LALS

Lexical Analysis Lemmatization Service

Lematizačný servis pre slovenský jazyk

Technický prehľad
(Technical Overview)

Verzia 1.0

Obsah

1 Prehľad ... 3
2 Princíp fungovania ... 3
2.1 Viacznačné slová ... 3
2.2 Význam pre vyhľadávanie ... 3
3 API prehľad ... 4
3.1 Endpointy ... 4
3.1.1 Parametre ... 4
3.2 Endpoint /text/ ... 4
3.2.1 Príklad request ... 4
3.2.2 Príklad response ... 4
3.2.3 Formát výstupu ... 4
3.3 Endpoint /html/ ... 5
3.3.1 Vlastnosti ... 5
3.3.2 Príklad ... 5
3.3.3 Využitie ... 5
3.4 Spracovanie diakritiky ... 5
3.4.1 unaccent=true ... 5
3.4.2 postunaccent=true ... 5
4 Výkon ... 6
4.1 Benchmark prostredie ... 6
4.2 Typický výkon ... 6
5 Nasadenie ... 6
5.1 Požiadavky ... 6
6 Integrácia ... 7
6.1 Typické použitie ... 7
6.2 Integrácia ... 7
6.3 Obmedzenia ... 7
7 Rozšírenie: inteligentné rozšírenie dotazu (Query Expansion) ... 8
7.1 Princíp ... 8
7.2 Endpoint ... 8
7.2.1 Formát vstupu ... 8
7.2.2 Formát výstupu ... 8
7.2.3 Príklady použitia ... 8
Príklad (významové rozšírenie pre slovo „slovensko“) ... 8
Príklad (morfologické rozšírenie pre slovo „test“) ... 8
Príklad (sloveso „testovať“) ... 9
7.3 Podpora vstupu bez diakritiky ... 10
7.3.1 Príklad (slovo „stat“) ... 10
7.4 Využitie ... 10
7.5 Podporované typy rozšírenia: ... 10
7.6 Poznámky ... 11
8 Zhrnutie ... 12
8.1 Prečo LALS namiesto stemmingu? ... 12
9 Ďalšie informácie ... 12

LALS – Technický prehľad

1 Prehľad

LALS (Lexical Analysis Lemmatization Service) je vysokovýkonná REST služba určená na spracovanie slovenského textu.

Rieši problém slovenskej morfológie bez potreby zložitého NLP spracovania.

Služba prevádza vstupný text na základné tvary slov (lemma) a pri nejednoznačných slovách vracia viacero možných lemma tvarov.

Je vhodná pre:

- fulltextové vyhľadávanie
- dokumentové systémy (DMS)
- analýzu textu
- NLP pipeline

Služba je optimalizovaná pre produkčné nasadenie s dôrazom na:

- nízku latenciu
- vysokú priepustnosť
- jednoduchú integráciu

---

2 Princíp fungovania

LALS spracuje vstupný text a pre každé slovo:

- identifikuje možné lemma tvary
- využíva slovník (~3 milióny slovných tvarov)
- pre neznáme slová používa algoritmické spracovanie

2.1 Viacznačné slová

Pri nejednoznačných slovách služba vracia všetky relevantné lemma kandidáty.

Príklad:

```
miest → miesto, mesto
```

Slovo „miest“ môže byť:

- genitív množného čísla slova „miesto“ (priestor)
- genitív množného čísla slova „mesto“ (osídlenie)

Preto služba vracia obe možné lemy.

2.2 Význam pre vyhľadávanie

Tento prístup:

- znižuje riziko, že relevantný dokument nebude nájdený (false negatives)
- môže mierne zvýšiť počet menej relevantných výsledkov (false positives)

Pri enterprise vyhľadávacích systémoch je zvyčajne väčší problém nenájsť relevantný dokument, než zobraziť niekoľko menej relevantných výsledkov navyše.

👉 Tento kompromis je vhodný pre väčšinu vyhľadávacích systémov.

3 API prehľad

3.1 Endpointy

```
POST /text/
POST /html/
```

POST /text/ - endpoint pre spracovanie voľného textu.

POST /html/ - endpoint pre spracovanie textu v HTML.

3.1.1 Parametre

Parameter	Typ	Popis
data	text	vstupný text alebo HTML
unaccent	boolean	spracovanie s ignorovaním diakritiky
postunaccent	boolean	spracovanie textu s diakritikou + odstránenie diakritiky po spracovaní pre podporu vyhľadávania s ignorovaním diakritiky
synonyms	boolean	rozšírenie o synonymá

3.2 Endpoint /text/

3.2.1 Príklad request

```
POST /text/
data=Výročná správa o činnosti Slovenskej televízie
```

3.2.2 Príklad response

```
0 ; 7 ; výročná ; výročný
8 ; 14 ; správa ; správať
15 ; 16 ; o
17 ; 25 ; činnosti ; činnosť
26 ; 36 ; slovenskej ; slovenský
37 ; 46 ; televízie ; televízia
```

3.2.3 Formát výstupu

Každý riadok obsahuje:

1. začiatok slova v texte
2. koniec slova v texte
3. pôvodné slovo
4. lemma tvary (ak sa líšia od pôvodného slova)

👉 Ak je lemma rovnaké ako pôvodné slovo, nie je opakovane uvedené.

3.3 Endpoint /html/

LALS podporuje spracovanie jednoduchého HTML bez potreby jeho predchádzajúceho čistenia.

3.3.1 Vlastnosti

- HTML značky sú ignorované
- spracováva sa iba textový obsah
- zachovávajú sa pozície slov voči pôvodnému vstupu

3.3.2 Príklad

Vstup:

```
<div>včera bolo ráno</div>
```

Výstup:

```
5 ; 10 ; včera
16 ; 20 ; bolo ; byť
21 ; 25 ; ráno
```

3.3.3 Využitie

- spracovanie dokumentov konvertovaných do HTML
- príprava dát pre vyhľadávanie
- práca s formátovaným textom

3.4 Spracovanie diakritiky

LALS podporuje režim práce s diakritikou:

- unaccent
- postunaccent

3.4.1 `unaccent=true`

Používa sa pre spracovanie textu bez diakritiky (napr. e-maily, chat).

- vstup sa spracuje bez ohľadu na diakritiku
- vhodné pre nekvalitný vstup, v ktorom chceme vyhľadávať len s ignorovaním diakritiky

---

3.4.2 `postunaccent=true`

Používa sa pre spracovanie textu, ktorý obsahuje diakritiku, avšak vyhľadávať chceme s ignorovaním diakritiky

- vstup sa spracuje s ohľadom na diakritiku
- vhodné pre kvalitný vstup, v ktorom chceme vyhľadávať s ignorovaním diakritiky

---

4 Výkon

Služba je optimalizovaná pre vysoký výkon:

- viac ako **400 000 slov / sekundu**
- latencia:
  - ~1 ms (krátke texty)
  - ~5–10 ms (dlhé texty)
- stabilné správanie pri súbežnej záťaži

4.1 Benchmark prostredie

- CPU: 12-core virtual server
- RAM: 6 GB
- Java: OpenJDK 17
- OS: Debian Linux Trixie

4.2 Typický výkon

- 400 000+ slov / sekundu
- priemerná latencia < 5 ms
- nízka pamäťová náročnosť
- stabilné správanie pri súbežnej záťaži

---

5 Nasadenie

LALS je distribuovaný ako Debian balík (.deb) a beží ako samostatná systemd služba poskytujúca HTTP API na konfigurovateľnom porte.

Podrobné kroky nasadenia sú popísané v dokumente LALS Deployment Guide.

5.1 Požiadavky

- Java 17+
- odporúčané: 6 GB RAM

---

6 Integrácia

LALS je navrhnutý ako jednoduchý komponent pre integráciu do existujúcich systémov.

6.1 Typické použitie

- fulltextové vyhľadávanie
- dokumentové systémy
- spracovanie textu
- NLP pipeline

6.2 Integrácia

- REST API
- bez potreby tréningu modelu
- bez externých závislostí

👉 Integrácia je možná v ľubovoľnom systéme podporujúcom HTTP komunikáciu.

6.3 Obmedzenia

- vracia viac možných lemma tvarov pri nejednoznačných slovách
- zameraná výhradne na slovenský jazyk

---

7 Rozšírenie: inteligentné rozšírenie dotazu (Query Expansion)

Query Expansion je voliteľný prémiový modul dostupný v rámci samostatnej komerčnej licencie.

Na rozdiel od lematizácie, ktorá vracia lemma tvary slov,
query expansion vracia širšie jazykové vzťahy (napr. odvodeniny, synonymá).

7.1 Princíp

Pre zadaný výraz služba vracia skupiny súvisiacich výrazov podľa typu jazykového vzťahu.

7.2 Endpoint

```
POST /text/ES
```

7.2.1 Formát vstupu

Na vstupe je očakávaný parameter „data“ so slovom alebo slovným spojením, ku ktorému chceme získať návrhy.

7.2.2 Formát výstupu

Každý riadok obsahuje:

1. typ rozšírenia (číselný kód)
2. množinu návrhov (slová alebo frázy)

👉 Všetky položky za typom predstavujú návrhy
👉 Jeden z návrhov je vždy zhodný s pôvodným výrazom
👉 Každý riadok reprezentuje jednu skupinu výrazov so spoločným jazykovým vzťahom.

7.2.3 Príklady použitia

Príklad (významové rozšírenie pre slovo „slovensko“)

Vstup:

```
POST /text/ES
data=slovensko
```

Výstup:

```
4 ; slovák ; slovenka ; slovenská republika ; slovensko ; slovenský
21 ; slovensko ; slovenský štát
```

Príklad (morfologické rozšírenie pre slovo „test“)

Vstup:

```
POST /text/ES
data=test
```

Výstup:

```
8  ; test ; testíček ; testík
10 ; test ; testový
16 ; test ; testisko
20 ; test ; testovať
```

Príklad (sloveso „testovať“)

Vstup:

```
POST /text/ES
data=testovať
```

Výstup:

```
11 ; testácia ; testovať
11 ; testovanie ; testovať
13 ; testovaný ; testovať
14 ; otestovať ; testovať
20 ; test ; testovať
```

7.3 Podpora vstupu bez diakritiky

7.3.1 Príklad (slovo „stat“)

👉 Jeden vstup bez diakritiky môže viesť k viacerým jazykovým interpretáciám:

`stat → štát / stať / stáť / sťať`

Vstup:

```
POST /text/ES
unaccent=true
data=stat
```

Výstup:

```
8 ; štát ; štátik
10 ; štát ; štátne ; štátny
11 ; stáť ; státie
11 ; stať ; statie
11 ; šťanie ; šťať
11 ; sťať ; sťatie
13 ; šťaný ; šťať ; šťatý
13 ; stáť ; stojaci
13 ; sťať ; sťatý
13 ; šťať ; štiaci
14 ; stáť ; stávať
14 ; stať ; stávať
14 ; šťať ; vyšťať
14 ; sťať ; stínať
19 ; stať ; udiať
19 ; sťať ; zrúbať
```

7.4 Využitie

- rozšírenie používateľského dotazu
- zvýšenie úspešnosti vyhľadávania
- návrhy alternatívnych výrazov
- podpora pokročilého vyhľadávania

👉 Výstup je štruktúrovaný podľa typov jazykových vzťahov,
čo umožňuje selektívne použitie alebo váhovanie jednotlivých typov rozšírenia.

7.5 Podporované typy rozšírenia:

- rovnaký význam
- zdrobnelina
- zveličenina

- podstatné meno / prídavné meno / príslovka
- prvok / prídavné meno
- zviera / prídavné meno
- priezvisko / prídavné meno
- remeslo / osoba / prídavné meno
- vlastnosť / osoba / prídavné meno / príslovka

- osoba / privlastňovacie prídavné meno
- miesto / obyvateľ / prídavné meno
- povolanie / osoba / prídavné meno

- sloveso (párový vid)
- sloveso / podstatné meno (činnosť)
- sloveso / podstatné meno (súvisiace)
- sloveso / prídavné meno
- sloveso / prídavné meno (schopnosť)

7.6 Poznámky

- rozšírenie môže zvýšiť počet výsledkov
- odporúča sa používať selektívne
- vhodné kombinovať s váhovaním

---

8 Zhrnutie

LALS poskytuje jednoduchý a výkonný spôsob spracovania slovenského textu v produkčných systémoch.
Služba je navrhnutá pre produkčné použitie v systémoch, kde je prioritou nájsť všetky relevantné informácie.

Vďaka podpore:

- viacznačných lemma kandidátov
- spracovania HTML
- práce s diakritikou
- inteligentného rozšírenia dotazu

umožňuje:

- zlepšiť kvalitu vyhľadávania
- znížiť počet nenájdených relevantných výsledkov
- jednoducho integrovať spracovanie textu do existujúcich systémov

8.1 Prečo LALS namiesto stemmingu?

- slovenská morfológia je príliš zložitá pre stemming
- stemming často znižuje presnosť vyhľadávania
- LALS vracia reálne slovníkové lemma tvary
- podporuje viacvýznamové slová
- správa sa predvídateľne v produkcii

---

9 Ďalšie informácie

Nasadenie služby je popísané v dokumente:
„LALS deployment guide.pdf“

Integrácia do Apache Solr je popísaná v dokumente:
„LALS – Solr Integration Guide.pdf“

Benchmark služby LALS:
„LALS benchmark.pdf“