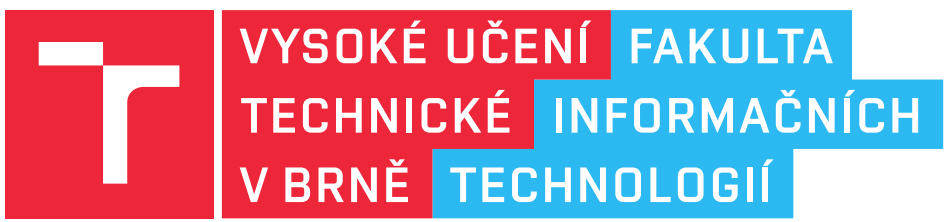
Rozšíření systému pro získávání, zpracování a analýzu rozsáhlých kolekcí textů z webu

Jiří Matějka

Vysoké učení technické v Brně, Fakulta informačních technologií
Božetěchova 1/2 612 66 Brno
xmatej52@stud.fit.vutbr.cz



14. 6. 2018

Motivace

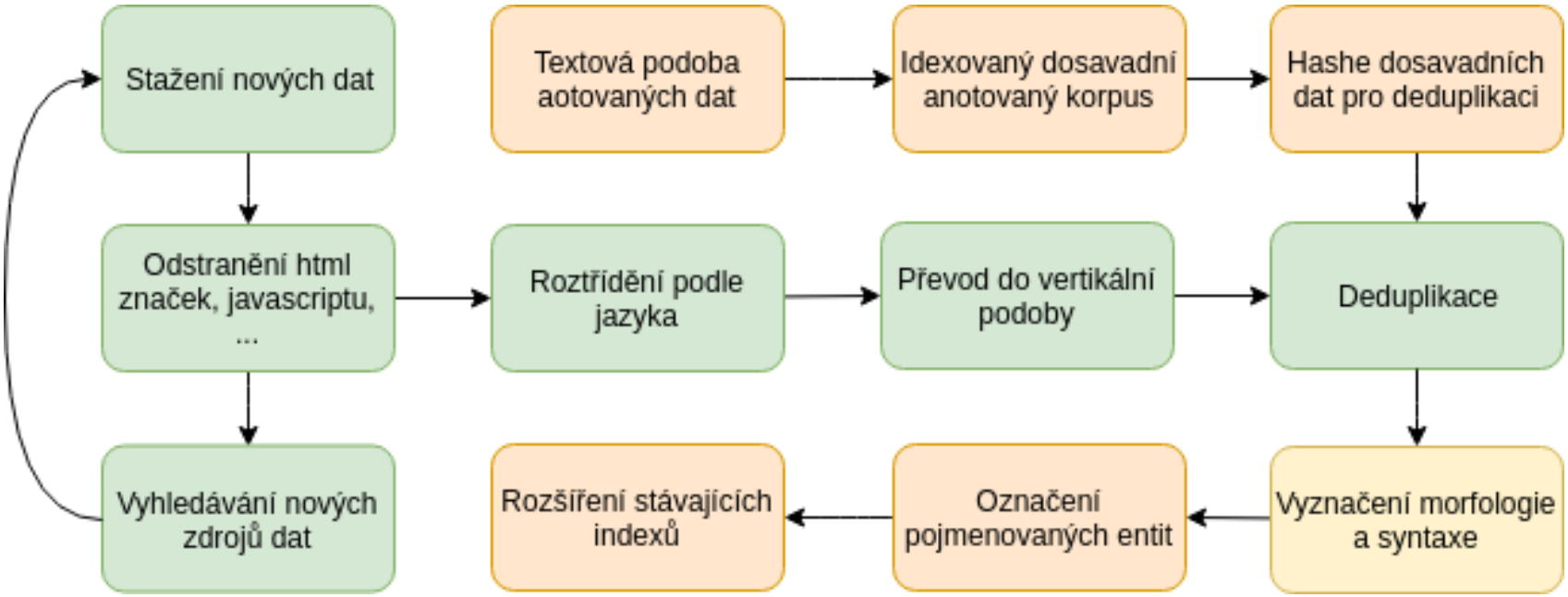
Dosavadní systém:

- jednorázové vytvoření korpusu z předem shromážděných dat (zatím ověřeno pouze na angličtině)

Omezení:

- jednotlivé kroky zpracování korpusových dat vyžadují ruční zásahy
- nejsou ošetřeny chybové stavy prvotních fází zpracování
- není možné korpusová data a indexy snadno aktualizovat
- nelze odhadnout, jak dlouho bude zpracování nových dat probíhat
- není možné zpracovat dokumenty psané ve více jazycích
- v případě chyby se její zdroj hledá velice obtížně

Zpracování


Stažení nových dat
Textová podoba aotovaných dat
Idexovaný dosavadní anotovaný korpus
Hashe dosavadních dat pro deduplikaci
Odstranění html značek, javascriptu, ...
Roztřídění podle jazyka
Převod do vertikální podoby
Deduplikace
Vyhledávání nových zdrojů dat
Rozšíření stávajících indexů
Označení pojmenovaných entit
Vyznačení morfologie a syntaxe

Rozšíření dosavadního systému

Automatizace zpracování:

- vyhledávání nových zdrojů dat
- stahování, zpracování a archivace dat

Logování:

- monitorování běžících procesů zpracování
- vedení podrobných a přehledných logů
- vedení statistik
- pravidelné odesílání emailů o průběhu zpracování

Zpracování dat:

- zpracování dokumentů psaných ve více jazycích
- lepší způsob archivace dat
- archiv zpracován více procesy

Spuštění

Spuštění vertikalizace nad již staženými daty (145 GiB):

- 412 minut (nový systém, 12 procesů) $\times$ 354 minut (původní systém, 12 procesů)
- bylo ztraceno minimum dat $\times$ data psaná v jiném než českém jazyce byla zahozena

Spuštění automatického zpracování (3 týdny samostatné činnosti):

- počet RSS a ATOM zdrojů vzrostl ze 116 000 na 185 000
- každý den nalezeno v průměru o 15 000 více článků ke stažení
- každý den staženo v průměru 10x více článků
- během této doby nebyla hlášena žádná chyba (tzn. žádné zpracování neskončilo s chybou)

Budoucí vývoj

Rozšíření automatického zpracování:

- deduplikace vertikalizovaných dat
- parsing
- tvorba a aktualizce indexů

Spuštění na více serverech:

- tvorba serveru řídícího zpracování
- tvorba klientů provádějící zpracování

Děkuji za pozornost.

Otázka oponenta

Můžete stručně popsat prototyp vytvořeného řešení přiložený na DVD a uvést, proč jste se následně rozhodl dané řešení opustit?

Prototyp:

- schopen zpracovat přibližně 70 % českých dat.
- manuální možnost tvorby indexů

Nedostatky prototypu:

- schopen zpracovat pouze česká data
- manuální obsluha
- nástroje mají nepoužitelné a nebo žádné logovací soubory
- velká doba zpracování
- nevhodný způsob provedení lemmatizace
- velká chybovost