Výkonnostná benchmark správa – LALS

1. Prehľad

Tento dokument sumarizuje výsledky výkonnostného testovania služby LALS (Slovak Lemmatization Service). Cieľom benchmarku bolo vyhodnotiť latenciu, priepustnosť, škálovateľnosť a stabilitu systému pri rôznych úrovniach súbežnej záťaže.

Služba je navrhnutá ako REST API pre lematizáciu slovenského textu, s podporou spracovania textov s diakritikou aj bez diakritiky.

---

2. Testovacie prostredie

- Nasadenie: virtuálny server
- CPU: 12 jadier
- Pamäť: 4 GB RAM
- Typ služby: Java REST služba
- Metóda: POST

---

3. Konfigurácia testov

Každý test bol vykonaný s nasledujúcimi parametrami:

- Počet požiadaviek na vlákno: 1000
- Warmup požiadavky na vlákno: 100
- Veľkosť datasetu: 5 položiek (dlhé texty)
- Validácia odpovede: zapnutá

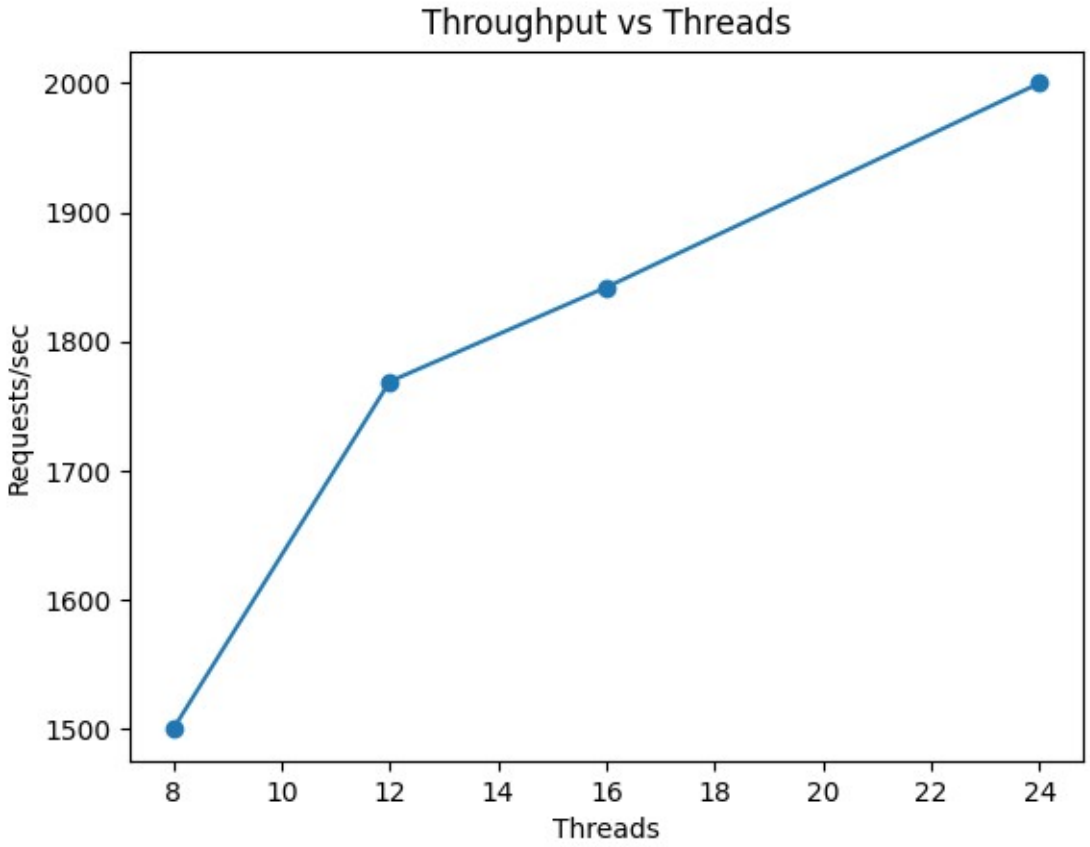
Testované režimy:

- Štandardný režim (s diakritikou)
- Režim bez diakritiky (unaccent)

---

4. Výsledky benchmarku

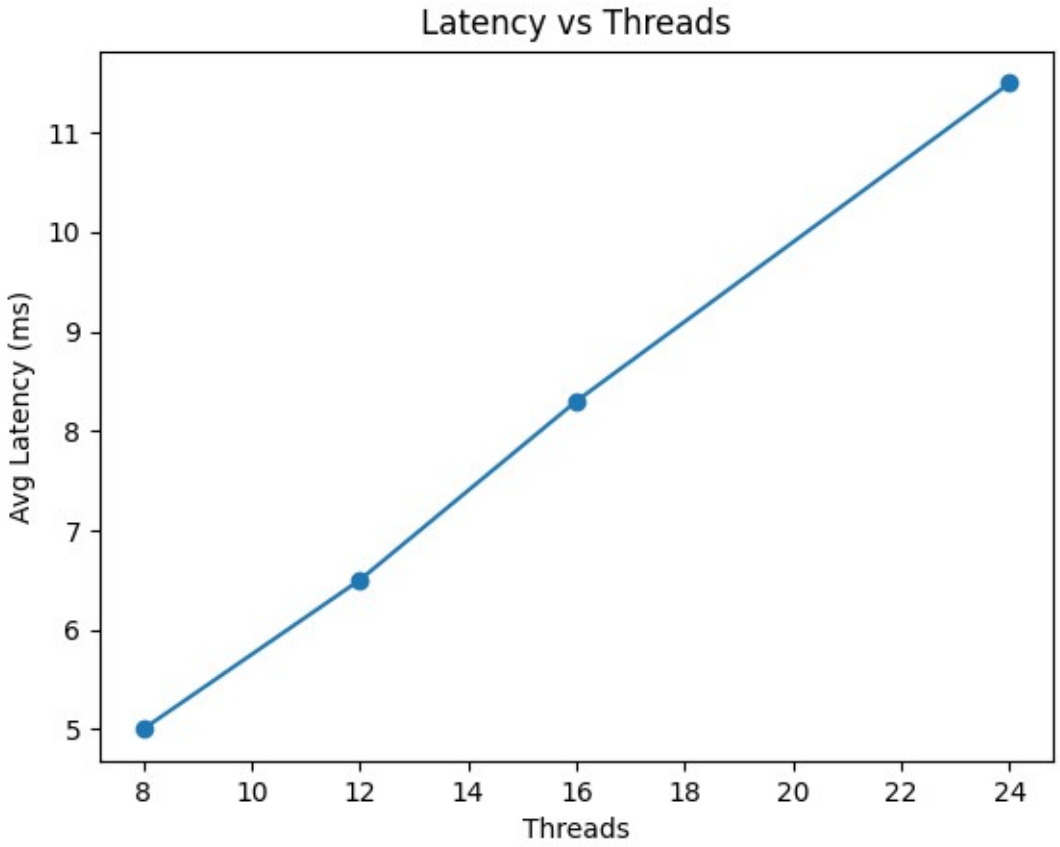
1. Závislosť priepustnosti od počtu vlákien



Graf 1: Závislosť priepustnosti od počtu vlákien

Graf zobrazuje rast priepustnosti služby LALS v závislosti od počtu paralelne bežiacich vlákien. Výsledky ukazujú, že služba efektívne škáluje približne do úrovne počtu dostupných CPU jadier, pričom po prekročení tejto hranice sa rast priepustnosti spomaľuje.

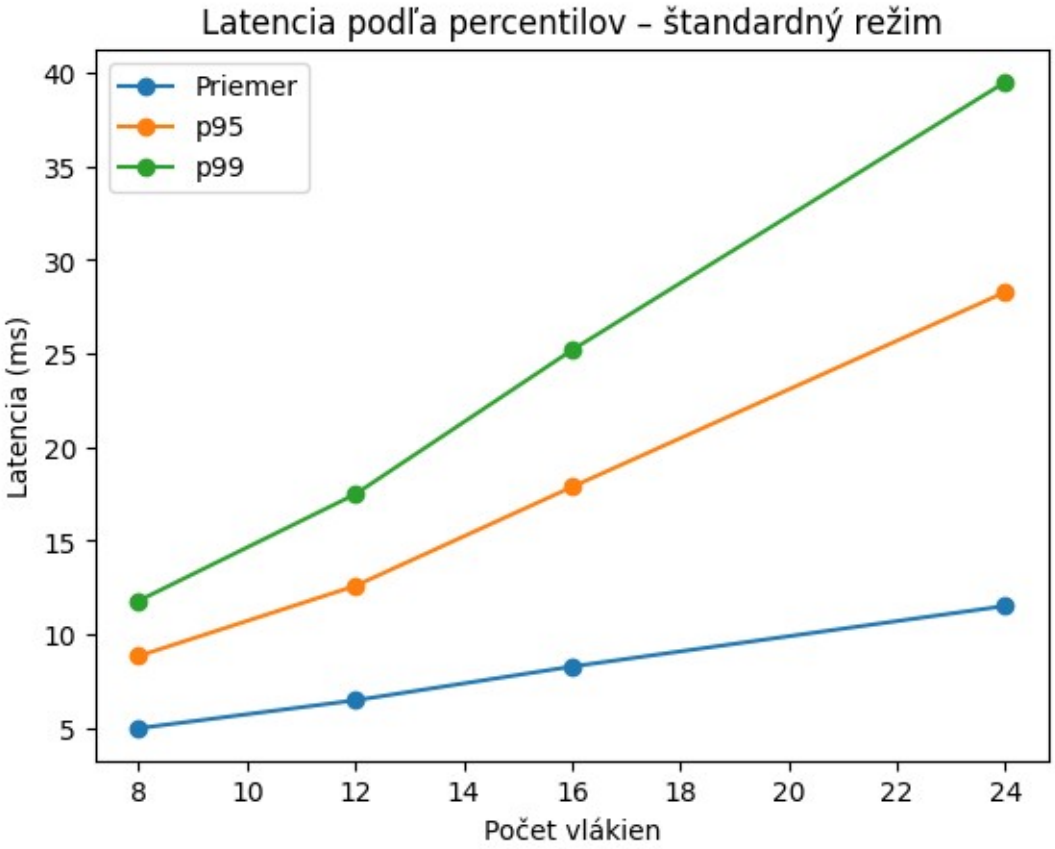
2. Závislosť priemernej latencie od počtu vlákien



Graf 2: Závislosť priemernej latencie od počtu vlákien

Graf ilustruje vývoj priemernej latencie pri rastúcej súbežnej záťaži. Latencia rastie plynule a bez náhlych výkyvov, čo potvrdzuje stabilné a predvídateľné správanie služby aj pri vyššom zaťažení.

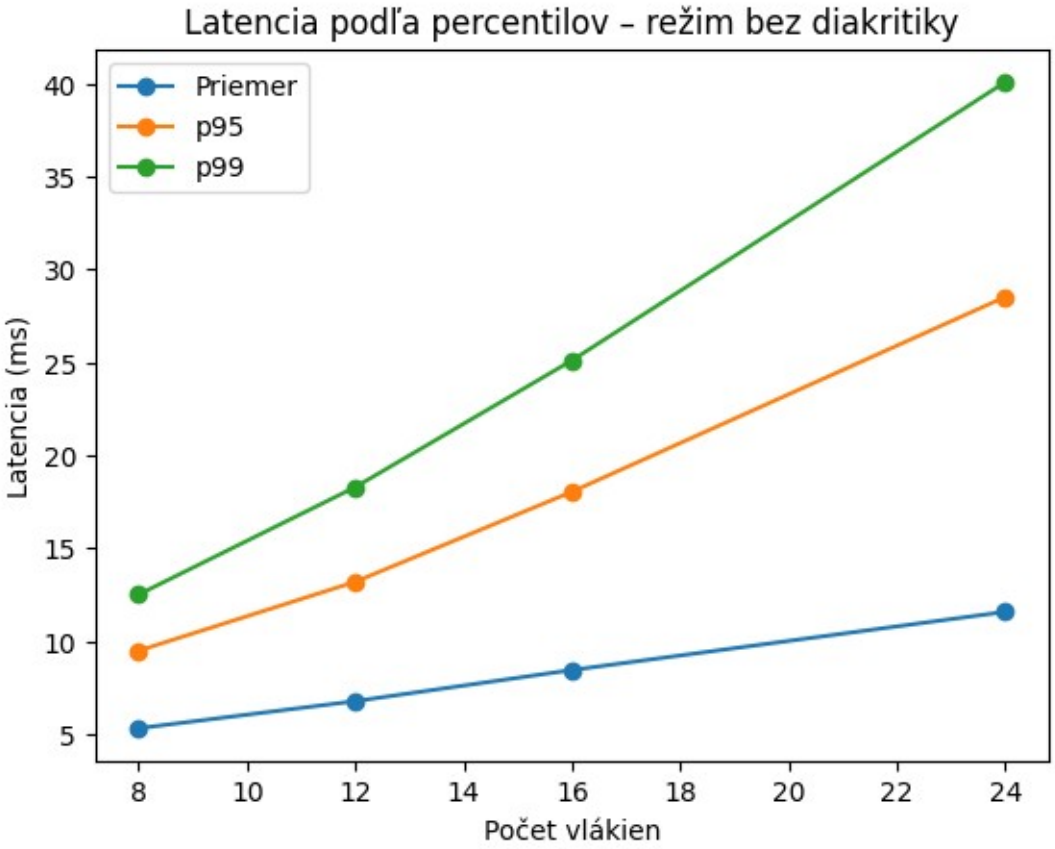
3. Latencia podľa percentilov – štandardný režim



Graf 2: Latencia podľa percentilov – štandardný režim

Graf zobrazuje vývoj priemernej latencie, p95 a p99 v štandardnom režime spracovania. Rozdiel medzi priemernou hodnotou a vyššími percentilmi zostáva kontrolovaný, čo znamená, že aj chvostové latencie sú pri produkčnej záťaži prijateľné.

4. Latencia podľa percentilov – režim ignorovania diakritiky



Graf 4: Latencia podľa percentilov – režim ignorovania diakritiky

Graf porovnáva priemernú latenciu, p95 a p99 pri spracovaní textov bez diakritiky. Výsledky potvrdzujú, že aktivácia režimu unaccent má len minimálny dopad na výkon a nemení stabilný charakter systému.

4.1 Prehľad priepustnosti a latencie

Počet vlákien	Požiadavky/s	Slová/s	Priemerná latencia	p95	p99
8	~1500	~320 000	~5.0 ms	~8.8 ms	~11.8 ms
12	~1769	~376 000	~6.5 ms	~12.6 ms	~17.5 ms
16	~1842	~392 000	~8.3 ms	~17.9 ms	~25.2 ms
24	~2000	~426 000	~11.5 ms	~28.5 ms	~40.0 ms

4.2 Porovnanie režimu bez diakritiky (unaccent)

Počet vlákien	Požiadavky/s	Slová/s	Priemerná latencia
12	~1680	~357 000	~6.8 ms
16	~1805	~384 000	~8.4 ms
24	~2008	~428 000	~11.6 ms

Pozorovanie:
Režim bez diakritiky spôsobuje len minimálne spomalenie (približne 5–7 %), ktoré sa pri vyššej záťaži prakticky stráca.

---

5. Kľúčové zistenia

5.1 Vysoká priepustnosť

- Maximálna priepustnosť presahuje **426 000 slov za sekundu**
- Až **2000 požiadaviek za sekundu** pri spracovaní dlhých textov

To radí LALS medzi vysoko výkonné komponenty pre spracovanie textu vo veľkom rozsahu.

5.2 Nízka latencia

- Krátke texty: ~1 ms priemerná latencia
- Dlhé texty: ~5–11 ms priemerná latencia pri záťaži

Latencia zostáva v jednotkách milisekúnd aj pri paralelnom spracovaní.

5.3 Škálovateľnosť

- Takmer lineárne škálovanie do približne 12 vlákien (počet CPU jadier)
- Po prekročení počtu jadier dochádza k postupnému znižovaniu efektivity
- Stabilné správanie aj pri vyššej záťaži

5.4 Stabilita

- 100 % úspešnosť vo všetkých testoch
- Bez chýb a timeoutov
- Bez výrazných výkyvov latencie

5.5 Predvídateľné správanie

- Latencia rastie plynule so záťažou
- p99 latencia zostáva pod 40 ms aj pri vysokej záťaži

---

6. Odporúčaná konfigurácia

Na základe výsledkov:

- Odporúčaný počet vlákien: **10–12**
- Očakávaný výkon:
  - ~370 000 – 380 000 slov/s
  - ~6–7 ms priemerná latencia

Táto konfigurácia poskytuje optimálny pomer medzi výkonom a latenciou.

---

7. Architektonická interpretácia

Služba vykazuje nasledovné vlastnosti:

- CPU-bound spracovanie
- bezstavový (stateless) dizajn
- efektívna paralelizácia
- minimálna synchronizačná réžia

Výkon je primárne limitovaný počtom dostupných CPU jadier.

---

8. Vhodnosť použitia

LALS je vhodný pre:

- predspracovanie textu pre fulltextové vyhľadávanie (napr. Solr/Lucene)
- pipeline pre spracovanie dokumentov
- analýzu e-mailov a komunikácie
- podnikové vyhľadávacie systémy
- eGovernment a archívne systémy

---

9. Záver

Služba LALS dosahuje vysokú priepustnosť, nízku latenciu a stabilné správanie aj pri súbežnej záťaži.

Schopnosť spracovať viac ako 400 000 slovenských slov za sekundu pri zachovaní nízkej latencie ju radí medzi produkčne použiteľné NLP komponenty pre veľké systémy.

Podpora spracovania textov s diakritikou aj bez diakritiky s minimálnym dopadom na výkon predstavuje významnú praktickú výhodu pri spracovaní reálnych dát.

---

10. Súhrn

LALS poskytuje vysokovýkonnú lematizáciu slovenského jazyka s latenciou v jednotkách milisekúnd pri realistickej záťaži a priepustnosťou presahujúcou 400 000 slov za sekundu na 12-jadrovom serveri, pričom si zachováva stabilitu a minimálny výkonový dopad pri spracovaní textov bez diakritiky.

---

11. Porovnanie s alternatívami

Pri hodnotení LALS je dôležité porovnať jeho vlastnosti s bežne používanými prístupmi a nástrojmi na spracovanie prirodzeného jazyka.

11.1 Všeobecné NLP knižnice (napr. spaCy, NLTK)

Charakteristika:

- univerzálne nástroje pre viac jazykov
- široká funkcionalita (tokenizácia, POS tagging, NER, atď.)

Obmedzenia pre slovenský jazyk:

- obmedzená kvalita modelov pre slovenčinu
- slabšia podpora morfológie
- nižší výkon pri spracovaní veľkých objemov textu

Porovnanie s LALS:

- LALS je výrazne rýchlejší pre samotnú lematizáciu
- vyššia presnosť pre slovenské tvary a priezviská
- jednoduchšie nasadenie ako samostatná služba

11.2 Akademické nástroje (napr. UDPipe, MorphoDiTa)

Charakteristika:

- kvalitné lingvistické modely
- podpora viacerých jazykov vrátane slovenčiny
- často využívané v akademickom prostredí

Obmedzenia:

- vyššia latencia
- nižšia priepustnosť pri veľkých objemoch dát
- zložitejšia integrácia do produkčných systémov

Porovnanie s LALS:

- LALS dosahuje výrazne vyššiu priepustnosť (státisíce slov/s)
- nižšia latencia (jednotky milisekúnd)
- jednoduché REST API vhodné pre produkčné použitie

11.3 Jednoduché prístupy (stemming, regex normalizácia)

Charakteristika:

- veľmi rýchle
- jednoduchá implementácia

Obmedzenia:

- nízka jazyková presnosť
- nevhodné pre slovenčinu (bohatá morfológia)

- neschopnosť spracovať rôzne tvary slov správne

Porovnanie s LALS:

- LALS poskytuje výrazne vyššiu kvalitu výsledkov
- stále si zachováva veľmi vysoký výkon

11.4 Výhody LALS

- špecializácia na slovenský jazyk
- rozsiahly slovník (milióny tvarov)
- podpora textov bez diakritiky
- algoritmické spracovanie neznámych slov
- veľmi nízka latencia
- vysoká priepustnosť
- jednoduché REST API

11.5 Zhrnutie porovnania

LALS predstavuje kompromis medzi:

- kvalitou lingvistického spracovania
- výkonom
- jednoduchosťou integrácie

V porovnaní s univerzálnymi NLP nástrojmi a akademickými riešeniami ponúka výrazne lepší výkon pre slovenský jazyk, pričom si zachováva dostatočnú presnosť pre praktické nasadenie v produkčných systémoch.

11.6 Kontakt

[Róbert Baláž]
[bohem303@gmail.com]
[0950 259 480]
[https://bob303.duckdns.org]