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

Matějka Jiří

30. 1. 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Motivace

Co umíme:

- Umíme vytvořit jednorázově korpus ze stažených.
- Sémanticky označit korpus.

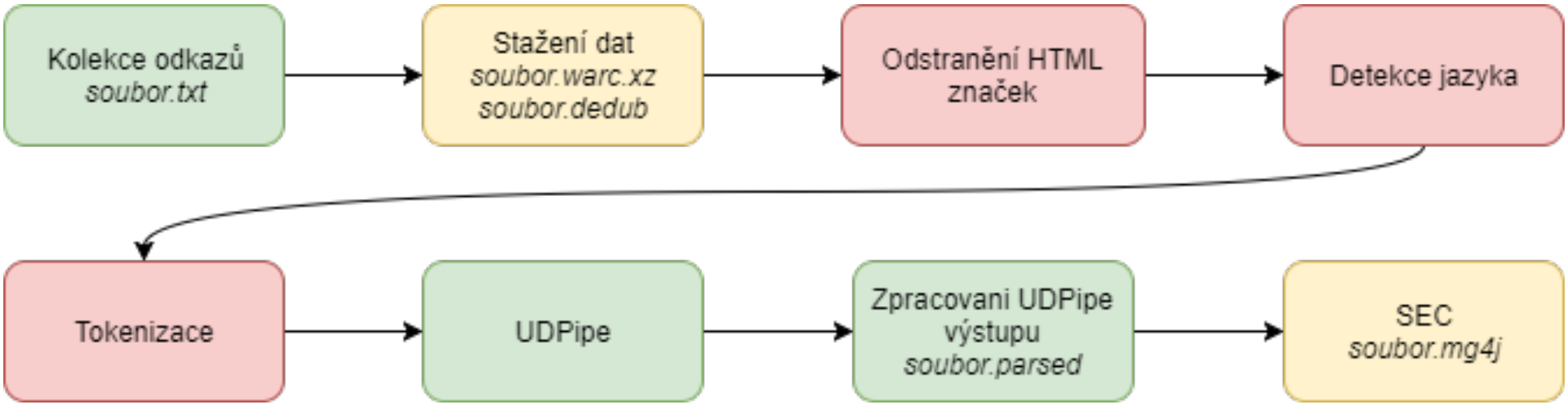
Kde je tedy problém:

- Neumíme tvorbu korpusu provádět automaticky.
- Sémanticky označený korpus neumíme aktualizovat.
- Špatná detekce jazyka a deduplikace dat.
- Výpočetní náročnost celého procesu zpracování.

Odkud brát nová data:

- Atom a RSS zdroje,
- blogy,
- komentáře u článků,
- sociální sítě,
- ...

Zpracování


Kolekce odkazů
soubor.txt
Stažení dat
soubor.warc.xz
soubor.dedub
Odstranění HTML značek
Detekce jazyka
Tokenizace
UDPipe
Zpracovani UDPipe výstupu
soubor.parsed
SEC
soubor.mg4j

Aktuální problémy během zpracování

- Spousta skriptů (často i nečitelných),
- vertikalizace není úplně vychytaná (např. neošetřené chybové stavy),
- zpracování není automatizované,
- spousta logů, u chyb často není možné zjitit přesně příčinu.

Časový plán na 2. semestr

- Sjednocení skriptů do větších celků (použití přepínačů, sdílené moduly) – během února,
- podrobně analyzovat vertikalizátor a buď opravit chyby nebo najít vhodnější řešení – během března (možná i duben),
- zautomatizovat zpracování – duben,
- sledování zpracování, měření náročnosti na procesor a paměť, tvorba statistik – duben, květen.

Co by bylo ještě užitečné

- Logy snadno parsovatelné regulárnímy výrazy nebo skripty,
- vedení podrobných statistik už během zpracování, možnost zasílat je pravidelně emailem,
- využití emailové notifikace na upozornění při kritických chybách.