7. Stretnutie tímu

Dátum: 08. 11. 2021

Miesto: Discord

Prítomní: všetci

Poznámky:

- story zavesenie klienta je hotova, David nam ukazal/vysvetlil kod
- story sekvencne scrapovanie je hotove, Dominik to urobil cez xpath a lxml
- Jakub ukazal docker s mongom
- Tana ukazala ako funguje readability, a ze nezavisi od jazyka stranky
- navrh na story do dalsieho sprintu integracia mongo a elastic prostrednictvom nastroja
- Adam ukazal, ako implementoval naplnanie Mongo priamo pri scrapovani, treba dorobit, aby sa odpovede ine ako 200 ukladali hned a nie az nakonci programu. Ostava nahodit crime maps a ulozit linky
- Moric robil so scrapy xpath a este mu ostava zmerat casovu zlozitost
- zajtrajsie stretnutie urobime prezencne ale Jakub M. sa pripoji online

7. Stretnutie s vedúcim

Dátum: 16. 11. 2021

Miesto: Miestnosť 5.27

Prítomní: všetci

Poznámky:

- Informovanie veduceho o redukcii html body tagu prostrednictvom lxml, readability, xpath v scrapy
- Budeme potrebovat 2TB disku, 4CPU, 8GB RAM -> veduci sa bude snazit vybavit

Review šprintu č. 2:

Adam - upravil scraper, vsetko sa uklada do MongoDB, ale treba este nejake minimalne upravy (6h 30 min este hodina a pol k dobru)

Jakub - Vytvoril docker-compose pre MongoDB a elasticsearch, skusal nahodit vsetky clanky do elasticsearch z MongoDB, ale elasticsearch mal problem s parsovanim, navrh na alokovanie viac casu a dokoncenie v dalsom sprinte… Pozriet sa na river plugin pre integraciu MongoDB a elasticsearch (6h vs 6,5h)

Jakub M. - robil stranku prvu polovicu sprintu (dokoncena, trvalo alkovanu 1 hodinu), vysledok scrapy xpath je o minuty rychlejsi ako Dominove a zda sa naefektivnejsi sposob selekcie tagov (trvalo 6 hodin prekrocil cas o 1 hodinu)

Dominik - Redukovanie dát s lxml. Readability velmi pomale, takze skusil beautiful soup. Na nase pomery bola velmi pomala a niektore clanky boli duplikatne, pretoze to kniznica robi cez rekurziu. Idealne vyzerala lxml aj veduci schvalil... Robil vyskum na RSS nejake dvojca na RSS neexistuje (7h vs 7h).

Táňa - Readability cele zalozene na tagoch a classach, pridava nejake vahy, takze nie su zavisle na jazykoch. Trafilatura pouzije readability v urcitom kroku, co je nic moc.

Robila porovnanie a trafilature vracia holy text a nie tagy.

Vyskusala slovensky clanok a nebol ziaden problem.

Alokovane 3 hodiny a trvalo 6 hodin

David - 11h 30 min vs 13h, klient je zaveseny. Spravil frontend a backend do jedneho na github, nech sa na tom lahsie pracuje. Potrebne spravit wraper na api calls, je mozne pouzit redux, ale tiez ho mozno nebudeme potrebovat pouzit.

Čo hodnotíme pozitívne?

- Naplánovali sme si 62 hodín, čo vychádza, že každý z nás mal 1.6 hodiny rezervu. Nakoniec sme odrobili viac, a to 68 hodín, čo je oproti predošlým 43.5 posun vpred. Úlohy sme si pridali v priebehu šprintu.
- Náročnosti jednotlivých príbehov sa celkom vyrovnali.

Čo chceme zlepšiť?

- Chceme príbehy špecifikovať menej, aby sme príbehy nemuseli rušiť, ak narazíme na nejaký problém.
- Zistili sme, že náročnosť úlohy je výrazne odlišná, ak na úlohe pracuje človek, ktorý s danou technológiou má skúsenosti a človek, ktorý má minimálne skúsenosti.

- Ak príbeh pridelíme človeku s minimálnymi skúsenosťami, chceme daný príbeh radšej rozdeliť na menšie.

Názov príbehu	**Riešiteľ**	**Odhadovaný počet bodov**	**Je príbeh ukončený?**	**Pridelený počet bodov**
Parsovanie tagov pomocou CSS selector	Jakub Müller	5	áno	13
Parsovanie tagov pomocou regex	Táňa Poláková	8	zrušený	3
Upratať Jiru	Táňa Poláková	2	áno	3
Analýza bezstratovej kompresie textu	Jakub Hlavačka	8	áno	8
Lokálna MongoDB so získanými dátami	Adam Šípka	8	áno	5
Request URL s odpoveďou inou ako 200 uložiť do súboru	Adam Šípka	3	áno	1
Pridať info o projekte a opísať členov tímu	Jakub Müller	3	áno	1
Rozšíriť docker compose	Dominik Horváth	13	áno	8
Zavesenie klienta	Dávid Silady	13	áno	13
Parsovanie tagov pomocou lxml, readability, trafilatura	Táňa Poláková	3	áno	8
Pamäťová zložitosť Elasticsearch na vzorke dát	Jakub Hlavačka	5	nie	5
Rozšírenie google news scrapperu na lokáciu UK	Dominik Horváth	1	áno	1
Vyhľadanie RSS dvojčaťa k HTML článku - všeobecný postup	Dominik Horváth	8	áno	5

Napĺňanie MongoDB priamo pri scrapovaní	Adam Šípka	8	nie	8

Meno	**Súčet počtu odhadovaných bodov**	**Súčet počtu pridelených bodov**	**Podiel práce (%)**
Jakub Hlavačka	16	13	16.66667
Dominik Horváth	22	13	16.66667
Jakub Müller	9	13	16.66667
Táňa Poláková	13	13	16.66667
Dávid Silady	13	13	16.66667
Adam Šípka	19	13	16.66667

Nový šprint – č. 3

Trvanie: 09. 11. 2021 – 23. 11. 2021

Názov príbehu	**Riešiteľ**	**Odhadovaný počet bodov**
Napĺňanie MongoDB priamo pri scrapovaní	Adam Šípka	2
Zabrániť zacykleným buildom	Dominik Horváth	5
Rozšíriť úložisko pomocou nepriradeného disku	Dominik Horváth	5
Integrácia scrapera a vylepšeného parsera	Jakub Müller	5
Integrácia MongoDB a Elasticsearch	Jakub Hlavačka	8
Úprava dokumentácie zo stretnutí a šprintov tak, aby mohli ísť na stránku tímu	Táňa Poláková	8

Zabezpečenie komunikácie medzi klientom a serverom	David Silady	8
Analýza možností implementácie testovania	Adam Šípka	5
Návrh Flask API	Táňa Poláková	5
Implementácia dummy Flask API	Dominik Horváth	5