Export úloh 30. 11. 2021

Obsah

IN PROGRESS 5

[AMS-36] User databáza + zbieranie aktivity používateľa Created: 13/Oct/21 Updated: 22/Nov/21 5

[AMS-11] Lofi pouzivatelskeho rozhrania Created: 05/Oct/21 Updated: 22/Nov/21 6

TO DO 7

[AMS-117] Komunikácia medzi aplikačným serverom a API serverom Created: 23/Nov/21 Updated: 23/Nov/21 7

[AMS-116] Komunikácia medzi API serverom a Elastic Search Created: 23/Nov/21 Updated: 23/Nov/21 8

[AMS-115] Zobrazenie výsledkov vyhľadávania Created: 23/Nov/21 Updated: 23/Nov/21. 9

[AMS-114] Komunikácia medzi MongoDB a API serverom Created: 23/Nov/21 Updated: 23/Nov/21 10

[AMS-112] Fuzzy search Created: 23/Nov/21 Updated: 23/Nov/21 11

[AMS-109] Analýza log sash na update Elasticsearch z MongoDB a analýza elastic ylm Created: 22/Nov/21 Updated: 22/Nov/21 12

[AMS-108] Zabezpecit elasticsearch a mongodb kontajner Created: 21/Nov/21 Updated: 21/Nov/21 13

[AMS-86] PDF report z nájdených článkov Created: 02/Nov/21 Updated: 02/Nov/21 14

[AMS-85] Prihlásenie sa na odber nejakého vyhľadaného človeka Created: 02/Nov/21 Updated: 02/Nov/21 15

[AMS-84] Označenie článku ako videného (prečítaného) Created: 02/Nov/21 Updated: 02/Nov/21 16

[AMS-83] Hodnotenie risku pri nájdených článkoch Created: 02/Nov/21 Updated: 02/Nov/21 17

[AMS-82] Zvýraznenie hľadaných slovíčok Created: 02/Nov/21 Updated: 02/Nov/21 18

[AMS-77] Filter na web stránke Created: 02/Nov/21 Updated: 02/Nov/21 19

[AMS-72] Technický návrh developerského prostredia Created: 02/Nov/21 Updated: 23/Nov/21 20

[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 Updated: 02/Nov/21 21

[AMS-67] Pozrieť sa na zahodené články, kt. neobsahovali paragrafy Created: 02/Nov/21 Updated: 02/Nov/21 22

[AMS-42] Porovnanie výsledkov nášho vyhľadávania s výsledkami Google News Created: 13/Oct/21 Updated: 27/Oct/21 23

[AMS-41] Vyriešiť, čo s článkami, ktoré budú v pôvodnom zdroji nekompletné Created: 13/Oct/21 Updated: 27/Oct/21 24

[AMS-37] Zistiť, ako fungujú iné zdroje Google News, prípadne aký mechanizmus používa Google News. Created: 13/Oct/21 Updated: 23/Nov/21 25

[AMS-33] Zistiť, čo spája kriminálne články Created: 13/Oct/21 Updated: 27/Oct/21 26

[AMS-19] Analyza - multilingvistika Created: 05/Oct/21 Updated: 27/Oct/21 27

[AMS-12] Nasu API dat na portal verejnych API v ramci propagacie Created: 05/Oct/21 Updated: 27/Oct/21 28

DONE 29

[AMS-118] Rozbehnúť PostgreSQL Created: 23/Nov/21 Updated: 28/Nov/21 Resolved: 28/Nov/21 29

[AMS-113] Indexovanie stiahnutých dát Created: 23/Nov/21 Updated: 29/Nov/21 Resolved: 29/Nov/21 30

[AMS-111] Dokumentácia k riadeniu projektu Created: 23/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21 32

[AMS-110] Dokumentácia k inžinierskemu dielu Created: 23/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21 33

[AMS-103] Rozšírenie docker compose o MongoDB a Elasticsearch Created: 15/Nov/21 Updated: 21/Nov/21 Resolved: 21/Nov/21 34

[AMS-102] Deployment scrapera Created: 15/Nov/21 Updated: 29/Nov/21 Resolved: 28/Nov/21 36

[AMS-95] Implementácia dummy Flask API Created: 09/Nov/21 Updated: 15/Nov/21 Resolved: 15/Nov/21 38

[AMS-94] Návrh Flask API Created: 09/Nov/21 Updated: 15/Nov/21 Resolved: 15/Nov/21 39

[AMS-93] Analýza možností implementácie testovania Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21 44

[AMS-92] Zabezpečenie komunikácie medzi klientom a serverom Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21 47

[AMS-91] Úprava dokumentácie zo stretnutí a šprintov tak, aby mohli ísť na stránku tímu Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21 48

[AMS-90] Integrácia MongoDB a Elasticsearch Created: 09/Nov/21 Updated: 16/Nov/21 Resolved: 16/Nov/21 49

[AMS-89] Integrácia scrapera a vylepšeného parsera Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 22/Nov/21 51

[AMS-88] Rozšíriť úložisko pomocou nepriradeného disku Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 14/Nov/21 53

[AMS-87] Zabrániť zacykleným buildom Created: 07/Nov/21 Updated: 22/Nov/21 Resolved: 15/Nov/21 54

[AMS-73] Vyhľadanie RSS dvojčaťa k HTML článku - všeobecný postup Created: 02/Nov/21 Updated: 23/Nov/21 Resolved: 09/Nov/21 55

[AMS-71] Napĺňanie MongoDB priamo pri scrapovaní Created: 02/Nov/21 Updated: 23/Nov/21 Resolved: 14/Nov/21 56

[AMS-70] Rozšírenie google news scrapperu na lokáciu UK Created: 02/Nov/21 Updated: 02/Nov/21 Resolved: 02/Nov/21 57

[AMS-60] Pamäťová zložitosť elastic search na vzorke dát Created: 01/Nov/21 Updated: 12/Nov/21 Resolved: 12/Nov/21 58

[AMS-59] Parsovanie tagov pomocou lxml, readability, trafilatura Created: 30/Oct/21 Updated: 09/Nov/21 Resolved: 09/Nov/21 60

[AMS-57] Zavesenie klienta Created: 27/Oct/21 Updated: 07/Nov/21 Resolved: 07/Nov/21 63

[AMS-53] Rozšíriť docker compose Created: 27/Oct/21 Updated: 31/Oct/21 Resolved: 31/Oct/21 64

[AMS-51] Pridať info o projekte a opísať členov tímu Created: 27/Oct/21 Updated: 02/Nov/21 Resolved: 01/Nov/21 66

[AMS-50] Request URL s odpoveďou inou ako 200 uložiť do súboru Created: 27/Oct/21 Updated: 23/Nov/21 Resolved: 01/Nov/21 67

[AMS-49] Lokálna MongoDB so získanými dátami Created: 27/Oct/21 Updated: 23/Nov/21 Resolved: 01/Nov/21 68

[AMS-48] Analýza bezstratovej kompresie textu Created: 27/Oct/21 Updated: 01/Nov/21 Resolved: 01/Nov/21 71

[AMS-47] Parsovanie tagov pomocou regex Created: 27/Oct/21 Updated: 02/Nov/21 Resolved: 02/Nov/21 73

[AMS-46] Parsovanie tagov pomocou CSS Selector Created: 27/Oct/21 Updated: 08/Nov/21 Resolved: 08/Nov/21 74

[AMS-44] Stránka tímu Created: 18/Oct/21 Updated: 22/Nov/21 Resolved: 01/Nov/21 ... 76

[AMS-43] Otvorenie portov na serveri Created: 18/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21 77

[AMS-39] Vyriešiť, aby fiitkar nemohol urobiť "sudo su" na virtuálnom stroji (aby nemal šancu sa zmeniť na roota) Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 19/Oct/21 . 78

[AMS-35] Prieskum databáz Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21 79

[AMS-34] Vytvoriť github projekt Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21 80

[AMS-32] Prihláška na TP CUP Created: 13/Oct/21 Updated: 02/Nov/21 Resolved: 02/Nov/21 81

[AMS-31] Set up elastic search v docker container Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 16/Oct/21 82

[AMS-29] Získanie dát pre prototyp Created: 13/Oct/21 Updated: 23/Nov/21 Resolved: 26/Oct/21 84

[AMS-21] Vytvorit Slack server Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 11/Oct/21 86

[AMS-20] Analyza parametrov Google News Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 13/Oct/21 87

[AMS-15] Realny scrapping - vytvorenie vzorky Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 10/Oct/21 88

[AMS-13] Zistit, ktore miestnosti su na karticky Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21 89

[AMS-10] Revizia poziadaviek Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 13/Oct/21 90

[AMS-9] Spísať denník z prvého stretnutia s vedúcim Created: 03/Oct/21 Updated: 22/Nov/21 Resolved: 05/Oct/21 91

[AMS-8] Pripady pouzitia Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 05/Oct/21 92

[AMS-7] Prieskum ziskavania dat, praca s kniznicou Created: 01/Oct/21 Updated: 23/Nov/21 Resolved: 05/Oct/21 93

[AMS-6] Specifikacia poziadaviek Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 03/Oct/21 94

[AMS-5] Vysoka architektura Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 03/Oct/21 95

[AMS-4] Zaobstarat stroj v škole Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 13/Oct/21 96

[AMS-3] Zoznam trestnych cinov, ktore budu pouzite ako queries na google news Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 04/Oct/21 97

IN PROGRESS

[AMS-36] User databáza + zbieranie aktivity používateľa Created: 13/Oct/21 Updated: 22/Nov/21	
Status:	IN PROGRESS
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	Conceptual Model.bmp Tables.bmp
Sprint:	
Rank:	0|i0001y:zr

Comments

Comment by Táňa Poláková [17/Oct/21]

Model, ktory velmi na hrubo zobrazuje aka by mohla byt pointa user databazy.

Comment by Táňa Poláková [17/Oct/21]

Podrobne zobrazenie tabuliek a ich atributov. Je potrebne doriesit, ako by sme prilinkovali clanky.

[AMS-11] Lofi pouzivatelskeho rozhrania Created: 05/Oct/21 Updated: 22/Nov/21	
Status:	IN PROGRESS
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	David Silady
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	Screenshot_93.png
Sprint:	
Rank:	0|i0001y:zi

Description

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

TO DO

[AMS-117] Komunikácia medzi aplikačným serverom a API serverom Created: 23/Nov/21 Updated: 23/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	David Silady
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 4
Story point estimate:	5
Rank:	0|i000nb:

[AMS-116] Komunikácia medzi API serverom a Elastic Search Created: 23/Nov/21 Updated: 23/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Müller
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 4
Story point estimate:	13
Rank:	0|i000n3:

[AMS-115] Zobrazenie výsledkov vyhľadávania Created: 23/Nov/21 Updated: 23/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 4
Story point estimate:	13
Rank:	0|i000mv:

[AMS-114] Komunikácia medzi MongoDB a API serverom Created: 23/Nov/21 Updated: 23/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Adam Šípka
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 4
Story point estimate:	13
Rank:	0|i000mn:

Description

+ stránkovanie

[AMS-112] Fuzzy search Created: 23/Nov/21 Updated: 23/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000m7:

[AMS-109] Analýza log sash na update Elasticsearch z MongoDB a analýza elastic ylm

Created: 22/Nov/21 Updated: 22/Nov/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000lz:

[AMS-108] Zabezpecit elasticsearch a mongodb kontajner Created: 21/Nov/21 Updated: 21/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000lr:

[AMS-86] PDF report z nájdených článkov Created: 02/Nov/21 Updated: 02/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000hr:

[AMS-85] Prihlásenie sa na odber nejakého vyhľadaného človeka

Created: 02/Nov/21 Updated: 02/Nov/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000hj:

Description

vždy keď sa pridá nový článok o hľadanom človeku

[AMS-84] Označenie článku ako videného (prečítaného) Created: 02/Nov/21 Updated: 02/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000hb:

[AMS-83] Hodnotenie risku pri nájdených článkoch Created: 02/Nov/21 Updated: 02/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000h3:

Description

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

[AMS-82] Zvýraznenie hľadaných slovíčok Created: 02/Nov/21 Updated: 02/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000gv:

[AMS-77] Filter na web stránke Created: 02/Nov/21 Updated: 02/Nov/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
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-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:	
Rank:	0|i000fr:

[AMS-72] Technický návrh developerského prostredia

Created: 02/Nov/21 Updated: 23/Nov/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	David Silady
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 4
Story point estimate:	8
Rank:	0|i000nr:

[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 Updated: 02/Nov/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000dr:

[AMS-67] Pozrieť sa na zahodené články, kt. neobsahovali paragrafy

Created: 02/Nov/21 Updated: 02/Nov/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i000dj:

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

Created: 13/Oct/21 Updated: 27/Oct/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0008n:

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

Created: 13/Oct/21 Updated: 27/Oct/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0008f:

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 Updated: 23/Nov/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Adam Šípka
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0001y:zx

Comments

Comment by Adam Šípka [23/Nov/21]

Väčšina iných zdrojov je buď obmedzená na určitý počet requestov za deň, alebo je spoplatnená. Prípadne sa jedná o príliš špecifické zdroje, ako napríklad New York Times, čiže by sme mali jednostranne zamerané dáta.

[AMS-33] Zistiť, čo spája kriminálne články Created: 13/Oct/21 Updated: 27/Oct/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0006f:

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 Updated: 27/Oct/21	
Status:	To Do
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
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 funuguje 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	

Sprint:	
Rank:	0|i0001y:zz

Description

API na Google translate, preklad klucovych slov

[AMS-12] Nasu API dat na portal verejnych API v ramci propagacie

Created: 05/Oct/21 Updated: 27/Oct/21

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

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Unassigned
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0001z:

DONE

[AMS-118] Rozbehnúť PostgreSQL Created: 23/Nov/21 Updated: 28/Nov/21 Resolved: 28/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	2 hours		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 4
Story point estimate:	3
Rank:	0|i000ml:

Comments

Comment by Jakub Hlavačka [28/Nov/21]
POSTGRES_DB=users, prihlasovacie udaje su v /home/fiitkar/docker-file/postgres_db.env

[AMS-113] Indexovanie stiahnutých dát Created: 23/Nov/21 Updated: 29/Nov/21 Resolved: 29/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	7 hours, 30 minutes		
Original estimate:	Not Specified		

Attachments:	image-20211129-155001.png
Sprint:	AMS Sprint 4
Story point estimate:	8
Rank:	0|i000me:

Description

Rozbehnutie docker kontajneru pre elastic search.

Zaindexovanie UK clankov za posledne dva tyzdne.

Comments

Comment by Jakub Hlavačka [29/Nov/21]

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

Github actions nepodporuje deploy ale iba build. Pretoze by bolo zbytocne zakazdym indexovat.

Zakomentoval som tento kontajner v produkcnej verzii, pretoze by sa pri kazdom docker-compose up indexoval

Comment by Jakub Hlavačka [29/Nov/21]

Aktuálne je zaindexovaných 51774 dokumentov.

[AMS-111] Dokumentácia k riadeniu projektu Created: 23/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	6 hours		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 3
Rank:	0|i0008v:2

[AMS-110] Dokumentácia k inžinierskemu dielu Created: 23/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	4 hours		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 3
Rank:	0|i0008v:9

[AMS-103] Rozšírenie docker compose o MongoDB a Elasticsearch

Created: 15/Nov/21 Updated: 21/Nov/21 Resolved: 21/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Σ Remaining Estimate:	3 hours	**Remaining Estimate:**	3 hours
Σ Time Spent:	2 hours	**Time Spent:**	2 hours
Σ Original Estimate:	5 hours	**Original estimate:**	5 hours

Sub-tasks:

Key	Summary	Type	Status	Assignee
AMS-105	Analyza River plugin	Subtask	Done	Jakub Hlavačka
AMS-106	Analyza shard, nodes	Subtask	Done	Jakub Hlavačka
AMS-107	Vytvorenie articles db na virtualnej ...	Subtask	Done	Jakub Hlavačka

Sprint:	AMS Sprint 3
Story point estimate:	5
Rank:	0|i0008v:i

Comments

Comment by Jakub Hlavačka [21/Nov/21]

Zistit ako funguje logstash.

Comment by Jakub Hlavačka [21/Nov/21]

https://www.elastic.co/elasticon/conf/2016/sf/quantitative-cluster-sizing - Tu pisu, ze je vhodne otestovat na zaindexovanych datach scenare a zistit, ze kedy jednotlive shardy pomahaju a kedy nie… Shary mozu zahltit CPU, pretoze vyhladavanie v kazdom sharde funguje na samotnom CPU vlakne.

https://www.elastic.co/guide/en/elasticsearch/reference/current/size-your-shards.html#shard-size-recommendation - Tu spominaju, ze by sme sa mali snazit drzat velkost jedneho shardu medzi 10GB az 50GB

https://www.elastic.co/guide/en/elasticsearch/reference/current/size-your-shards.html#shard-count-recommendation - Tu sa spomina, ze ak jeden node prekroci 20 shardov pre 1GB heap memory, tak sa ma pridat dalsi node

https://discuss.elastic.co/t/how-many-nodes-should-an-elasticsearch-cluster-have/227674 - Tu opisuju, ze kolko nodov by mal mat cluster a vetu each node is one host (either physical or virtual) som pochopil tak, ze mi budeme mat jeden node.

Vysledok: Cluster ma jeden node, jeden shard a jednu repliku

Navrh na dalsi task https://www.elastic.co/guide/en/elasticsearch/reference/current/size-your-shards.html#use-ds-ilm-for-time-series - Zistit nieco o elastic ILM, pretoze pouzivame time series data a zjavne sa to hodi.

Dalsi task zistit, ci sa nam hodi logstash na updateovanie elasticsearch vzdy, ked pride novy zaznam do mognodb zo scrapera. Ak ano, tak ho pouzit.

[AMS-102] Deployment scrapera Created: 15/Nov/21 Updated: 29/Nov/21 Resolved: 28/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	1 day, 5 hours		
Original estimate:	Not Specified		

Sprint:	AMS Sprint 4
Story point estimate:	13
Rank:	0|i000mh:

Description

+ systémové premenné na ovládanie kontajnera

Spustenie scrapera na posledné dva týždne (zatiaľ) z UK.

Comments

Comment by Dominik Horvath [28/Nov/21]

parser bol upraveny na konfiguraciu pomocou argumentov. Deploy bol vyrieseny pomocou scrapyd modulu. Tento modul vytvori jednoduchy HTTP API server nad nasim crawlerom. Cela aplikacia je dockerizovana a bezi na virtualnom stroji timu.

Ovladanie je teda prostrednictvom HTTP API, momentalne pouzitelne len priamo z virtualneho stroja - NGNINX konfiguracia z nejakeho dovodu zabranuje POST requestom z externych sieti. Toto je praca do dalsej story.

Priklad ovladania:

Po prihlaseni na virtualny stroj je mozne spustit nas crawler nasledovnym sposobom:

```
curl localhost:6800/schedule.json -d spider=news_spider -d project=ams -d
search_from=2021-11-14 -d search_to=2021-11-28 -d locale=en-gb -d
crimes_file=list_of_crimes.txt
```

Toto spusti crawlovanie dat v danom casovom rozpati so specifikovanym suborom zlocinov.

Cely kod je k dispozicii na githube.

Poznamka: v README na githube su argumenty posielane pomocou **-a** prepinaca a pri vzorke kodu v tomto komentari je pouzity **-d** prepinac. Nie je to chyba, **-d** pri curl-e iba posiela POST data a samotny server ich automaticky potom prehodi na **-a**. Cize pri lokalnom developmente, ked scrapy nebezi na serveri, pouzivat **-a** argumenty a sposob specifikovany v README.

[AMS-95] Implementácia dummy Flask API

Created: 09/Nov/21 Updated: 15/Nov/21 Resolved: 15/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	5 hours		
Original estimate:	5 hours		

Issue links:	**Blocks**			
	is blocked by	~~AMS-94~~	Návrh Flask API	Done
Sprint:	AMS Sprint 3			
Story point estimate:	5			
Rank:	0|i000jj:			

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-94] Návrh Flask API Created: 09/Nov/21 Updated: 15/Nov/21 Resolved: 15/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	6 hours		
Original estimate:	5 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 3
Story point estimate:	5
Rank:	0|i0008x:i

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-2f2c9f9fcdbf

Comment by Táňa Poláková [14/Nov/21]

Versioning: robi sa kvoli supportu starsich zakaznikov v pripade, ze sa nejake poziadavky zmenia a my spravim

Existuju 4 strategie versioningu:

1. URI path versioning - /api/v1/Customers/1 … /api/v2/Customers/1

2. URL parameter versioning - /api/Categories?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 p
sme pridali na zaklade zmenenych poziadaviek (viac: https://www.youtube.com/watch?v=M2KCu0Oq3JE).

Comment by Táňa Poláková [14/Nov/21]

Dokumentaciu je vhodne pisat/generovat, az ked su requesty implementovane.

Postman
Vyhody
intuitivne graficke rozhranie dostupne aj cez prehliadac
podpora kolaboracie
testovanie API
Nevyhody
vysledna dokumentacia je dostupna cez postman URL a nie cez nasu vlastnu - to sa da len pri PRO verzii
response json sa neda dokumentovat
Ukazka dokumentacie
Ako funguje
nainstalujeme si Postman (https://www.postman.com/downloads/)
do timoveho workspace sa vieme dostat cez invite link dostupny na nasom drive
vytvorime si collection, kam budu spadat nase implementovane requesty, mozeme vytvorit aj viac kolekcii ak na
nasa Collection je vlastne subor nasich requestov, ktoru vieme ako celok zdokumentovat - vieme jej dat meno, d
nam asi bude stacit len Collection “Articles“
v ramci kolekcie si vytvorime vsetky implementovane requesty - rozdelene aj podla metod
pri kazdom requeste vieme pridat description, query parametre, ulozit, ako vyzera priklad json response
Flask-RESTPlus
Rozsirenie pre Flask, dostupne cez pip: pip install flask-restplus
https://www.imaginarycloud.com/blog/flask-python/
Vyhody

- vacsinu 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 vzhlad dokumentacie sa mi nepaci (😃)

- testujeme iba cez konzolu

Ukazka kodu

Ukazka dokumentacie

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

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

[AMS-93] Analýza možností implementácie testovania Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Adam Šípka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	5 hours		
Original estimate:	5 hours		

Sprint:	AMS Sprint 3
Story point estimate:	5
Rank:	0|i0008v:4

Comments

Comment by Adam Šípka [23/Nov/21]

Typy testovania:

1. Functional Testing
2. Usability testing -zatiaľ netreba
3. Interface testing
4. Compatibility testing - zatiaľ netreba
5. Performance testing - zatiaľ asi netreba
6. Security testing - zatiaľ asi netreba

Functional Testing
Toto testovanie kontroluje UI, API, databázu, bezpečnosť, klient/server komunikáciu a fungovanie aplikácie.

a) Unit testing

Účelom je otestovať každú funkciu poskytnutím vhodného vstupu a overením výstupu vzhľadom na funkčné požiadavky. Testujeme malé časti softvéru, vykonáva sa ako prvé
Frameworky na testovanie Node.js aplikácií:

- Mocha
- Jest
- Chai
- Jasmine
- AVA

Pre Python:

- unittest + Nose2 a Testify (optional)
- PyTest
- DocTest

Ako písať ľahko testovateľný kód:

- deterministické funkcie: výstupná hodnota záleží od vstupnej a nie od skrytých premenných v danej funkcií
- Inversion of Control technika: oddeliť decision making kód (kedy a za akých podmienok sa niečo vykoná) a action kód (čo sa za vykoná)
- Dependency Injection: Objekt (client) prijíma ďalšie objekty (services/dependencies) od ktorých je závislý. Kód kde je service vložený (injector) určí clientovi, akú službu bude používať, namiesto toho, aby si to client určil sám.

```
public class SmartHomeController
{
    private readonly IDateTimeProvider _dateTimeProvider; // Dependency

    public SmartHomeController(IDateTimeProvider dateTimeProvider)
    {
        // Inject required dependency in the constructor.
        _dateTimeProvider = dateTimeProvider;
    }

    public void ActuateLights(bool motionDetected)
    {
        // Delegating the responsibility
        DateTime time = _dateTimeProvider.GetDateTime();

        ...
    }
}
```

b) Integration testing

Jednotlivé moduly sú skombinované a otestované ako celok. Zamerané na presun dát medzi modulmi.
Ak medzi niektorými už existuje prepojenie/funkcionalita a ostatné nie sú hotové, môžeme

existovať existujúce. Cieľom je odhaliť chyby v interakcií medzi modulmi, keďže tie sú zväčša písané rozdielnymi ľuďmi.

[AMS-92] Zabezpečenie komunikácie medzi klientom a serverom

Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	David Silady
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	1 day, 1 hour		
Original estimate:	1 day		

Sprint:	AMS Sprint 3
Story point estimate:	8
Rank:	0|i000iv:

Description

Vytvoriť wrapper pre fetch. Je potrebné, aby fungoval aj na development prostredí (development cross-origin requests).

[AMS-91] Úprava dokumentácie zo stretnutí a šprintov tak, aby mohli ísť na stránku tímu

Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 23/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	1 day		
Original estimate:	1 day		

Sprint:	AMS Sprint 3
Story point estimate:	8
Rank:	0|i000kt:

Description

Denníky zo stretnutí, upratanie šprintov, exporty úloh, šprint review, …

Comments

Comment by Táňa Poláková [23/Nov/21]

Denníky zo stretnutí, šprint reviews aj exporty úloh sú hotové a sú zatiaľ umiestnené na tímovom drive.

[AMS-90] Integrácia MongoDB a Elasticsearch

Created: 09/Nov/21 Updated: 16/Nov/21 Resolved: 16/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Σ Remaining Estimate:	0 minutes	**Remaining Estimate:**	0 minutes
Σ Time Spent:	1 day, 1 hour	**Time Spent:**	1 day, 1 hour
Σ Original Estimate:	1 day	**Original estimate:**	1 day

Sub-tasks:

Key	Summary	Type	Status	Assignee
AMS-97	Vytvorenie docker-compose file s Mong...	Subtask	Done	Jakub Hlavačka
AMS-98	Napisat dokumentaciu do README.md	Subtask	Done	Jakub Hlavačka
AMS-99	Skonzultovat komenty v storke	Subtask	Done	Jakub Hlavačka
AMS-100	Vytvorit config pre articles_index	Subtask	Done	Jakub Hlavačka
AMS-101	vytvorit komprimovanu collection z ma...	Subtask	Done	Jakub Hlavačka

Sprint:	AMS Sprint 3
Story point estimate:	8
Rank:	0|i0008w:

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-89] Integrácia scrapera a vylepšeného parsera

Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 22/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Müller
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	1 day		
Original estimate:	5 hours		

Attachments:	improved_parser_measurments.txt
Sprint:	AMS Sprint 3
Story point estimate:	5
Rank:	0|i000kr:

Description

Je potrebné vylepšiť parser z predošlého šprintu (konkrétne z tasku “Parsovanie tagov pomocou CSS Selector”) a integrovať ho do nášho scrapera.

Comments

Comment by Jakub Müller [23/Nov/21]

Parser bol integrovaný bez nejakých problémov, keďže bol vytvorený pomocou knižnice Scrapy, ktorá je použitá aj pri samotnom scraperi.

Comment by Jakub Müller [23/Nov/21]

Vylepšenie parsera sa týkalo odstránenia nepotrebných dát z tagov paragrafov a headingov, ktoré boli vyselektované pomocou XPATH selektora. Tieto nepotrebné dáta predstavovali atribúty daných tagov (teda napríklad classy, ID-čka a rôzne iné špecifikácie), ktoré nepotrebujeme, pretože nijako neovplyvňujú formu nášho výsledného HTML súboru.

Comment by Jakub Müller [23/Nov/21]

Na odstránenie týchto atribútov bolo vyskúšaných viacero spôsobov. Bolo skúmané, či sa nedajú jednoducho odignorovať už pri samotnom selektovaní tagov pomocou **XPATH** selektora. Toto sa ukázalo ako nemožné. Následne bolo skúšané odstraňovanie atribútov pomocou **regex**-ov a aj pomocou **listových operácií** v Pythone.

Nakoniec sa ukázalo, že riešenie pomocou listových operácií bolo najlepšie. Následne sa tiež kontrolovalo, či daný tag má vôbec nejaký obsah a ak nie, tak sa celý odstránil.

Comment by Jakub Müller [23/Nov/21]

Toto riešenie bolo testované na rovnakom súbore ako pôvodný parser (1_articles.jl, ktorý mal veľkosť približne 3,5GB). Pôvodný parser dokázal zredukovať daný súbor o cca **92%** a vylepšený parser ho dokázal zredukovať o **93%**, teda o približne 1% viac, pričom sa nestratili žiadne potrebné informácie z daných HTML súborov. Tento vylepšený parser taktiež dokázal spracovať tieto články v približne rovnakom čase ako pôvodný a to za cca **3 minúty**.

Detailné výsledky testovania vylepšeného parseru sú uvedené v prílohe **improved_parser_measurments.txt**

[AMS-88] Rozšíriť úložisko pomocou nepriradeného disku

Created: 09/Nov/21 Updated: 23/Nov/21 Resolved: 14/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	2 hours		
Original estimate:	5 hours		

Sprint:	AMS Sprint 3
Story point estimate:	5
Rank:	0|i0008y:

Comments

Comment by Dominik Horvath [14/Nov/21]

Disk is mounted under **/data**

Comment by Dominik Horvath [23/Nov/21]

Disk je mountnuty pod /data na virtualnom stroji timu. Obsahuje 20GB volneho miesta a bol naformatovany na suborovy system ext4.

[AMS-87] Zabrániť zacykleným buildom

Created: 07/Nov/21 Updated: 22/Nov/21 Resolved: 15/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Dominik Horvath	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	3 hours		
Time Spent:	2 hours		
Original estimate:	5 hours		

Sprint:	AMS Sprint 3
Story point estimate:	5
Rank:	0|i000bu:

Description

Github actions timeout setting aby nam nebezali buildy zbytocne dlho

[AMS-73] Vyhľadanie RSS dvojčaťa k HTML článku - všeobecný postup

Created: 02/Nov/21 Updated: 23/Nov/21 Resolved: 09/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	4 hours		
Time Spent:	4 hours		
Original estimate:	1 day		

Sprint:	AMS Sprint 04
Rank:	0|i0008z:

Comments

Comment by Dominik Horvath [23/Nov/21]

Bohuzial, nic v podobnej sfere v dnesnej dobe neexistuje. Existujuce citacky RSS streamov potrebuju priamy odkaz na stream konkretneho webu, nedokazu ziskat stream URL len z “base“ URL domeny. Taktiez kazdy portal obsahuje vlastne struktury pre RSS stream URL, ich tvar nie je standardizovany.

V dnesnej dobe je taktiez narocne najst kompletny RSS stream, cize RSS stream obsahujuci aj telo samotneho clanku, nielen metadata o nom. Vzhladom na to, ze vacsina RSS streamov obsahuje len odkaz na samotny clanok, tento postup by nam neposkytol ziadnu vyhodu oproti terajsiemu pristupu.

[AMS-71] Napĺňanie MongoDB priamo pri scrapovaní

Created: 02/Nov/21 Updated: 23/Nov/21 Resolved: 14/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Adam Šípka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	1 day		
Original estimate:	1 day		

Sprint:	AMS Sprint 04, AMS Sprint 3
Rank:	0|i0008y:i

Description

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

Comments

Comment by Adam Šípka [23/Nov/21]

Podarilo sa mi upraviť scraper tak, aby sa získané dáta hneď ukladali do Monogo databázy. Momentálne sa jedná len o lokálnu databázu umiestnenú na mojom osobnom zariadení. Nebude problém neskôr zmeniť, stačí v súbore settings.py upraviť nasledovné premenné (+ pridať heslo):

```
# db server and port (local for now)
MONGODB_SERVER = "localhost"
MONGODB_PORT = 27017
```

Databáza je tvorená z 3 kolekcií (articles, crimemaps a errorlinks). Okrem automatického indexu ***_id***, ktoré tvorí mongo automaticky, som v každej kolekcií nastavil field ***link*** na unikátny index, aby sa zabránilo duplikátom.

[AMS-70] Rozšírenie google news scrapperu na lokáciu UK

Created: 02/Nov/21 Updated: 02/Nov/21 Resolved: 02/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	1 hour		
Original estimate:	1 hour		

Sprint:	AMS Sprint 04
Rank:	0|i000e7:

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-60] Pamäťová zložitosť elastic search na vzorke dát

Created: 01/Nov/21 Updated: 12/Nov/21 Resolved: 12/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	6 hours, 30 minutes		
Original estimate:	6 hours		

Attachments:	image-20211108-145633.png main.py
Sprint:	AMS Sprint 04, AMS Sprint 3
Rank:	0|i0008x:

Description

Zobrat dump od Adam Šípka a spravit index v elastic search, pricom jednotlive 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://hevodata.com/learn/integrating-elasticsearch-and-mongodb/

Comment by Jakub Hlavačka [08/Nov/21]

Mam zatial vytvoreny docker-compose, kde je mongodb a elasticsearch s ich vlastnymi volumes. Mongo ma col
clanky z articles_1.zip . K tasku prikladam script, pomocou ktoreho som cital z mongodb a snazil sa vytvarat ind

Posting list sa mi zatial nepodarilo naimplementovat, pretoze elasticsearch analyzer ma obmedzenie pre tokenizo

Aby som aspon hrubym odhadom zistil pamatovu zlozitost vytvoreneho indexu v elasticsearch, tak som vkladal
jeho kluc bola id z mongodb… Pricom po nejakom 650 clanku vybehol nasledovny 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-mongod
- 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-59] Parsovanie tagov pomocou lxml, readability, trafilatura

Created: 30/Oct/21 Updated: 09/Nov/21 Resolved: 09/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	0 minutes		
Time Spent:	6 hours		
Original estimate:	3 hours		

Attachments:	2_sk_article.html lxml_measurement - Copy.txt p_lxml_measurement.txt readability_measurement.txt trafilatura_measurement.txt
Sprint:	AMS Sprint 04
Rank:	0|i00093:

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 maze tagy na zaklade regexov:

{{'unlikelyCandidatesRe': re.compile('combx|comment|community|disqus|extra|foot|header|menu|remark|rss|shoutbox|sidebar|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 urcitych 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 najlepsi kandidat na zaklade score a potom sa hladaju susedne elementy, ktore by s najlepsim kandidatom mohli suvisiet.

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. Zaujimavostou bolo, ze v kode pouzili readability a na zaklade dlzky extrahovaneho textu sa rozhodli, ci pouziju svoj alebo readability algoritmus.

Rozdiel je, ze trafilatura 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-dvojnasobna-vrazda-v-lednickych-rovniach-kedy-sa-zacne-proces-s-obzalovanym/)

2_sk_article.html

Trafilatura tiez.

[AMS-57] Zavesenie klienta Created: 27/Oct/21 Updated: 07/Nov/21 Resolved: 07/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	David Silady
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	1 hour, 30 minutes		
Time Spent:	1 day, 3 hours, 30 minutes		
Original estimate:	1 day, 5 hours		

Sprint:	AMS Sprint 04
Story point estimate:	13
Rank:	0|i000bj:

Description

Zavesiť klienta buď cez Express alebo priamo cez NGINX. Testovacie API calls, redux, …

[AMS-53] Rozšíriť docker compose Created: 27/Oct/21 Updated: 31/Oct/21 Resolved: 31/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Σ Remaining Estimate:	6 hours	**Remaining Estimate:**	6 hours
Σ Time Spent:	7 hours	**Time Spent:**	7 hours
Σ Original Estimate:	1 day, 5 hours	**Original estimate:**	1 day, 5 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 04				
Story point estimate:	13				
Rank:	0|i0009n:				

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 kontajnery.

Nove cesty v nginx reverse proxy:

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

[AMS-51] Pridať info o projekte a opísať členov tímu

Created: 27/Oct/21 Updated: 02/Nov/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Müller
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	3 hours		
Time Spent:	1 hour		
Original estimate:	4 hours		

Sprint:	AMS Sprint 04
Story point estimate:	4
Rank:	0|i000a7:

[AMS-50] Request URL s odpoveďou inou ako 200 uložiť do súboru

Created: 27/Oct/21 Updated: 23/Nov/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Adam Šípka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	2 hours		
Time Spent:	1 hour		
Original estimate:	3 hours		

Sprint:	AMS Sprint 04
Story point estimate:	3
Rank:	0|i0009p:

Description

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

Comments

Comment by Adam Šípka [23/Nov/21]

Úspešne sa mi podarilo splniť task. URL aj zločiny sa počas behu programu ukladajú do dictionary. Ten je po ukončení programu uložený v JSON formáte.

[AMS-49] Lokálna MongoDB so získanými dátami

Created: 27/Oct/21 Updated: 23/Nov/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Adam Šípka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	3 hours		
Time Spent:	5 hours		
Original estimate:	1 day		

Attachments:	mongodb zistenia_01.txt mongodb zistenia_02.txt
Sprint:	AMS Sprint 04
Story point estimate:	8
Rank:	0|i0009r:

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.

Comments

Comment by Adam Šípka [23/Nov/21]

Veľkosť pôvodných dát: 28,3 GB (3,53 + 4,75 + 4,94 + 4,56 + 4,66 + 5,87)
Veľkosť databázy (bez extra kompresie): 9,028870144 GB
Veľkosť databázy (dodatočná kompresia): 5,791744000 GB
Počet záznamov: 228,774 (articles_all)

Bez komprimovania (úplne basic databáza):
"db" : "articledata", meno databázy
"collections" : 1, počet kolekcií (niečo ako tabuľka v SQL databázach)
"views" : 0,

"objects" : 228774, počet všetkých dokumentov vo všetkých kolekciách
"avgObjSize" : 125179.16999746475, priemerná veľkosť 1 dokumentu
"dataSize" : 28637739437, veľkosť všetkých dokumentov (neskomprimovaných)
"storageSize" : 9025990656, priestor alokovaný všetkým kolekciám v databáze vrátane voľného miesta (bez indexov)
"freeStorageSize" : 2019328,
"indexes" : 1, počet indexov (1 automaticky vytvorený - niečo ako id, unikátny pre každý dokument)
"indexSize" : 2879488, priestor alokovaný pre indexy
"indexFreeStorageSize" : 163840,
"totalSize" : 9028870144, súčet indexSize a storageSize (veľkosť databázy)
"totalFreeStorageSize" : 2183168,
"scaleFactor" : 1,
"fsUsedSize" : 511983788032, využitá veľkosť disku kde mongo ukladá veci
"fsTotalSize" : 984432504832, celková veľkosť disku kde mongo ukladá veci
"ok" : 1

Databáza kde bola kolekcia vytvorená s dodatočnou kompresiou dát:
db.createCollection('collectionName', {storageEngine: {wiredTiger: {configString: 'block_compressor=zlib'}}})
"db" : "articlescompressed",
"collections" : 1,
"views" : 0,
"objects" : 228774,
"avgObjSize" : 125179.16999746475,
"dataSize" : 28637739437,
"storageSize" : 5788897280,
"freeStorageSize" : 1695744,
"indexes" : 1,
"indexSize" : 2846720,
"indexFreeStorageSize" : 118784,
"totalSize" : 5791744000,
"totalFreeStorageSize" : 1814528,
"scaleFactor" : 1,
"fsUsedSize" : 517487702016,
"fsTotalSize" : 984432504832,
"ok" : 1

vytvorenie databázy a kolekcie cez mongo shell (s extra kompresiou)
use dbName
db.createCollection('collectionName', {storageEngine: {wiredTiger: {configString: 'block_compressor=zlib'}}})

import dát cez command line
mongoimport --db dbName --collection collectionName --file fileName

Mongo vygeneruje unikátne id pre každý dokument sám, ale asi bude lepie použiť url ako unikátne id, aby sme zabránili duplikátom. Alebo sa to vyrieši ešte pri scrapovaní dát, navštívené linky už nebudeme scrapovať.

[AMS-48] Analýza bezstratovej kompresie textu

Created: 27/Oct/21 Updated: 01/Nov/21 Resolved: 01/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Σ Remaining Estimate:	2 hours	**Remaining Estimate:**	2 hours
Σ Time Spent:	6 hours	**Time Spent:**	6 hours
Σ Original Estimate:	1 day	**Original estimate:**	1 day

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 04
Story point estimate:	8
Rank:	0|i0009h:

Description

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

[AMS-47] Parsovanie tagov pomocou regex Created: 27/Oct/21 Updated: 02/Nov/21 Resolved: 02/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	6 hours		
Time Spent:	2 hours		
Original estimate:	1 day		

Sprint:	AMS Sprint 04
Story point estimate:	8
Rank:	0|i0009b:

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-46] Parsovanie tagov pomocou CSS Selector

Created: 27/Oct/21 Updated: 08/Nov/21 Resolved: 08/Nov/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Müller
Resolution:	Done	**Votes:**	0
Labels:	None		
Σ Remaining Estimate:	0 minutes	**Remaining Estimate:**	0 minutes
Σ Time Spent:	1 day, 5 hours	**Time Spent:**	6 hours
Σ Original Estimate:	1 day, 4 hours	**Original estimate:**	5 hours

Attachments:	example.html scrapy_xpath_measurments.txt.txt

Sub-tasks:

Key	Summary	Type	Status	Assignee
AMS-74	Sekvenčné scrapovanie	Subtask	Done	Dominik Horvath

Sprint:	AMS Sprint 04
Story point estimate:	5
Rank:	0|i000bn:

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_measurments.txt** a príklad výsledného HTML súboru, z ktorého bol vyselektovaný iba text je v prílohe **example.html**

[AMS-44] Stránka tímu Created: 18/Oct/21 Updated: 22/Nov/21 Resolved: 01/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Müller
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0001y:zv

[AMS-43] Otvorenie portov na serveri Created: 18/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	David Silady
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0008v:

Description

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

[AMS-39] Vyriešiť, aby fiitkar nemohol urobiť "sudo su" na virtuálnom stroji (aby nemal šancu sa zmeniť na roota)

Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 19/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|hzzzzz:i

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-35] Prieskum databáz Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Müller
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i00012:

Description

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

[AMS-34] Vytvoriť github projekt

Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Müller
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0000l:i

Description

Viac repozitárov pod jedným projektom

[AMS-32] Prihláška na TP CUP Created: 13/Oct/21 Updated: 02/Nov/21 Resolved: 02/Nov/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Unassigned
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i00067:

[AMS-31] Set up elastic search v docker container

Created: 13/Oct/21 Updated: 22/Nov/21 Resolved: 16/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	docker-compose.yml image-20211016-144056.png
Sprint:	
Rank:	0|i00000:i

Comments

Comment by Jakub Hlavačka [16/Oct/21]

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

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-29] Získanie dát pre prototyp Created: 13/Oct/21 Updated: 23/Nov/21 Resolved: 26/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
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 kategorizácia zoznamu zloč... / Subtask / Done / Dominik Horvath
Sprint:	
Rank:	0|hzzzzz:r

Comments

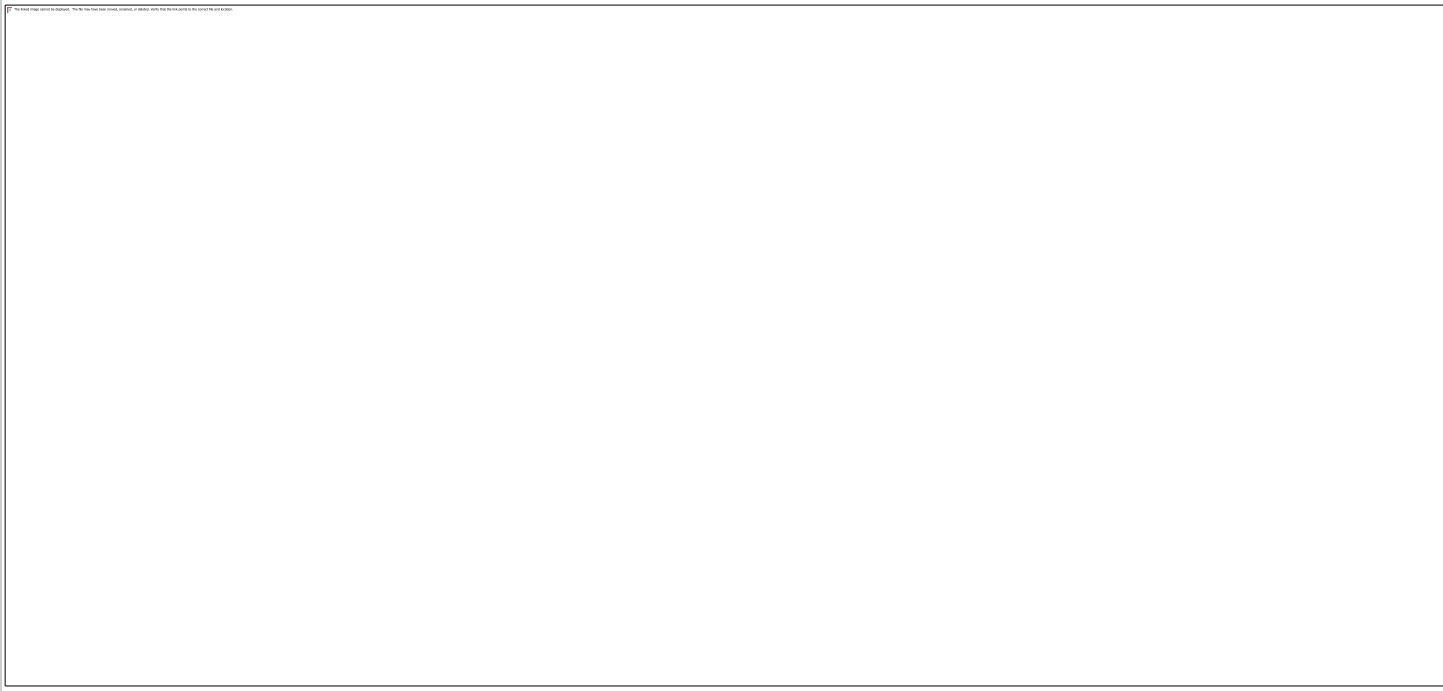
Comment by Jakub Hlavačka [19/Oct/21]

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]



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…

Comment by Dominik Horvath [23/Nov/21]

Scrapper program bol upravený a rozsireny o stahovanie dat pomocou gNewsParseru vytvorenom v https://tim8-

[AMS-21] Vytvorit Slack server Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 11/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i00004:

Comments

Comment by Táňa Poláková [11/Oct/21]
Vytvorila som Team na platforme Microsoft Teams

[AMS-20] Analyza parametrov Google News Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 13/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Adam Šípka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i00002:

Comments

Comment by Adam Šípka [13/Oct/21]

http://books.gigatux.nl/mirror/googlehacks/0596008570/googlehks2-CHP-4-SECT-3.html

[AMS-15] Realny scrapping - vytvorenie vzorky

Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 10/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
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

Sprint:	
Rank:	0|i00009:

Description

Ziskanie vzorky: Jeffrey epstein, poslednych 5 rokov.

[AMS-13] Zistit, ktore miestnosti su na karticky

Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 18/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	David Silady
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0001y:x

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-10] Revizia poziadaviek Created: 05/Oct/21 Updated: 22/Nov/21 Resolved: 13/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i00001:

[AMS-9] Spísať denník z prvého stretnutia s vedúcim

Created: 03/Oct/21 Updated: 22/Nov/21 Resolved: 05/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Táňa Poláková	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0000m:

[AMS-8] Pripady pouzitia Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 05/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Dominik Horvath
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	Snímka obrazovky 2021-10-05 124128.png
Sprint:	
Rank:	0|i0000q:

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

[AMS-7] Prieskum ziskavania dat, praca s kniznicou

Created: 01/Oct/21 Updated: 23/Nov/21 Resolved: 05/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Adam Šípka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0000l:

Comments

Comment by Adam Šípka [23/Nov/21]

Podarilo sa mi vytvoriť prvotný script na zber dát, ktorý pracuje s knižnicou GoogleNews a newspaper3k. Avšak nie vždy sa nám podarilo získať samotný text článku, hlavne ak sa jednalo o iný jazyk ako angličtina. Ďalším problémom bola rýchlosť získavania samotných tiel článkov, ktorá celý proces podstatne spomalila.

[AMS-6] Specifikacia poziadaviek Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 03/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Jakub Hlavačka
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Sprint:	
Rank:	0|i0000j:

Comments

Comment by Jakub Hlavačka [03/Oct/21]

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

[AMS-5] Vysoka architektura Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 03/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	David Silady
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	ams_high_arch.png
Sprint:	
Rank:	0|i0000n:

[AMS-4] Zaobstarat stroj v škole Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 13/Oct/21	
Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Táňa Poláková
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	virtualne-stroje-info2021.txt
Sprint:	
Rank:	0|i00000:

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-3] Zoznam trestnych cinov, ktore budu pouzite ako queries na google news

Created: 01/Oct/21 Updated: 22/Nov/21 Resolved: 04/Oct/21

Status:	Done
Project:	Adverse Media Screening
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Jakub Hlavačka	**Assignee:**	Jakub Müller
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Attachments:	list_of_crimes.txt
Sprint:	
Rank:	0|i0000t:

Generated at Tue Nov 30 12:07:44 UTC 2021 by Táňa Poláková using Jira 1001.0.0-SNAPSHOT#100183-sha1:c958b34e2455eba0e2823012e669c17b0ecee2fc.