Zápisnica 10

Dátum	28.11.2018 – 12:00
Miesto	FIIT STU, 4.26
Zapisovateľ	Dávid Csomor

Úvod – pripomienky

- Riešenie bug-u zoraďovania v indexe, prepojenosti úloh v TFS
    - stačí zoraďovať podľa tfidf, count
- Vytvoriť user-stories pre chyby (?)
- Features v TFS zoradil product owner podľa dôležitosti
- Treba sa dohodnúť v akom „stave“, v akej „forme“ a aké dáta budeme mať
    - Spojiť korpusy „štáty“ a „mestá“ pod názov „GEO“ (+- 500)
- Potreba prepracovať crawler nakoľko už začína byť blokovaný
- Treba zapracovať do skriptu, aby sťahovalo po častiach, nech môžeme sťahovať z viacerých strojov
- Odporúčanie:
    - najprv sa pozrieť, či aktuálna URL nebola crawlovaná posledný týždeň
        - Ak áno: zobrať HTML z cache
        - Ak nie: stiahnuť
- Vybrať kategórie, ktoré pridať do korpusov
- Treba používať relatívne metriky (relatívny počet výskytu nejakého slova vs absolútnemu výskytu)
- Odtlačok článku je spredmetnený v zistení:
    - Počte podstatných mien
    - Koľko slov článku je patrí do TOP 1000
    - ...

Úlohy

- Vybrať kategórie, ktoré pridať do korpusov
- Spraviť „exact match“ pri vyhľadávaní v invertovanom indexe
- Prepracovať crawler
    - aby nebol blokovaný stránkami, ktoré crawluje
- Pridať na úvodnú stránku nejakú zaujímavú štatistiku