

Report zimný semester

Cieľom práce bolo preskúmať možnosti rekonštrukcie súborov pomocou file carvingu a navrhnúť a otestovať jednoduché algoritmy na spájanie fragmentov textového súboru.

Priebeh práce počas semestra

- Oboznámil som sa s tematikou file carvingu. Pozrel som sa na postupy a algoritmy, ktoré používajú rôzne aplikácie na trhu, hlavne pre textové súbory.
- Implementoval som jednoduchý algoritmus, ktorý spája text rozdelený na rovnako veľké časti pomocou externého slovníka. Pri spájaní dvoch kúskov sa pozerá na novovytvorené slovo a hľadá ho v slovníku. Vytvoril som skript, ktorý vie takýto slovník vytvoriť priamo zo slov, ktoré sú v celku v rozhádzanom textovom súbore.
- Implementoval som dve ďalšie varianty overovania spojení: Slovník 2-slovných a 3-slovných spojení, ktoré sa pozerajú aj na slová okolo spojeného slova a v slovníku hľadajú takúto postupnosť slov. Rovnako som vytvoril skripty, ktoré spracujú dáta do formy použiteľnej slovníkmi.
- Začal som sa zaoberať použitím umelej inteligencie na vyhodnocovanie spojení. Oboznámil som sa so štúdiami, ktoré sa venovali problematike. Začal som s implementáciou modelu. Tento spôsob vyhodnocovania plánujem dokončiť v budúcom semestri.
- Otestoval som základnú logiku a efektivitu metód. Vyskúšal som implementované metódy na vstupoch v rôznych prírodných jazykoch. Tiež som vyskúšal metódy na zjednodušenom kóde v jazyku C. Pri každom jazyku som sa pozrel na to, ako ovplyvňuje dĺžka kúskov a celková dĺžka vstupu efektivitu algoritmov.

Výstup z testovania

- Pri obyčajnom jednoslovnom slovníku som zistil, že má veľký problém v jazykoch, kde sa skloňuje - keďže môj algoritmus má obmedzené veľký slovník, nie všetky tvary slov sa v slovníku nachádzali. Preto bol najúspešnejší v angličtine. Čím boli kúsky dlhšie, tým rýchlejšie prišiel algoritmus na nejaké riešenie. Na použitie interného slovníka žiaľ treba príliš veľký text, ktorý potom následne touto jednoduchou metódou nie je realistické spojiť. Pri jazyku C bol problém, že existovalo príliš veľa správnych možností, ako text spojiť, dokonca aj takých, čo sa skompilujú.

- Pri sekvenčných slovníkoch boli oveľa rýchlejšie vyhodnotené nesprávne prípady, čo umožnilo nájsť riešenia aj pre dlhšie texty. Problém bola vysoká chybovosť výsledkov pri menších dĺžkach segmentov, ako aj to, že pokiaľ v dátach nebola nejaká korektná postupnosť slov, ale vstupný text ju obsahoval, nájsť riešenie nie je možné. V jazyku C pretrvávajú problémy z jednoduchého slovníka.