Vilniaus universitetas

Ataskaitinė informatikos krypties doktorantų konferencija 2023-09-27

Rolandas Gricius (VU DMSTI doktorantas, Išmaniųjų technologijų tyrimų grupė)

Preliminari darbo tema.

Turinio atpažinimas suskaitmenintuose struktūrizuotuose dokumentuose.

Recognising the contents in digitised structured documents.

Darbo vadovas.

Prof. dr. Igoris Belovas.

Doktorantūros studijų laikotarpis.

2021 m. spalio mėn. 1 d. – 2025 m. rugsėjo mėn. 30 d.

Ataskaitinis laikotarpis.

2023 m. balandžio mėn. 1 d. – 2023 m. rugsėjo mėn. 30 d.

Visų studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2021/2022)	2	3
II (2022/2023)	2	1
III (2023/2024)		
IV (2024/2025)		
Iš viso:	4	4

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citav. rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2021/2022)				1						
II (2022/2023)	1	1				0	Įteikta			
III (2023/2024)					1					
IV (2024/2025)	1				1					
Iš viso:	3	1	1	1	2	0				

Ataskaitinio pusmečio darbo planas ir jo vykdymo suvestinė

Egzaminai 2022/2023 (II pusmetis)		
Planas	Įvykdyta	Būklė
-	-	-

Dalyvavimas konferencijose 2022/2023 (II pusmetis)		
Planas	Įvykdyta	Konferencijos tipas
-	-	-

Publikacijos 2022/2023 (II pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
Journal of King Saud University - Computer and Information Sciences	R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models. Journal of King Saud University - Computer and Information Sciences.	Įteikta: 2023-09-25	Žurnalas turi cituojamumo rodiklį (impact factor) CA WoS duomenų bazėje.

Informacija apie tarptautinius renginius ir publikacijas, kuriose pateikti pagrindiniai disertacijos rezultatai

Dalyvavimas tarptautinėse konferencijose	
	Aprašas
1.	R. Gricius, I. Belovas "Generation of Synthetic Invoices for the Training of Machine Learning Models". International Conference on Pattern Recognition Applications and Methods (ICPRAM) 2023, Lisabona, Portugalija, 2023-02-22 – 24 d.

	Publikacijos (tik su citavimo rodikliu)	
	Bibliografinis aprašas	Būklė
1.	-	-

Tyrimo objektas, tikslas ir uždaviniai

- **Tyrimo objektas – tekstas ir jo išdėstymas sąskaitose-faktūrose (angl. invoices), gautas tiesiogiai arba po OCR procedūros**
- **Tikslas – naudojant natūralios kalbos apdorojimo metodus, atpažinti ir ištraukti tolesniam apdorojimui sąskaitos duomenis, reikšmingus:**
  - **teisėtumui – privalomus pagal teisės aktus duomenis**
  - **apskaitai – data, pirkėjo ir pardavėjo duomenys, sandorio ir mokesčių sumos**
  - **sandorio vykdymui – pristatymo duomenys, apmokėjimo detalės**
- **Uždaviniai – sudaryti duomenų rinkinį tyrimui, atlikti teorinį tyrimą identifikuojant metodus, empirinį tyrimą palyginant jų veikimą ir modifikuoti pritaikant Lietuvos specifikai ir surinktiems duomenims**

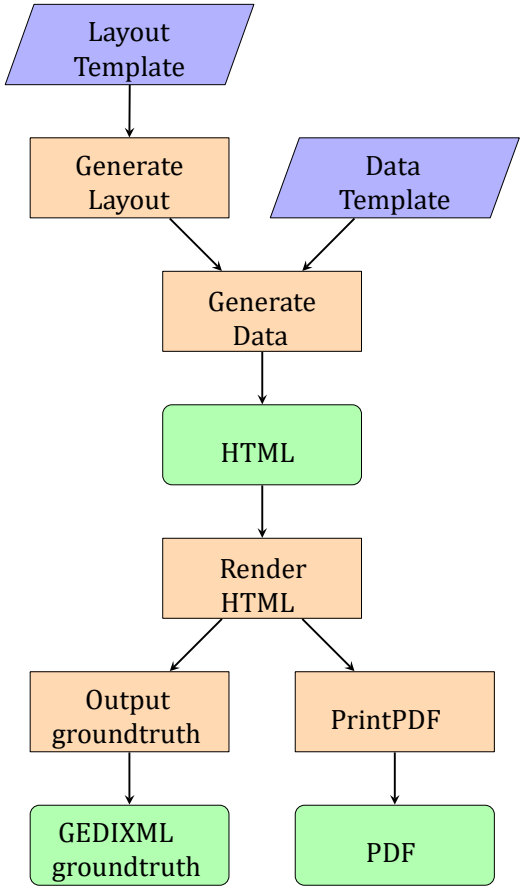
Duomenų rinkinys tyrimui

- Esami sąskaitų rinkiniai nedideli, ne visuomet anotuoti, todėl netinka giliam mokymui
- Viešai prieinamų lietuviškų duomenų rinkinių iš viso nėra
- Dauguma tyrimų naudoja neviešinamus duomenų rinkinius, todėl rezultatus sunku palyginti
- Priimtas sprendimas duomenis tyrimui generuoti

Esami duomenų rinkiniai

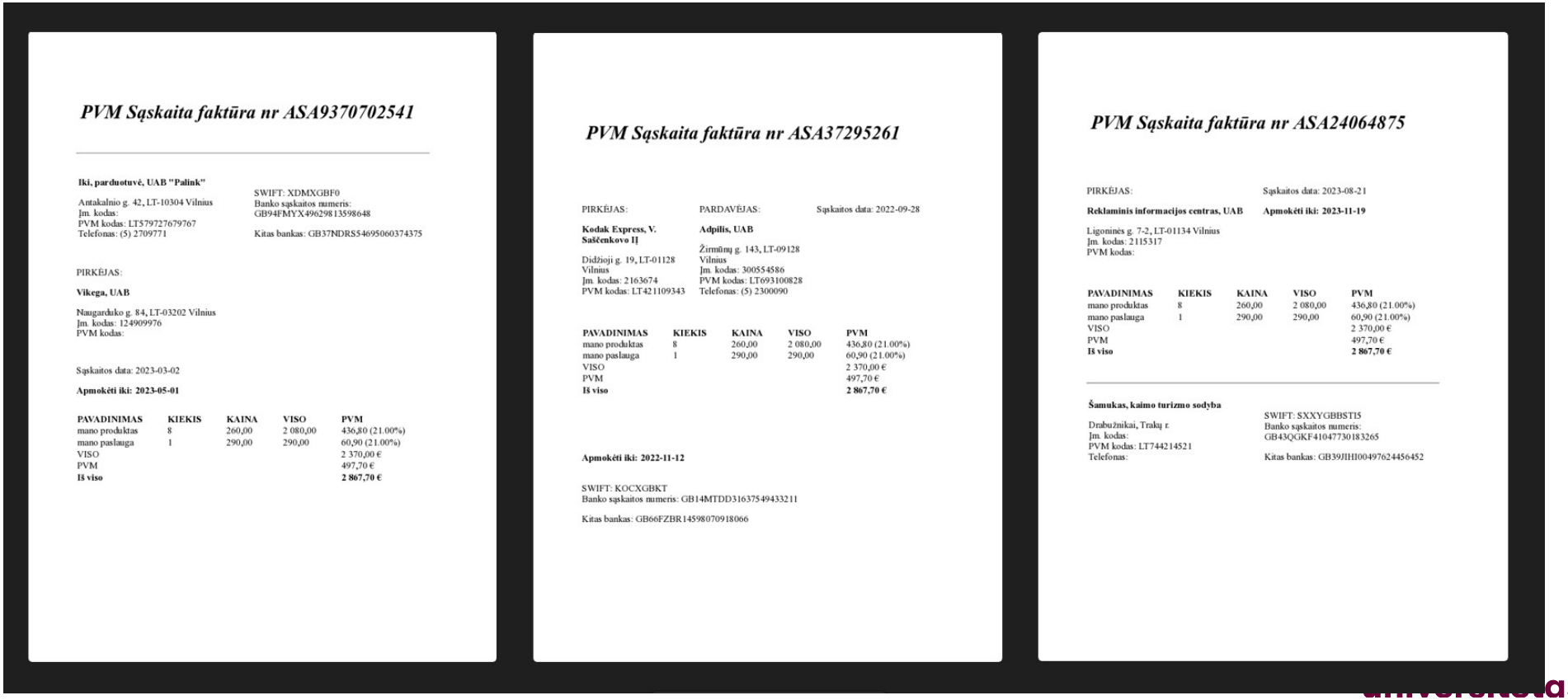
Dataset	SROIE	RVL-CDIP invoices	ZUGFeRD	IDSEM
Size	1000	25000	100	75000
Language	English	English	English, German, French	Spanish
With image	Yes	Yes	Yes	Yes
With text	No	No	Yes	Yes
With entity annotations	Yes	No	Yes	Yes
Drawbacks	No extracted text, moderate size	No extracted text, no annotations, small resolution	Very small dataset	Only 9 invoice templates, electricity invoices

Dokumentų generavimas



- Parengta programinė įranga duomenų rinkiniui generuoti, pritaikyta lietuviškų sąskaitų generavimui
- Atskirtas dokumento išdėstymo ir turinio generavimas
- Sugeneruotas bandomasis 10000 duomenų rinkinys, 1 dok./0,7 sek

Dokumentų generavimas: pavyzdžiai



PVM Sąskaita faktūra nr ASA9370702541

Iki, parduotuvė, UAB "Palink"

Antakalnio g. 42, LT-10304 Vilnius
Jm. kodas:
PVM kodas: LT579727679767
Telefonas: (5) 2709771

SWIFT: XDMXGBF0
Banko sąskaitos numeris:
GB94FMYX49629813598648

Kitas bankas: GB37NDRS54695060374375

PIRKĖJAS:

Vikega, UAB

Naugarduko g. 84, LT-03202 Vilnius
Jm. kodas: 124909976
PVM kodas:

Sąskaitos data: 2023-03-02

Apmokėti iki: 2023-05-01

PAVADINIMAS	KIEKIS	KAINA	VISO	PVM
mano produktas	8	260,00	2 080,00	436,80 (21.00%)
mano paslauga	1	290,00	290,00	60,90 (21.00%)
VISO				2 370,00 €
PVM				497,70 €
Iš viso				**2 867,70 €**

PVM Sąskaita faktūra nr ASA37295261

PIRKĖJAS:

Kodak Express, V. Saščenkovo IĮ

Didžioji g. 19, LT-01128
Vilnius
Jm. kodas: 2163674
PVM kodas: LT421109343

PARDAVĖJAS:

Adpilis, UAB

Žirmūnų g. 143, LT-09128
Vilnius
Jm. kodas: 300554586
PVM kodas: LT693100828
Telefonas: (5) 2300090

Sąskaitos data: 2022-09-28

PAVADINIMAS	KIEKIS	KAINA	VISO	PVM
mano produktas	8	260,00	2 080,00	436,80 (21.00%)
mano paslauga	1	290,00	290,00	60,90 (21.00%)
VISO				2 370,00 €
PVM				497,70 €
Iš viso				**2 867,70 €**

Apmokėti iki: 2022-11-12

SWIFT: KOCXGBKT
Banko sąskaitos numeris: GB14MTDD31637549433211

Kitas bankas: GB66FZBR14598070918066

PVM Sąskaita faktūra nr ASA24064875

PIRKĖJAS:

Reklaminis informacijos centras, UAB

Ligoninės g. 7-2, LT-01134 Vilnius
Jm. kodas: 2115317
PVM kodas:

Sąskaitos data: 2023-08-21

Apmokėti iki: 2023-11-19

PAVADINIMAS	KIEKIS	KAINA	VISO	PVM
mano produktas	8	260,00	2 080,00	436,80 (21.00%)
mano paslauga	1	290,00	290,00	60,90 (21.00%)
VISO				2 370,00 €
PVM				497,70 €
Iš viso				**2 867,70 €**

Šamukas, kaimo turizmo sodyba

Drabužnikai, Trakų r.
Jm. kodas:
PVM kodas: LT744214521
Telefonas:

SWIFT: SXXYGBBSTI5
Banko sąskaitos numeris:
GB43QGKF41047730183265

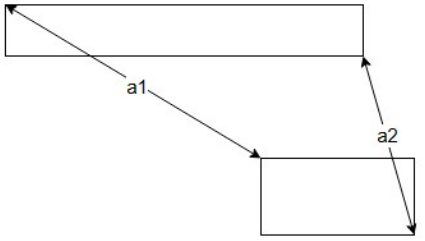
Kitas bankas: GB39JIHI00497624456452

Dokumentų generavimas: metrikos

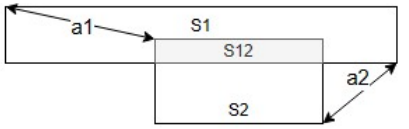
Dataset	SROIE	G Invoices	NewDocGen
Alignment ↑	0.48	0.14	0.18
Overlap ↑	0.998	0.997	0.91
SELF-BLEU ↓	0.44	0.29	0.18

- **Alignment:** $\frac{a1^{*} + a2^{*}}{2(w+h)}$

  * Manhattan distances



- **Overlap:** $1 - \frac{2S12}{S1 + S2}$



- **Self-BLEU:** $\begin{matrix} the\ cat \\ vs \\ cat\ on\ the\ mat; another\ cat;\ dog \end{matrix}$

Dokumentų generavimas

- **Pagal rezultatus parengtas straipsnis Elsevier leidyklos Web of Science reitinguojamame žurnale *Journal of King Saud University - Computer and Information Sciences*. R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models.**

Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

- Parengta programinė įranga duomenų rinkiniui generuoti, pritaikyta lietuviškų sąskaitų generavimui
- Sugeneruotas bandomasis 10000 sąskaitų duomenų rinkinys
- Įvertinta duomenų rinkinio generavimo kokybė, lyginant su kitais generuotais rinkiniais ir realiais dokumentais pagal tris metrikas, vertinančias teksto išdėstymo ir turinio įvairovę.
- Parengtas ir įteiktas straipsnis Elsevier leidyklos Web of Science reitinguojamame žurnale Journal of King Saud University - Computer and Information Sciences. R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models.
- Nuolat pildoma svarbiausių publikacijų preliminaria disertacijos tematika bazė. Straipsniai yra rūšiuojami, atliekama jų analitinė apžvalga

Kito pusmečio darbo planas.

1. Empirinis tyrimas
   - Turinio atpažinimo suskaitmenintuose struktūrizuotuose dokumentuose skirtingų algoritmų palyginimas
   - Įgyvendintų algoritmų modifikavimas, ar naujų algoritmų kūrimas, sprendžiant apibrėžtus uždavinius
2. Mokslinių tyrimų disertacijos tema analitinės apžvalgos pildymas naujai atsirandančiais straipsniais
3. Pakoreguoti įteiktą straipsnį apie sąskaitų generavimo sprendimą pagal recenzentų atsiliepimus

Vilniaus
universitetas

Ačiū už dėmesį

Rolandas Gricius

VU DMSTI doktorantas
rolandas.gricius@mif.stud.vu.lt