slovenskej televízie za rok 2010.

7.1.2 Spracovanie počas indexovania

Tokenizer odošle obsah poľa na službu LALS:

POST http://localhost:60998/html/

data= Výročná správa o činnosti Slovenskej televízie za rok 2010.

7.1.3 Odpoveď služby LALS

5 ; 12 ; výročná ; výročný

13 ; 19 ; správa ; správať

20 ; 21 ; o

22 ; 30 ; činnosti ; činnosť

31 ; 41 ; slovenskej ; slovenský

42 ; 51 ; televízie ; televízia

52 ; 54 ; za

55 ; 58 ; rok

59 ; 63 ; 2010

7.1.4 Interpretácia

Pre každé slovo služba vracia:

- pôvodný tvar slova
- jeho lemma tvar (alebo viacero lemma tvarov)

👉 Tokenizer tieto informácie využije tak, že:

- pre každý výskyt slova v texte vytvorí jeden logický token
- k tomuto tokenu priradí všetky jeho relevantné tvary (pôvodný + lemma)

7.1.5 Príklad

Pre slovo:
výročná

sa do indexu uloží:
výročná, výročný

👉 To znamená, že:

- ide o jeden výskyt slova v texte
- ale je možné ho nájsť pomocou viacerých tvarov

7.1.6 Dôležité vysvetlenie

👉 Nejde o nezávislé nesúvisiace termy.
👉 Ide o rôzne jazykové tvary toho istého slova, ktoré sú viazané na rovnakú pozíciu v texte.

7.2 Vyhľadávanie

Používateľ zadá dotaz:
televízia

7.2.1 Spracovanie dotazu

Dotaz je spracovaný rovnakým tokenizerom:
POST http://localhost:60998/text/
data=televízia

7.2.2 Výsledok analýzy

0 ; 9 ; televízia

7.3 Vyhodnotenie dotazu

V dokumente sa nachádza tvar:
televízie

Počas indexovania bol však pre tento výskyt uložený aj lemma tvar:
televízia

👉 Výsledok:

- dotaz „televízia“ sa úspešne zhoduje s textom „televízie“
- systém správne nájde dokument aj pri rôznych tvaroch slova

7.4 Highlighting

Keďže:

- dokument je uložený vo formáte HTML
- LALS zachováva pozície slov

je možné použiť highlighting priamo vo formátovanom texte.

👉 Výsledkom je výrazne vyššia kvalita vyhľadávania slovenského textu bez potreby zložitej logiky na strane aplikácie.

7.4.1 Príklad výsledku

Výročná správa o činnosti Slovenskej televízie za rok 2010.

7.5 Výsledok

Použitím LALS tokenizeru:

- rôzne tvary slovenských slov sú správne spárované
- vyhľadávanie je robustnejšie
- zvýrazňovanie funguje priamo vo formátovanom dokumente

7.6 Zhrnutie scenára

1. HTML dokument je spracovaný cez LALS (/html/)
2. pre každé slovo sa získajú jeho lemma tvary
3. tieto tvary sú uložené do indexu ako alternatívy pre daný výskyt slova
4. dotaz používateľa je spracovaný rovnakým spôsobom (/text/)
5. Solr nájde zhodu aj medzi rôznymi tvarmi slova
6. výsledok je zvýraznený priamo v HTML dokumente

---

8 Odporúčania

- používať `/html/` pre dokumenty (PDF/Word konvertované do HTML)
- používať `unaccent` pre používateľské dotazy bez diakritiky
- používať `postUnaccent` pri indexovaní, ak chceme ignorovať diakritiku

---

9 Najčastejšie chyby

- nesprávna verzia pluginu (nekompatibilná s Lucene)
- zabudnutý reštart Solr
- nebežiaca služba LALS
- nesprávny endpoint (`/html/` vs `/text/`)

---

10 Zhrnutie

LALS tokenizer pre Apache Solr umožňuje:

- jednoduchú integráciu lematizácie do Solr
- spracovanie slovenského textu bez dodatočnej logiky
- podporu HTML dokumentov a highlightingu
- flexibilnú prácu s diakritikou

👉 **Integrácia spočíva v:**

- skopírovaní JAR súboru
- konfigurácii schémy
- spustení služby LALS