

Export úloh 16. 11. 2021

Obsah

IN PROGRESS	4
[AMS-89] Integrácia scraperu a vylepšeného parseru Created: 09/Nov/21	4
[AMS-93] Analýza možností implementácie testovania Created: 09/Nov/21	5
[AMS-11] Lofi používateľského rozhrania Created: 05/Oct/21 Updated: 12/Oct/21	6
[AMS-36] User databáza + zbieranie aktivity používateľa Created: 13/Oct/21 Updated: 17/Oct/21	7
TO DO	8
[AMS-102] Kontajnerizácia scraperu Created: 15/Nov/21	8
[AMS-103] Rozšírenie docker compose o MongoDB a Elasticsearch Created: 15/Nov/21 ...	9
[AMS-92] Zabezpečenie komunikácie medzi klientom a serverom Created: 09/Nov/21	10
[AMS-91] Úprava dokumentácie zo stretnutí a šprintov tak, aby mohli ísť na stránku tímu Created: 09/Nov/21	11
[AMS-67] Pozrieť sa na zahodené články, kt. neobsahovali paragrafy Created: 02/Nov/21	12
[AMS-68] Analýza médií, ktoré dostaneme vo výstupoch a aké search queries treba poslať, aby sme mali výsledky ako z Google Created: 02/Nov/21	13
[AMS-72] Vytvorenie developerského prostredia Created: 02/Nov/21	14
[AMS-86] PDF report z nájdených článkov Created: 02/Nov/21	15
[AMS-85] Prihlásenie sa na odber nejakého vyhládaného človeka Created: 02/Nov/21	16
[AMS-84] Označenie článku ako videného (prečítaného) Created: 02/Nov/21	17
[AMS-83] Hodnotenie risku pri nájdených článkoch Created: 02/Nov/21	18
[AMS-82] Zvýraznenie hľadaných slovíčok Created: 02/Nov/21	19
[AMS-77] Filter na web stránke Created: 02/Nov/21	20
[AMS-42] Porovnanie výsledkov nášho vyhľadávania s výsledkami Google News Created: 13/Oct/21	21
[AMS-12] Nasu API dát na portal verejných API v rámci propagácie Created: 05/Oct/21 ..	22
[AMS-41] Vyriešiť, čo s článkami, ktoré budú v pôvodnom zdroji nekompletné Created: 13/Oct/21	23
[AMS-37] Zistiť, ako fungujú iné zdroje Google News, prípadne aký mechanizmus používa Google News. Created: 13/Oct/21	24
[AMS-33] Zistiť, čo spája kriminálne články Created: 13/Oct/21	25
[AMS-19] Analyza - multilingvistika Created: 05/Oct/21	26
[AMS-71] Napĺňanie MongoDB priamo pri scrapovaní Created: 02/Nov/21 Resolved: 14/Nov/21	27
DONE	28

[AMS-60] Pamäťová zložitosť elastic search na vzorke dát Created: 01/Nov/21 Resolved: 12/Nov/21	28
[AMS-87] Zabrániť zacykleným buildom Created: 07/Nov/21 Resolved: 15/Nov/21.....	30
[AMS-88] Rozšíriť úložisko pomocou nepriradeného disku Created: 09/Nov/21 Resolved: 14/Nov/21	31
[AMS-90] Integrácia MongoDB a Elasticsearch Created: 09/Nov/21 Resolved: 16/Nov/21	32
[AMS-94] Návrh Flask API Created: 09/Nov/21 Resolved: 15/Nov/21	33
[AMS-95] Implementácia dummy Flask API Created: 09/Nov/21 Resolved: 15/Nov/21 ...	36
[AMS-46] Parsovanie tagov pomocou CSS Selector Created: 27/Oct/21 Resolved: 08/Nov/21	37
[AMS-57] Zavesenie klienta Created: 27/Oct/21 Resolved: 07/Nov/21	39
[AMS-59] Parsovanie tagov pomocou lxml, readability, trafiletura Created: 30/Oct/21 Resolved: 09/Nov/21	40
[AMS-73] Vyhľadanie RSS dvojčat'a k HTML článku - všeobecný postup Created: 02/Nov/21 Resolved: 09/Nov/21.....	43
[AMS-70] Rozšírenie google news scrapperu na lokáciu UK Created: 02/Nov/21 Resolved: 02/Nov/21	44
[AMS-51] Pridať info o projekte a opísať členov tímu Created: 27/Oct/21 Resolved: 01/Nov/21	45
[AMS-50] Request URL s odpoveďou inou ako 200 uložiť do súboru Created: 27/Oct/21 Resolved: 01/Nov/21	46
[AMS-49] Lokálna MongoDB so získanými dátami Created: 27/Oct/21 Resolved: 01/Nov/21	47
[AMS-48] Analýza bezstratovej kompresie textu Created: 27/Oct/21 Resolved: 01/Nov/21	48
[AMS-47] Parsovanie tagov pomocou regex Created: 27/Oct/21 Resolved: 30/Nov/21	49
[AMS-53] Rozšíriť docker compose Created: 27/Oct/21 Resolved: 31/Oct/21	50
[AMS-32] Prihláška na TP CUP Created: 13/Oct/21 Resolved: 02/Nov/21	51
[AMS-44] Stránka tímu Created: 18/Oct/21 Resolved: 01/Nov/21	52
[AMS-29] Získanie dát pre prototyp Created: 13/Oct/21 Resolved: 26/Oct/21	53
[AMS-43] Otvorenie portov na serveri Created: 18/Oct/21 Resolved: 18/Oct/21.....	54
[AMS-35] Prieskum databáz Created: 13/Oct/21 Resolved: 18/Oct/21	55
[AMS-39] Vyriešiť, aby fiitkar nemohol urobiť "sudo su" na virtuálnom stroji (aby nemal šancu sa zmeniť na roota) Created: 13/Oct/21 Resolved: 19/Oct/21	56
[AMS-31] Set up elastic search v docker container Created: 13/Oct/21 Resolved: 16/Oct/21	57
[AMS-34] Vytvoriť github projekt Created: 13/Oct/21 Resolved: 18/Oct/21	58

[AMS-13] Zistit, ktore miestnosti su na karticky Created: 05/Oct/21 Resolved: 18/Oct/21	59
[AMS-4] Zaobstarat stroj v škole Created: 01/Oct/21 Updated: 05/Oct/21 Resolved: 13/Oct/21	60
[AMS-10] Revizia poziadaviek Created: 05/Oct/21 Resolved: 13/Oct/21	61
[AMS-15] Realny scrapping - vytvorenie vzorky Created: 05/Oct/21 Resolved: 10/Oct/21	62
[AMS-20] Analyza parametrov Google News Created: 05/Oct/21 Resolved: 13/Oct/21 ...	64
[AMS-21] Vytvorit Slack server Created: 05/Oct/21 Resolved: 11/Oct/21	65
[AMS-3] Zoznam trestnych cinov, ktore budu pouzite ako queries na google news Created: 01/Oct/21 Resolved: 04/Oct/21	66
[AMS-5] Vysoka architektura Created: 01/Oct/21 Resolved: 03/Oct/21	67
[AMS-6] Specifikacia poziadaviek Created: 01/Oct/21 Resolved: 03/Oct/21.....	68
[AMS-7] Prieskum ziskavania dat, praca s kniznicou Created: 01/Oct/21 Resolved: 05/Oct/21	69
[AMS-8] Pripady pouzitia Created: 01/Oct/21 Resolved: 05/Oct/21	70

IN PROGRESS

[AMS-89] Integrácia scraperu a vylepšeného parseru <small>Created: 09/Nov/21</small>	
Status:	In Progress
Project:	Adverse Media Screening
Type:	Story
Assignee:	Jakub Müller
Original estimate:	5 hours
Sprint:	AMS Sprint 2
Story point estimate:	5

[AMS-93] [Analýza možností implementácie testovania](#) Created: 09/Nov/21

Status:	In Progress
Project:	Adverse Media Screening

Type:	Story
Assignee:	Adam Šípka
Original estimate:	5 hours

Sprint:	AMS Sprint 2
Story point estimate:	5

[AMS-11] [Lofi používateľského rozhrania](#) Created: 05/Oct/21 Updated: 12/Oct/21

Status:	In Progress
Project:	Adverse Media Screening

Type:	Story
Assignee:	David Silady
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Okrem Dávida bude na tomto pracovať aj Jakub Muller.

<https://www.figma.com/file/zouHpeljsUKIQPuKYDeRB2/Lo-Fi-Prototypes?node-id=0%3A1>

Poznámky: je to ok, ale story ostáva rozpracovaná, keďže sa ešte LoFi môže zmeniť príchodom rôznych funkcionalít.

[AMS-36] [User databáza + zbieranie aktivity používateľa](#) Created: 13/Oct/21 Updated: 17/Oct/21

Status:	In Progress
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Time Spent:	Not Specified
Original estimate:	Not Specified

Attachments:	 Conceptual Model.bmp  Tables.bmp
---------------------	---

Comments

Comment by [Táňa Poláková](#) [17/Oct/21]

Modely su v obrazkoch.

Poznámka: nevenujeme sa tejto story zatiaľ, pretože ju momentálne nepovažujeme za až takú dôležitú.

TO DO

[AMS-102] [Kontajnerizácia scrapera](#) Created: 15/Nov/21

Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story
Assignee:	Unassigned
Original estimate:	Not Specified

Sprint:	
----------------	--

Description

+ systémové premenné na ovládanie kontajnera

[AMS-103] [Rozšírenie docker compose o MongoDB a Elasticsearch](#) Created: 15/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Hlavačka
Original estimate:	5 hours

Sub-tasks:	Key	Summary	Type	Status	Assignee
	AMS-104	Zabezpečenie MongoDB a Elasticsearch	Subtask	To Do	Jakub Hlavačka
	AMS-105	Analyza River plugin	Subtask	To Do	Jakub Hlavačka
	AMS-106	Analyza shard, nodes	Subtask	To Do	Jakub Hlavačka

Sprint:	AMS Sprint 2
Story point estimate:	5

[AMS-92] [Zabezpečenie komunikácie medzi klientom a serverom](#) Created: 09/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	David Silady
Time Spent:	Not Specified
Original estimate:	8 hours

Sprint:	AMS Sprint 2
Story point estimate:	8

Description

Obalený fetch.

[AMS-91] Úprava dokumentácie zo stretnutí a šprintov tak, aby mohli ísť na stránku tímu Created: 09/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Time Spent:	Not Specified
Original estimate:	8 hours

Sprint:	AMS Sprint 2
Story point estimate:	8

Description

Denníky zo stretnutí, upratanie šprintov, exporthy úloh, šprint review, ...

[AMS-67] [Pozriet' sa na zahodené články, kt. neobsahovali paragrafy](#) Created:
02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sprint:	
----------------	--

[AMS-68] [Analýza médií, ktoré dostaneme vo výstupoch a aké search queries treba poslať, aby sme mali výsledky ako z Google](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sprint:	
----------------	--

[AMS-72] [Vytvorenie developerského prostredia](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sprint:	
----------------	--

[AMS-86] [PDF report z nájdených článkov](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Description
Zatiaľ v backlogu.

[AMS-85] [Prihlásenie sa na odber nejakého vyhľadaného človeka](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sprint:	
----------------	--

Description

vždy keď sa pridá nový článok o hľadanom človeku. V backlogu.

[AMS-84] [Označenie článku ako videného \(prečítaného\)](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sprint:	
----------------	--

Description	V backlogu.
--------------------	-------------

[AMS-83] [Hodnotenie risku pri najdených článkoch](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sprint:	
----------------	--

Description

Používateľ si ohodnotí, aký risk má daný článok + nejaká poznámka k tomu. V backlogu.

[AMS-82] [Zvýraznenie hľadaných slovíčok](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sprint:	
----------------	--

Description	V backlogu.
--------------------	-------------

[AMS-77] [Filter na web stránke](#) Created: 02/Nov/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned

Sub-tasks:	Key	Summary	Type	Status	Assignee
	AMS-78	obmedzenie vyhľadávania na základe dá...	Subtask	To Do	
	AMS-79	vyhľadanie na základe zločinov	Subtask	To Do	
	AMS-80	geografická lokácia	Subtask	To Do	
	AMS-81	boolean query	Subtask	To Do	
Sprint:					

[AMS-42] Porovnanie výsledkov nášho vyhľadávania s výsledkami Google News

Created: 13/Oct/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned
Time Spent:	Not Specified
Original estimate:	Not Specified

[AMS-12] [Nasu API dat na portal verejnych API v ramci propagacie](#) Created: 05/Oct/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Pridané do backlogu.

[AMS-41] Vyriešiť, čo s článkami, ktoré budú v pôvodnom zdroji nekompletné

Created: 13/Oct/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Články môžu byť vymazané, prípadne iba ich obrázky. Návrhy: vytvoriť pdf, screenshot, ...

[AMS-37] Zistiť, ako fungujú iné zdroje Google News, prípadne aký mechanizmus používa Google News. Created: 13/Oct/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Adam Šípka
Time Spent:	Not Specified
Original estimate:	Not Specified

[AMS-33] [Zistiť, čo spája kriminálne články](#) Created: 13/Oct/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Napríklad aké frázy a slová sa nachádzajú v dokumentoch o podozrivých osobách a v iných článkoch nie (slovo "commit").

[AMS-19] [Analyza - multilingvistika](#) Created: 05/Oct/21

Status:	To Do
Project:	Adverse Media Screening

Type:	Story		
Assignee:	Unassigned		
Σ Remaining Estimate:	Not Specified	Remaining Estimate:	Not Specified
Σ Time Spent:	Not Specified	Time Spent:	Not Specified
Σ Original Estimate:	Not Specified	Original estimate:	Not Specified

Sub-tasks:	Key	Summary	Type	Status	Assignee
	AMS-38	Zistiť, ako funguje Google Translate...	Subtask	To Do	Adam Šípka
	AMS-40	inteligentný preklad slovíčok do slov...	Subtask	To Do	
	AMS-76	Iné možnosti prekladu článkov	Subtask	To Do	

Description

API na Google translate, preklad klucovych slov. Zatiaľ v backlogu.

[AMS-71] [Naplnanie MongoDB pri scrapovani](#) Created: 02/Nov/21 Resolved: 14/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Adam Šipka
Time Spent:	1 day
Original estimate:	8 hours
Time Spent:	8 hours

Sprint:	AMS Sprint 1, AMS Sprint 2
Story point estimate:	8

Description



Fields: meno, link, datum vydania, region, jazyk, telo clanku

DONE

[AMS-60] [Pamät'ová zložitosť elastic search na vzorke dát](#) Created: 01/Nov/21 Resolved: 12/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Hlavačka
Original estimate:	6 hours
Time Spent:	6 hours 30 minutes

Attachments:	 image-20211108-145633.png  main.py
Sprint:	AMS Sprint 1, AMS Sprint 2
Story point estimate:	6

Description

Zobrat dump od [Adam Šípka](#) a spravit index v elastic search, pricom jednotlivé values pre terms budu ID zo zaznamov v MongoDB

Comments

Comment by [Jakub Hlavačka](#) [07/Nov/21]

Neskôr by bolo vhodné použiť asi toto <https://hevo.com/learn/integrating-elasticsearch-and-mongodb/>

Comment by [Jakub Hlavačka](#) [08/Nov/21]

Mam zatiaľ vytvorený docker-compose, kde je mongodb a elasticsearch s ich vlastnými volumes. Mongo má collection articles, ktorá je komprimovaná a obsahuje všetky články z articles_1.zip . K tasku príkladám script, pomocou ktorého som čítal z mongodb a snažil sa vytvárať index v elasticsearch z článkov.

Posting list sa mi zatiaľ nepodarilo naimplementovať, pretože elasticsearch analyzer má obmedzenie pre tokenizovanie na 10000 slov...

Aby som aspoň hrubým odhadom zistil pamätovú zložitosť vytvoreného indexu v elasticsearch, tak som vkladal do dokumentu v indexe celý html body tag, pričom jeho kľúč bola id z mongodb... Pricom po nejakom 650 článku vybehol nasledovný error

Navrhujem dalsi postup:

- vyriesit problem s obmedzenim tokenizacie
- vytvarat uz rovno index (posting list), ktory budeme vyuzivat v AMS
- vytvorit dalsi task na analyzu nastroja z <https://hevodata.com/learn/integrating-elasticsearch-and-mongodb/>
- Alokovat viac casu !!!

EDIT: pridanie navrhu na dalsi tak s nastrojom.

Comment by [Jakub Hlavačka](#) [12/Nov/21]

Z 308MB clankov spravil elasticsearch 244MB.

[AMS-87] [Zabrániť zacykleným buildom](#) Created: 07/Nov/21 Resolved: 15/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Dominik Horvath
Original estimate:	5 hours
Time spent:	1 hour

Sprint:	AMS Sprint 2
Story point estimate:	5

Description

Github actions timeout setting aby nam nebezali buildy zbytocne dlho

[AMS-88] [Rozšířit úložisko pomocou nepriradeného disku](#) Created: 09/Nov/21 Resolved:
14/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Dominik Horvath
Original estimate:	5 hours
Time Spent:	2 hours

Sprint:	AMS Sprint 2
Story point estimate:	5

Comments

Comment by [Dominik Horvath](#) [14/Nov/21]

Disk is mounted under **/data**

[AMS-90] [Integrácia MongoDB a Elasticsearch](#) Created: 09/Nov/21 Resolved: 16/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Hlavačka
Original estimate:	8 hours
Time Spent:	9 hours

Sprint:	AMS Sprint 2
Story point estimate:	8

Description

výskum možností prepojenia, kontrola nastavení - aby neukladal elastic text, iba ID.

Comments

Comment by [Jakub Hlavačka](#) [12/Nov/21]

Neskor potrebne zabezpecenie elasticsearch <https://www.elastic.co/guide/en/elastic-stack-get-started/current/get-started-docker.html>

Comment by [Jakub Hlavačka](#) [12/Nov/21]

Skonzultovat skalovanie. (Pocet shards, pocet nodes, ...).

Comment by [Jakub Hlavačka](#) [12/Nov/21]

Pozriet sa na <https://hevodata.com/learn/integrating-elasticsearch-and-mongodb/>

Comment by [Jakub Hlavačka](#) [12/Nov/21]

https://github.com/FIIT-TEAM8/elasticsearch_mongo

Comment by [Jakub Hlavačka](#) [16/Nov/21]

Casovane meranie: total time in **seconds: 804.9638478755951**, cize cca **13 minut**.




Comment by [Jakub Hlavačka](#) [16/Nov/21]

Elasticsearch vyuziva nejaku jednoduchsiu kompresiu uz defaulte, nasiel som, ze podporuje este jeden sposob, lenze ten zredukuje ulozisko o nejake ~3%... Samotny index ma 126MB z 308MB dat, ked obsahuje jeden shard a jednu repliku. Tym padom ma jeden shard, v ktorom sa drzi LUCENE index cca ~60MB

[AMS-94] [Návrh Flask API](#) Created: 09/Nov/21 Resolved: 15/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Original estimate:	5 hours
Time Spent:	6 hours

Attachments:	 image-20211114-140311.png  image-20211114-142548.png  image-20211114-142800.png
Issue links:	Blocks blocks AMS-95 Implementácia dummy Flask API Done
Sprint:	AMS Sprint 2
Story point estimate:	5

Description

Nástroj na dokumentáciu + dokumentácia.

Comments

Comment by [Jakub Hlavačka](#) [10/Nov/21]

[Táňa Poláková](#) <https://betterprogramming.pub/restful-api-design-step-by-step-guide-2f2c9f9fcd9f>

Comment by [Táňa Poláková](#) [14/Nov/21]

Versioning: robi sa kvoli supportu starsich zakaznikov v pripade, ze sa nejake poziadavky zmenia a my spravime updates v nasej API.

Existuju 4 strategie versioningu:

1. URI path versioning - /api/v1/Customers/1 ... /api/v2/Customers/1
2. URL parameter versioning - /api/Customers?v=1.1
3. Content Negotiation - URLs sa nemenia, v accept header vidime, ktora verzia je ta, ktoru chceme:
 1. Accept: application/app.v1.categories
 2. Accept: application/app.v2.categories
4. Custom Header:
 1. x-App-version: 1.3

2. x-App-version: 2017-08-12

Continuous versioning - neexistuju ziadne verzie, iba json responses vzdy nejak prisposobime, aby obsahovalo polia, ktore zakaznik ocakava, ale aj nove, ktore sme pridali na zaklade zmenenych poziadaviek (viac: <https://www.youtube.com/watch?v=M2KCu0Oq3JE>).

Comment by [Táňa Poláková](#) [14/Nov/21]

Dokumentáciu je vhodné písať/generovať, až keď sú requesty implementované.

Postman

Vyhody

- intuitívne grafické rozhranie dostupné aj cez prehliadač
- podpora kolaborácie
- testovanie API

Nevyhody

- výsledná dokumentácia je dostupná cez postman URL a nie cez našu vlastnú - to sa dá len pri PRO verzii
- response json sa neda dokumentovať

Ukazka dokumentacie

Ako funguje

1. nainštalujeme si Postman (<https://www.postman.com/downloads/>)
2. do tímového workspace sa vieme dostať cez invite link dostupný na našom drive
3. vytvoríme si collection, kam budú spadať naše implementované requesty, môžeme vytvoriť aj viac kolekcii ak nájdeme viac logických celkov
4. naša Collection je vlastne súbor našich requestov, ktorú vieme ako celok zdokumentovať - vieme jej dať meno, description. Vieme vytvoriť viac Collections - nám asi bude stačiť len Collection "Articles"
5. v rámci kolekcie si vytvoríme všetky implementované requesty - rozdelené aj podľa metód
6. pri každom requeste vieme pridať description, query parametre, uložiť, ako vyzerá príklad json response

Flask-RESTPlus

Rozšírenie pre Flask, dostupné cez pip: pip install flask-restplus

<https://www.imaginarycloud.com/blog/flask-python/>

Vyhody

- väčšinu si naprogramujeme sami

- dokumentacia moze byt na nasej domene
- kolaboracia je mozna vlastne len tak, ze pristupujeme k spolocnemu kodu
- response json sa da dokumentovat

Nevyhody

- vacsinu si naprogramujeme sami
- nie je graficke rozhranie
- vysledny vzhľad dokumentacie sa mi nepaci (😞)
- testujeme iba cez konzolu

Ukazka kodu

Ukazka dokumentacie

https://flask-ic.herokuapp.com/documented_api/doc

[AMS-95] [Implementácia dummy Flask API](#) Created: 09/Nov/21 Resolved: 15/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Dominik Horvath
Original estimate:	5 hours
Time Spent:	5 hours

Issue links:	Blocks
	is blocked by AMS-94 Návrh Flask API Done
Sprint:	AMS Sprint 2
Story point estimate:	5

Comments



Comment by [Dominik Horvath](#) [15/Nov/21]

Dummy flask api hotova. Program momentalne vracia nahodny pocet clankov, vsetko jednoduche lorem-ipsum. Jednoducha dokumentacia je k dispozicii v README repozitaru [flask_serveru](#).

[AMS-46] [Parsovanie tagov pomocou CSS Selector](#) Created: 27/Oct/21 Resolved: 08/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Müller
Original estimate:	5 hours
Time Spent:	6 hours

Attachments:	 example.html  scrapy_xpath_meurments.txt.txt										
Sub-tasks:	<table border="1"><thead><tr><th>Key</th><th>Summary</th><th>Type</th><th>Status</th><th>Assignee</th></tr></thead><tbody><tr><td>AMS-74</td><td>Sekvenčné scrapovanie</td><td>Subtask</td><td>Done</td><td>Dominik Horvath</td></tr></tbody></table>	Key	Summary	Type	Status	Assignee	AMS-74	Sekvenčné scrapovanie	Subtask	Done	Dominik Horvath
Key	Summary	Type	Status	Assignee							
AMS-74	Sekvenčné scrapovanie	Subtask	Done	Dominik Horvath							
Sprint:	AMS Sprint 1										
Story point estimate:	5										

Description

Pomocou CSS Selector je potrebné vyskúšať parsovať tagy, ktoré obsahujú text

Comments

Comment by [Jakub Müller](#) [08/Nov/21]

Pri analýze CSS selektora z knižnice *scrapy* sa ukázalo, že pre náš účel nebude dostačujúci. Je to z toho dôvodu, že tento selektor dokáže selektovať iba jednotlivé tagy zvlášť a my potrebujeme všetky tagy, ktoré obsahujú text (teda <p> a všetky <h> tagy) a v pôvodnom poradí.

Preto sme sa namiesto CSS selektora z knižnice *scrapy* zamerali na selektor *XPATH* z tejto knižnice.

Tento selektor dokáže vyselektovať všetky tagy s textom (teda paragrafy a headingy) v takom poradí, v akom sú uvedené v pôvodnom HTML súbore.

Výsledný parser bol testovaný na vzorke dát `1_articles.jl`, ktorá má cca 3.7GB a obsahuje cca 27 000 článkov. Celú túto vzorku dát dokázal parser spracovať za necelé 3 minúty, pričom dokázal znížiť veľkosť výsledného súboru o cca 92%.

Detailnejšie výsledky parseru sú uvedené v prílohe **scrapy_xpath_meurments.txt** a príklad výsledného HTML súboru, z ktorého bol vyselektovaný iba text je v prílohe **example.html**

[AMS-74] [Sekvenčné scrapovanie](#) Created: 02/Nov/21 Resolved: 06/Nov/21

Status:	Done
Project:	Adverse Media Screening

Assignee:	Dominik Horvath
Type:	Subtask
Time Spent:	7 hours
Original estimate:	7 hours

Attachments:	 data.zip  timing.txt
Sprint:	AMS Sprint 1

Comments

Comment by [Dominik Horvath](#) [06/Nov/21]

Sekvenčne scrapovanie bolo skusane pomocou viacerých prístupov, a to BeautifulSoup knižnica, LXML knižnica spolu s [XPATH |<https://www.w3.org/TR/1999/REC-xpath-19991116/>] selector standardom a taktiež [python-fork reader-view](#) pluginu obsiahnutom v prehliadaci firefox.

BeautifulSoup knižnica nebola dostacujuca. Na jednu stranu bola velmi pomala + vyžadovala zložity kod na extrahovanie len tých atributov, ktoré su pre nas zaujimave.

Fork readerview pluginu fungoval perfektne, jeho vytvorene vzorky fungovali 99% pripadov tak ako v prehliadaci, avsak bol velmi pomaly - datovu sadu 1_articles.zip spracovaval asi 4 hodiny, kde sa nedostal ani do polovice clankov. Nasledne som program vypol.

LXML knižnica spolu s XPATH selectorom fungovala najlepsie. Bola vytvorena tato vyhľadavacia query:

```
"//p | //h1 | //h2 | //h3 | //h4 | //h5 | //h6"
```

ktora zoberie vsetky stanovene HTML tagy v povodnom poradi, ako boli v dokumente. Knižnica bola velmi rychla a redukcia dat signifikatna (priloha timing.txt)

V prilohu **data.zip** sa nachadza spracovanych prvych 100 zaznamov z datovej sady 1_articles.zip a taktiež zdrojovy kod. V priecinku **parsed** sa nachadza 100 HTML suborov, jeden pre kazdu spracovanu stranku. V subore **articles_sample_links.txt** sa pre moznost kontroly nachadzaju odkazy na povodne clanky. HTML subory maju v nazve cislo riadku, na ktorom sa v subore **articles_sample_links.txt** nachadza odkaz k povodnemu clanku.

[Jakub Müller](#) [Táňa Poláková](#) FYI

[AMS-57] [Zavesenie klienta](#) Created: 27/Oct/21 Resolved: 07/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	David Silady
Original estimate:	13 hours
Time spent:	11 hours, 30 minutes

Sprint:	AMS Sprint 1
Story point estimate:	13






Description

Zavesit' klienta bud' cez Express alebo priamo cez NGINX. Testovacie API calls, redux, ...

[AMS-59] [Parsovanie tagov pomocou lxml, readability, trafiletura](#) Created: 30/Oct/21
Resolved: 09/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Original estimate:	3 hours
Time Spent:	6 hours

Attachments:	 2_sk_article.html  lxml_measurement - Copy.txt  p_lxml_measurement.txt  readability_measurement.txt  trafiletura_measurement.txt
Sprint:	AMS Sprint 1
Story point estimate:	3

Description

pozriet aj ine jazyky, pozriet dokumentaciu, preco kniznica suvisi s jazykom

Comments

Comment by [Táňa Poláková](#) [07/Nov/21]

Readability - funguje iba na zaklade nazvov tagov a tried, nie na zaklade jazyka textu - funkcia summary() extrahuje hlavny obsah HTML stranky. Ako prve vymaze vsetky script alebo style tagy. Nasledne maza tagy na zaklade regexov:

```
{ {'unlikelyCandidatesRe':  
re.compile('combx|comment|community|disqus|extra|foot|header|menu|remark|rss|shoutbox|sideb  
ar|sponsor|ad-break|agegate|pagination|pager|popup|tweet|twitter',re.I),  
'okMaybeItsACandidateRe': re.compile('and|article|body|column|main|shadow',re.I), } }  
  
if (  
    REGEXES["unlikelyCandidatesRe"].search(s)  
    and (not REGEXES["okMaybeItsACandidateRe"].search(s))  
    and elem.tag not in ["html", "body"]  
) :  
    log.debug("Removing unlikely candidate - %s" % describe(elem))  
    elem.drop_tree()
```

Potom pridava score tagom na zaklade urcnych kriterii:


```

def class_weight(self, e):
    weight = 0
    for feature in [e.get("class", None), e.get("id", None)]:
        if feature:
            if REGEXES["negativeRe"].search(feature):
                weight -= 25

            if REGEXES["positiveRe"].search(feature):
                weight += 25

            if self.positive_keywords and
self.positive_keywords.search(feature):
                weight += 25

            if self.negative_keywords and
self.negative_keywords.search(feature):
                weight -= 25

            if self.positive_keywords and self.positive_keywords.match("tag-" +
e.tag):
                weight += 25

            if self.negative_keywords and self.negative_keywords.match("tag-" +
e.tag):
                weight -= 25

    return weight
def score_node(self, elem):
    content_score = self.class_weight(elem)
    name = elem.tag.lower()
    if name in ["div", "article"]:
        content_score += 5
    elif name in ["pre", "td", "blockquote"]:
        content_score += 3
    elif name in ["address", "ol", "ul", "dl", "dd", "dt", "li", "form",
"aside"]:
        content_score -= 3
    elif name in [
        "h1",
        "h2",
        "h3",
        "h4",
        "h5",
        "h6",
        "th",
        "header",
        "footer",
        "nav",
    ]:
        content_score -= 5
    return {"content_score": content_score, "elem": elem}

```

Vyberie sa najlepsí kandidát na základe score a potom sa hľadajú susedné elementy, ktoré by s najlepšími kandidátmi mohli súvisieť.

Comment by [Táňa Poláková](#) [08/Nov/21]

Readability - pri 20tich clankoch nesedel jeden extrahovany s povodnym clankom. Extrahovany hlavny obsah bol nezmysel - iba svg obrazky.

V kode nemaju velku vahu headings, co by bolo mozne upravit pre nase potreby. Vseobecne mi ale pride, ze v kode je bordel - vela zakomentovanych veci.

Comment by [Táňa Poláková](#) [08/Nov/21]

Trafilatura - vsetkych 20 clankov obsahovalo hlavny content - aj ten, ktory readability nezvladol. Zaujímavostou bolo, ze v kode pouzili readability a na zaklade dlzky extrahovaneho textu sa rozhodli, ci pouziju svoj alebo readability algoritmus.

Rozdiel je, ze trafiletura vracia iba holy text, bez tagov.

Comment by [Táňa Poláková](#) [08/Nov/21]

Readability funguje rovnako pre anglicke aj slovenske clanky.

Priklad:

(zdroj: <https://www.cas.sk/clanok/2603841/brutalna-dvojnashobna-vrazda-v-lednickych-rovniach-kedy-sa-zacne-proces-s-obzalovanim/>)

[2_sk_article.html](#)

Trafiletura tiez.

[AMS-73] [Vyhládanie RSS dvojčat'a k HTML článku - všeobecný postup](#) Created:
02/Nov/21 Resolved: 09/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Dominik Horvath
Original estimate:	8 hours
Time Spent:	4 hours

Sprint:	AMS Sprint 1
Story point estimate:	8

[AMS-70] [Rozšírenie google news scrapperu na lokáciu UK](#) Created: 02/Nov/21 Resolved: 02/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Dominik Horvath
Original estimate:	1 hour
Time Spent:	1 hour

Sprint:	AMS Sprint 1
Story point estimate:	1

Comments

Comment by [Dominik Horvath](#) [02/Nov/21]

Scrapper je pripraveny na parsovanie clankov z lokacie UK, v jazyku anglictina. Staci inicializovat triedu na parsovanie takto:

```
gnews_parser = GnewsParser()  
gnews_parser.setup_search("covid", '2021-09-01', '2021-09-02', locale="en-gb")
```

[AMS-51] [Pridat' info o projekte a opisat' clenov timu](#) Created: 27/Oct/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Müller
Original estimate:	4 hours
Time Spent:	1 hour

Sprint:	AMS Sprint 1
Story point estimate:	4

[AMS-50] [Request URL s odpoved'ou inou ako 200 uložit' do súboru](#) Created:
27/Oct/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Adam Šípka
Original estimate:	3 hours
Time Spent:	1 hour

Sprint:	AMS Sprint 1
Story point estimate:	3



Description

Čokoľvek, čo počas scrapovania dostane odpoveď inú ako 200 je potrebné uložiť do súboru. Okrem URL treba uložiť aj zločin.

[AMS-49] [Lokálna MongoDB so získanými dátami](#) Created: 27/Oct/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Adam Šípka
Original estimate:	8 hours
Time Spent:	5 hours

Attachments:	 mongodb zistenia_01.txt  mongodb zistenia_02.txt
Sprint:	AMS Sprint 1
Story point estimate:	8

Description

Naše získané dáta (všetky fields) dať do lokálnej MongoDB, zistiť, aké sú možnosti práce s ňou, koľko miesta zaberá na disku - uvidíme, ako naše dáta Mongo zredukuje. Zistiť, ako je tvorené ID.

Po porovnaní nám vyšlo, že sa to oplatí viac ako metóda ZIP

[AMS-48] [Analýza bezstratovej kompresie textu](#) Created: 27/Oct/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Hlavačka
Original estimate:	8 hours
Time Spent:	6 hours

Sub-tasks:	Key	Summary	Type	Status	Assignee
	AMS-61	bezstratova kompresia text	Subtask	Done	
	AMS-62	analyza vyhladavania nad komprimovany...	Subtask	Done	
	AMS-63	elastic search a komprimacia	Subtask	Done	
	AMS-64	realne kniznice	Subtask	Done	
	AMS-65	vyskusat nad datach	Subtask	Done	
	AMS-66	Ako mongoDB pracuje s vopred komprimo...	Subtask	Done	
Sprint:	AMS Sprint 1				
Story point estimate:	8				

Description

Analýza bezstratovej kompresie text; analýza vyhládavania nad komprimovaným textom; elastic search a komprimácia; encoding tetxtu; reálne knižnice; vyskúšať na dátach; ako MongoDB pracuje s vopred komprimovaným textom.

Poznámka vedúceho: úloha je splnená, jej podúloha bola preradená do novej story z dôvodu, že priamo nesúvisela s touto Story. Bola testovaná metóda ZIP

[AMS-47] [Parsovanie tagov pomocou regex](#) Created: 27/Oct/21 Resolved: 30/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Original estimate:	8 hours
Time Spent:	2 hours

Sprint:	AMS Sprint 1
Story point estimate:	8

Description

Pomocou regexov je potrebné vyskúšať rozparsovať tagy, ktoré obsahujú text. Dôležité je, aby mali správne poradie.

Comments

Comment by [Táňa Poláková](#) [30/Oct/21]

Analýza ukázala, že nie je efektívne pracovať s regexami. Úloha sa ukončila a nahradila ju nová <https://tim8-2021.atlassian.net/browse/AMS-59>

[AMS-53] [Rozšíriť docker compose](#) Created: 27/Oct/21 Resolved: 31/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Dominik Horvath
Original estimate:	13 hours
Time Spent:	7 hours

Sub-tasks:	Key	Summary	Type	Status	Assignee
	AMS-54	Kontajner Flask API	Subtask	Done	
	AMS-55	Kontajner Express server	Subtask	Done	
	AMS-56	NGINX konfigurácia	Subtask	Done	
Sprint:	AMS Sprint 1				
Story point estimate:	13				

Comments

Comment by [Dominik Horvath](#) [31/Oct/21]

Vytvorene 2 github repozitare: [node server](#) a [flask server](#). Oba repozitare obsahuju github actions procedury nakonfigurovane tak, aby sa spustili pri pushnuti / mergnuti do main branche.

Github actions vykonaju build kontajneru z aktualnej main branch, nahra hotovy image na docker-hub a nasledne nacita najnovsie docker image aj na timovy virtualny stroj, kde ich rovno spusti a nahradi stare kontajneru.

Nove cesty v nginx reverse proxy:

- /api => flask server
- /ams => node server

[AMS-32] [Prihláška na TP CUP](#) Created: 13/Oct/21 Resolved: 02/Nov/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Unassigned
Time Spent:	Not Specified
Original estimate:	Not Specified

[AMS-44] [Stránka tímu](#) Created: 18/Oct/21 Resolved: 01/Nov/21


Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Müller
Time Spent:	Not Specified
Original estimate:	Not Specified

[AMS-29] [Získanie dát pre prototyp](#) Created: 13/Oct/21 Resolved: 26/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story		
Assignee:	Dominik Horvath		
Σ Remaining Estimate:	Not Specified	Remaining Estimate:	Not Specified
Σ Time Spent:	Not Specified	Time Spent:	Not Specified
Σ Original Estimate:	Not Specified	Original estimate:	Not Specified

Attachments:	 image-20211019-151711.png				
Sub-tasks:	Key	Summary	Type	Status	Assignee
	AMS-30	Redukcia a kategorizacia zoznamu zloč...	Subtask	To Do	Dominik Horvath

Comments

Comment by [Jakub Hlavačka](#) [19/Oct/21]

<https://github.com/JKBGIT1/scraper>

V readme je sposob akym to spustit.

Treba osetrit veci, ktore su v komentoch spider.py ako TODO

Comment by [Jakub Hlavačka](#) [19/Oct/21]

<attachment>

Asi to ide moc rychlo...

Comment by [Jakub Hlavačka](#) [19/Oct/21]

Zjavne bude potrebne zabezpecit, aby to nerobilo requesty na server, kde uz dostal timeout...

[AMS-43] [Otvorenie portov na serveri](#) Created: 18/Oct/21 Resolved: 18/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	David Silady
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Otvorené:
443, 80, 22 (asi aj 53)
Je lepšie ostatné ani neotvárať. (Pochybujem, že by to vôbec šlo bez ďalších komplikácií - Docker)

[AMS-35] [Prieskum databáz](#) Created: 13/Oct/21 Resolved: 18/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Müller
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Všetko v containeroch. Analýza MongoDB, ... kompatibilita s elastic search

[AMS-39] Vyriešiť, aby fiitkar nemohol urobiť "sudo su" na virtuálnom stroji (aby nemal šancu sa zmeniť na roota) Created: 13/Oct/21 Resolved: 19/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

a nech sa ani nevie prepnut ani na ubuntu

Comments

Comment by [Táňa Poláková](#) [17/Oct/21]
<https://www.thegeekdiary.com/how-to-disable-sudo-su-for-users-in-sudoers-configuration-file/>

[AMS-31] [Set up elastic search v docker container](#) Created: 13/Oct/21 Resolved: 16/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Hlavačka
Time Spent:	Not Specified
Original estimate:	Not Specified

Attachments:	 docker-compose.yml  image-20211016-144056.png
---------------------	--

Comment by [Jakub Hlavačka](#) [16/Oct/21]

<attachment>

Pre spustenie je potrebne byt v adresari /home/fiitkar/docker-folder a odpalit command: *sudo docker-compose up -d*

Prepinac -d zabezpeci, ze sa proces spusti na pozadi.

Command na zastavenie a zmazanie docker kontajneru: *sudo docker-compose down*

Nerobit: *sudo docker-compose down -v* , pretoze to zmaze volume, v ktorom su ulozene naindexovane data.

Comment by [Jakub Hlavačka](#) [16/Oct/21]

[Dominik Horvath](#) mozes indexovat data na masine.

[AMS-34] [Vytvorit' github projekt](#) Created: 13/Oct/21 Resolved: 18/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Müller
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Viac repozitárov pod jedným projektom

[AMS-13] [Zistit, ktore miestnosti su na karticky](#) Created: 05/Oct/21 Resolved: 18/Oct/21

Status:	Done
Project:	Adverse Media Screening


Type:	Story
Assignee:	David Silady
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

6-te poschodie - coworking.
Treba sa 5 krát pipnuť a napísať pánovi (neviem komu).
Je tam stále restricted access - treba poznať správnych ľudí.
Vraj sa tam na tím FIIT-WIX škaredo pozerali.
Netreba tam mať rúško?

[AMS-4] [Zaobstarat stroj v škole](#) Created: 01/Oct/21 Updated: 05/Oct/21 Resolved: 13/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Time Spent:	Not Specified
Original estimate:	Not Specified
Attachments:	 virtualne-stroje-info2021.txt

Description

- VM na fiit(zdarma)
- 1GB Ram, 1CPU, min 50 GB - 100GB disk
- nie FreeBSD, ale skor Ubuntu

Comments

Comment by [Táňa Poláková](#) [03/Oct/21]

Kontaktovala som Ing. Juraja Petríka (5731@is.stuba.sk), ktorý by podľa úvodnej prezentácie mal mať na starosti virtuálne stroje.

Comment by [Táňa Poláková](#) [05/Oct/21]

[virtualne-stroje-info2021.txt](#)

Žiadosť o SSH key bola odoslaná na meno polakova18

[AMS-10] [Revizia poziadaviek](#) Created: 05/Oct/21 Resolved: 13/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Zpracovať poznámky od vedúceho zo story [AMS-6].

Schválené vedúcim, ale aj tak sa môžu ešte v priebehu vývoja softvéru meniť.

[AMS-15] [Realny scrapping - vytvorenie vzorky](#) Created: 05/Oct/21 Resolved: 10/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story		
Assignee:	Dominik Horvath		
Σ Remaining Estimate:	Not Specified	Remaining Estimate:	Not Specified
Σ Time Spent:	Not Specified	Time Spent:	Not Specified
Σ Original Estimate:	Not Specified	Original estimate:	Not Specified

Sub-tasks:	Key	Summary	Type	Status	Assignee
	AMS-16	Ziskanie vzorky RSS	Subtask	Done	Jakub Hlavačka
	AMS-17	API / scrapovanie	Subtask	Done	Dominik Horvath

Description

Ziskanie vzorky: Jeffrey epstein, poslednych 5 rokov.

Subtasks:

[AMS-16] [Ziskanie vzorky RSS](#) Created: 05/Oct/21 Updated: 08/Oct/21 Resolved: 08/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Subtask
Assignee:	Jakub Hlavačka
Time Spent:	Not Specified
Original estimate:	Not Specified

Attachments:	google_news_rss.py
---------------------	--------------------

Description

ziskanie RSS a zistenie, ci je to efektivne alebo nie

+ casovanie vsetkeho, co sa vykona

Comments

Comment by [Jakub Hlavačka](#) [08/Oct/21]

RSS vrati 100 clankov na danu query.

Total time: 0.046875

Malo by to byt v sekundach.

Ak sa vyberieme touto cestou, tak bude zjavne potrebne sa dostat k dalsim clankom a nie len prvym 100, ktore to vyhodi...

[AMS-17] [API / scrapovanie](#) Created: 05/Oct/21 Resolved: 10/Oct/21

Status: Done

Project: [Adverse Media Screening](#)

Type: Subtask

Assignee: [Dominik Horvath](#)

Time Spent: Not Specified

Original estimate: Not Specified

Attachments: output.csv

Comments

Comment by [Dominik Horvath](#) [10/Oct/21]

Vytvoreny custom parser: <https://github.com/Dominik1799/gnewsParser>

Vsetky kniznice, ktore sme nasli, pracovali len ako wrapper nad RSS streamom. Vyuzivali funkcie, ktore su pre nas problem zbytocne, a boli preto velmi pomale.

Vlastny parser uskutocnil 1827 requestov a ziskal 35772 zaznamov za 93 sekund processing time a cca 20 minut realneho casu. Vyhľadavacia query ziskala vsetky clanky v casovom období od 10.10.2016 do 10.10.2021 na klucove slova "Jeffrey Epstein".

[AMS-20] [Analyza parametrov Google News](#) Created: 05/Oct/21 Resolved: 13/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Adam Šípka
Time Spent:	Not Specified
Original estimate:	Not Specified

Comments

Comment by [Adam Šípka](#) [13/Oct/21]
<http://books.gigatux.nl/mirror/googlehacks/0596008570/googlehks2-CHP-4-SECT-3.html>

[AMS-21] [Vytvorit Slack server](#) Created: 05/Oct/21 Resolved: 11/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Táňa Poláková
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Pre efektívnejšiu komunikáciu a strenutia v online priestore.

Comments

Comment by [Táňa Poláková](#) [11/Oct/21]

Vytvorila som Team na platforme Microsoft Teams

[AMS-3] [Zoznam trestnych cinov, ktore budu pouzite ako queries na google news](#) Created: 01/Oct/21 Resolved: 04/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Müller
Time Spent:	Not Specified
Original estimate:	Not Specified
Attachments:	 list_of_crimes.txt

Description

Zoznam zlocinov v anglictine, ktore budu pouzite ako query v Google News.
Schválené vedúcim.

[AMS-5] [Vysoka architektura](#) Created: 01/Oct/21 Resolved: 03/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	David Silady
Time Spent:	Not Specified
Original estimate:	Not Specified
Attachments:	 ams_high_arch.png

Description

Popripade k nodom v grafe aj na com to bezi (databaza, server, technologia, atd.)

Poznámka vedúceho: minimalizovať náklady na vývoj aby sa vyskúšala životnosť stránky; môže to všetko bežať na jednom serveri.

[AMS-6] [Specifikacia požiadaviek](#) Created: 01/Oct/21 Resolved: 03/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Jakub Hlavačka
Time Spent:	Not Specified
Original estimate:	Not Specified

Description

Poznámka vedúceho:

- Nefunkčné požiadavky:
 - o Fungovať by to malo aj bez prihlásenia
 - o Mohlo by to vyhodit' niečo aj bez prihlásenia
 - o Registrácia ľudí odrádza
 - o Pre neprihlásených vyhodit' len prvých 5 článkov
 - o Responzívny design – web aj mobil
 - o História, archív pre používateľov – venovať sa až neskôr
- Funkčné požiadavky:
 - o Pridať aj právnické osoby
 - o Pri scrapovaní musíme rátať aj s tým, že nejaký článok sa môže objaviť aj po našom scrapovaní – čo s vecami, ktoré sa mohli zmeniť

Comments

Comment by [Jakub Hlavačka](#) [03/Oct/21]

https://docs.google.com/document/d/18XLz0RXSFB50VAKdrydNxtY9zznVDNJK0wyjOS_NG0s/edit?usp=sharing

[AMS-7] [Prieskum ziskavania dat, praca s kniznicou](#) Created: 01/Oct/21 Resolved: 05/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Adam Šípka
Time Spent:	Not Specified
Original estimate:	Not Specified


Description

Kniznica Google News.

Poznámka vedúceho: mali by sme tu venovať veľa pozornosti, aby sme to celé nemuseli preprogramovať

[AMS-8] [Pripady pouzitia](#) Created: 01/Oct/21 Resolved: 05/Oct/21

Status:	Done
Project:	Adverse Media Screening

Type:	Story
Assignee:	Dominik Horvath
Time Spent:	Not Specified
Original estimate:	Not Specified
Attachments:	 Snímka obrazovky 2021-10-05 124128.png

Description

1. User opens AMS web
2. User enters search query - a name of a person
3. System returns list of news articles that contain this name
4. user clicks on one of the results
5. system redirects the user to the news source

Poznámka vedúceho: je to ok.