Vilniaus universitetas

LIETUVOS TEISMO EKSPERTIZĖS CENTRAS

Ataskaitinė informatikos krypties doktorantų konferencija 2021-03-26

Andrius Chaževskas (VU DMSTI doktorantas, Išmaniųjų technologijų tyrimų grupė)

Vilniaus universitetas

Darbo tema.

Teksto semantinės analizės ir mašininio mokymosi algoritmų taikymo slaptažodžių parinkimui tyrimas.

Application of text semantic analysis and machine learning algorithms for passwords guessing.

Darbo vadovas.

Doc. dr. Igoris Belovas.

Doktorantūros studijų laikotarpis.

2020 m. spalio mėn. 1 d. – 2024 m. rugsėjo mėn. 30 d..

Ataskaitinis laikotarpis.

2020 m. spalio mėn. 1 d. – 2021 m. kovo mėn. 26 d..

Visų studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose		Publikacijos		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė
I (2020/2021) Pirmas pusmetis	1	1		1			
I (2020/2021) Antras pusmetis	1		1 (L)				
II (2021/2022) Pirmas pusmetis	1				1 (KD)		
II (2021/2022) Antras pusmetis	1		1 (L)		1 (KD / R)		
III (2022/2023) Pirmas pusmetis							
III (2022/2023) Antras pusmetis			1 (T)		1 (CA WoS)		
IV (2023/2024) Pirmas pusmetis							
IV (2023/2024) Antras pusmetis			1 (T)		1 (CA WoS)		

Ataskaitinių metų darbo planas ir jo vykdymo suvestinė

Egzaminai		Dalyvavimas konferencijose		Publikacijos	
Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta
Mašininis mokymasis	Išlaikyta: Mašininio mokymasi egzaminas	Dalyvavimas konferencijoje Lietuvoje	Dalyvauta Lietuvos matematikų draugijos LXI konferencijoje, skaitytas pranešimas	-	-

Vilniaus universitetas

Gauti pažymėjimai

ŠIAULIŲ UNIVERSITETAS
www.su.lt

PAŽYMĖJIMAS

Andrius Chaževskas

Vilniaus universitetas, Duomenų mokslo ir skaitmeninių technologijų institutas

2020 m. gruodžio 4 d. dalyvavo
Lietuvos matematikų draugijos LXI konferencijoje ir skaitė pranešimą

„Slaptažodžių parinkimo metodų grindžiamų mašininiu mokymusi apžvalga“

Konferencijos organizacinio komiteto pirmininkas **Prof. dr. Darius Šiaučiūnas**

Šiauliai, 2020 m. gruodžio 4 d.

Registracijos Nr. RP-8 - 251

Stanford | ONLINE

COURSE CERTIFICATE

Mar 13, 2021

Andrius Chaževskas

has successfully completed

Machine Learning

an online non-credit course authorized by Stanford University and offered through Coursera

Associate Professor Andrew Ng
Computer Science Department
Stanford University

SOME ONLINE COURSES MAY DRAW ON MATERIAL FROM COURSES TAUGHT ON-CAMPUS BUT THEY ARE NOT EQUIVALENT TO ON-CAMPUS COURSES. THIS STATEMENT DOES NOT AFFIRM THAT THIS PARTICIPANT WAS ENROLLED AS A STUDENT AT STANFORD UNIVERSITY IN ANY WAY. IT DOES NOT CONFER A STANFORD UNIVERSITY GRADE, COURSE CREDIT OR DEGREE, AND IT DOES NOT VERIFY THE IDENTITY OF THE PARTICIPANT.

Verify at coursera.org/verify/SHM89PYCVN7P
Coursera has confirmed the identity of this individual and their participation in the course.

Vilniaus universitetas

Visų mokslinių tyrimų ir disertacijos rengimo etapai

	Darbo pavadinimas	Atlikimo terminai	Pastabos
1.	Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje): 1.1. Analitinės apžvalgos atlikimas. 1.2. Disertacijos tyrimo objekto detalizavimas. 1.3. Mokslinių problemų susietų su tyrimo objektu identifikavimas ir tyrimo tikslo suformavimas.	2020 m. spalio mėn. – 2021 m. rugsėjo mėn.	Vykdomas, apibendrinti rezultatai mokslinėje ataskaitoje.
	Mokslinio tyrimo vykdymas:		
2.	**2.1.Tyrimo metodikos sudarymas:** 2.1.1. Uždavinių, skirtų tyrimo tikslui pasiekti, suformulavimas. 2.1.2. Tyrimo metodikos išsikeltiems uždaviniams spręsti parinkimas. 2.1.3. Teorinio ir empirinio tyrimų suplanavimas pagal pasirinktą metodiką.	2021 m. spalio mėn. – 2022 m. sausio mėn.	

Visų mokslinių tyrimų ir disertacijos rengimo etapai

	Darbo pavadinimas	Atlikimo terminai	Pastabos
	2.2. Teorinis tyrimas: 2.2.1. Mašininio mokymosi metodų naudojamų automatizuotame slaptažodžių parinkime tyrimas. 2.2.2. Semantinės slaptažodžių analizės ir šablonų parinkimo metodų tyrimas. 2.2.3.Slaptažodžių parinkimo algoritmų taikant semantinę analizę tyrimas.	2022 m. sausio mėn. – 2022 m. rugsėjo mėn.	
	2.3. Empirinis tyrimas: 2.3.1. Skirtingų algoritmų palyginimas. 2.3.2. Įgyvendintų algoritmų modifikacijos, ar naujų algoritmų kūrimas, sprendžiant apibrėžtus uždavinius. 2.3.3. Sukurtų modifikacijų eksperimentinis tyrimas analizuojant jų efektyvumą	2022 m. spalio mėn. – 2023 m. gegužės mėn.	
	2.4. Gautų rezultatų analizė ir apibendrinimas	2023 m. birželio mėn. – 2023 m. rugsėjo mėn.	

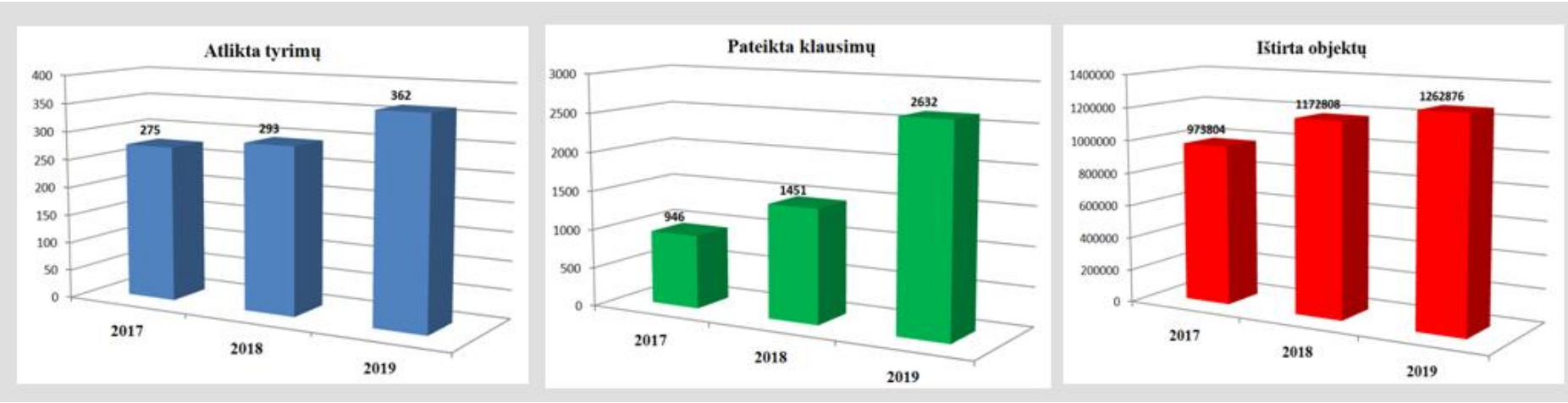
Visų mokslinių tyrimų ir disertacijos rengimo etapai

	Darbo pavadinimas	Atlikimo terminai	Pastabos
3.	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų ir kt.) parengimas: 3.1. Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas. 3.2. Analitinės disertacijos dalies parengimas. 3.3. Teorinės disertacijos dalies parengimas. 3.4. Eksperimentinės disertacijos dalies parengimas. 3.5. Bendrųjų išvadų formulavimas.	2023 m. spalio mėn. – 2024 m. gegužės mėn.	
4.	Daktaro disertacijos parengimas ir svarstymas padalinyje	2024 m. birželio mėn.	
5.	Daktaro disertacijos gynimas	2024 m. rugsėjo mėn.	

Tyrimo objektas, tikslas ir uždaviniai

Ekspertiniai tyrimai:

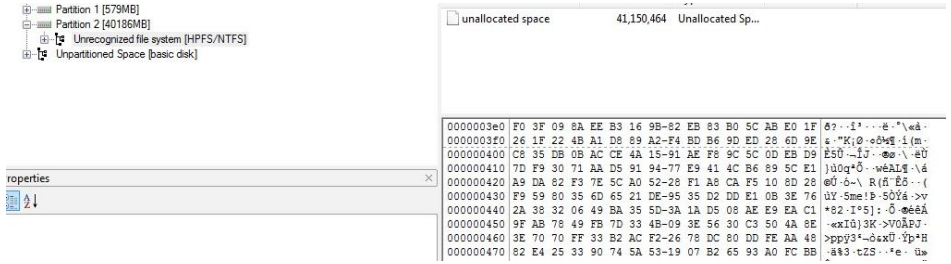
- **Teisminės ekspertizės (susijusios su IT) Lietuvoje.**
- **Pagrindiniai užsakovai.**
- **Tiriamieji objektai.**
- **Tyrimų statistika.**

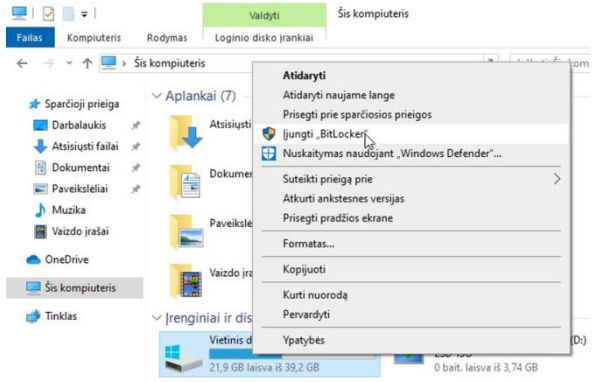


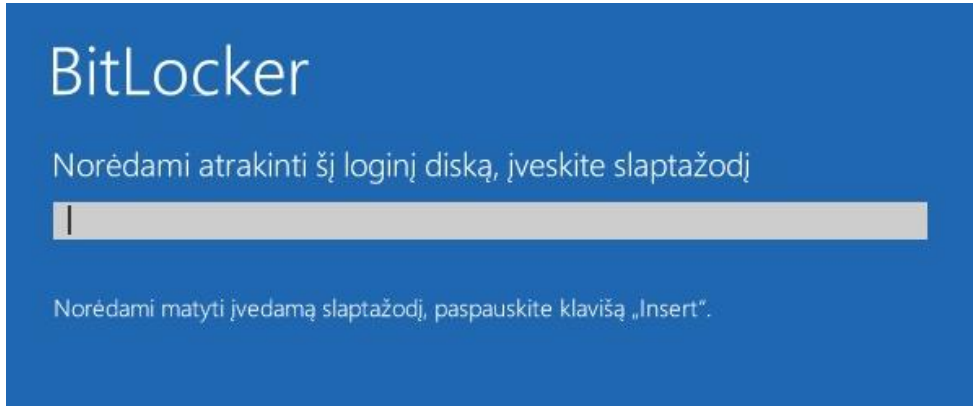
Vilniaus universitetas

Autentifikacija ir duomenų sauga

- **Skaitmeninės informacijos svarba ir saugumas.**
- **Autentifikavimas.**
- **Informacijos šifravimas.**
- **Slaptažodžiai.**
- **Pavyzdžiui** “Bitlocker” apsauga.







Vilniaus universitetas

Problemos

Kaip ištirti šifruotą informaciją?

Slaptažodžių parinkimo metodai:

- Žodynų taikymas;
- Nutekintų slaptažodžių duomenų bazių panaudojimas;
- Pilno perrinkimo atakos („brute-force“);
- Kombinuotos (mišrios) slaptažodžių parinkimo atakos, skirtinguose etapuose naudojant žodynų ir „brute force“ atakas.

Slaptažodžių parinkimo priemonės:

- Laboratorijos aparatūrinė įranga;
- Laboratorijos programinė įranga.

Laikas (kiek galime skirti laiko ir resursų parinkti slaptažodį).

“Brute force” atakos

Slaptažodžio ilgis	Simbolių sekos	Galimų slaptažodžių skaičius
1-8	Skaičiai	111111110
1-8	Skaičiai, mažosios raidės	2.90×10^{12}
1-8	Skaičiai, mažosios ir didžiosios raidės	2.22×10^{14}
1-8	Skaičiai, mažosios ir didžiosios raidės, spec. simboliai	6.16×10^{15}
1-10	Skaičiai	11111111110
1-10	Skaičiai, mažosios raidės	3.76×10^{15}
1-10	Skaičiai, mažosios ir didžiosios raidės	8.53×10^{17}
1-10	Skaičiai, mažosios ir didžiosios raidės, spec. simboliai	1.75×10^{19}

Pilno perrinkimo slaptažodžių parinkimo statistika.

Vartotojų slaptažodžiai

Nutekintų slaptažodžių duomenų bazių panaudojimas:

- https://github.com/lexcor/LT-SecList
- http://downloads.skullsecurity.org
- CityBee vartotojų duombazė

Vartotojų apklausos:

- https://www.bite.lt/apie/ziniasklaidai/slaptazodi-nulauzti-uztruktu-iki-10-sekundziu
- https://www.bite.lt/apie/ziniasklaidai/paskelbti-10-populiariausiu-2019-uju-slaptazodziu

Pirminis darbo tikslas

- **Nutekintų slaptažodžių duomenų bazių analizė rodo, kad žmonės yra linkę naudoti lengvai įsimenamus slaptažodžius, tai reiškia, kad jų pasirinkti slaptažodžiai paprastai turi logišką struktūrą ir nėra atsitiktinių simbolių rinkiniai.**

- **Šiuolaikiniai slaptažodžių parinkimo metodai, remiasi mašininiu mokymusi ir natūralios kalbos apdorojimu, siekiant išnaudoti šią informaciją.**

- **Šio tyrimo pirminis tikslas – naujausių slaptažodžių parinkimų metodų ir juose naudojamų strategijų ištyrimas, praktinis jų pritaikymas ir palyginimas. Gauti rezultatai bus panaudoti siekiant sukurti ir pritaikyti naujus slaptažodžių parinkimo, grįstų mašininio mokymusi, metodus (strategijas), atsižvelgiant į lietuvių kalbos specifiką (vartotojų naudojamų slaptažodžių ypatybes).**

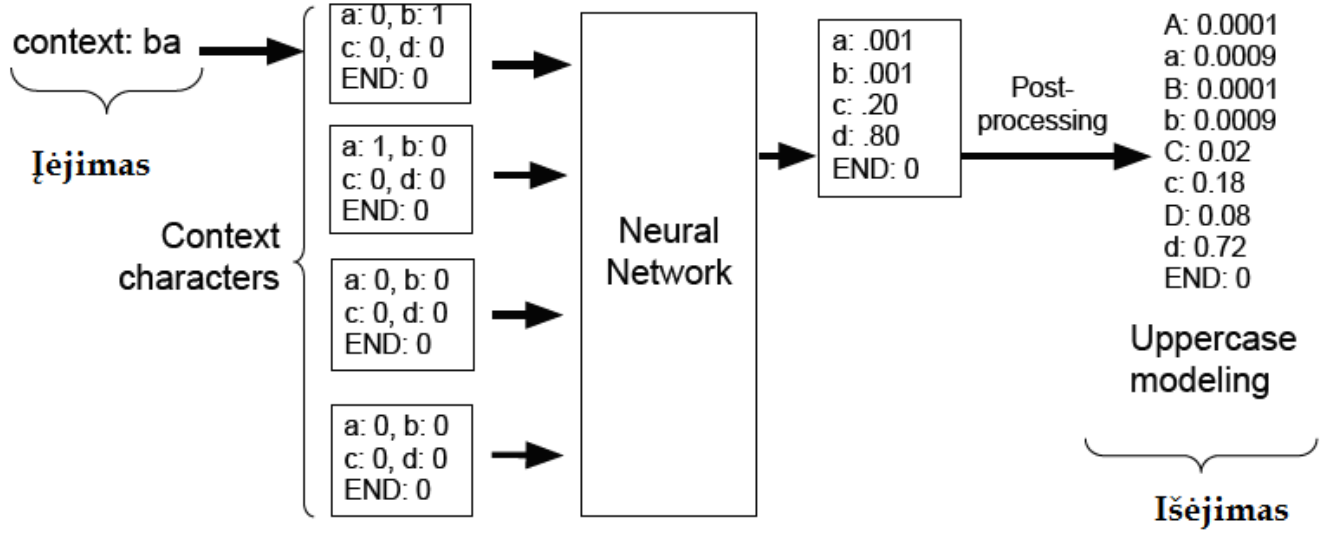
Natūralios kalbos apdorojimas

- Markovo grandinės (filtrai).
- PCFG (angl. k. probabilistic context-free grammars) - tikimybiniai gramatikos taisyklių rinkiniai.
- Semantinė analizė ir klasifikacija.

Neuroniniai tinklai

Pasikartojantys neuroniniai tinklai - Recurrent Neural Networks (RNNs);

Panašiai kaip Markovo modeliuose, LSTM pasikartojantys neuroniniai tinklai yra mokomi generuoti kitą slaptažodžio simbolį, atsižvelgiant į ankstesnius slaptažodžio simbolius.



Šaltinis: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_melicher.pdf, "Fast, Lean, and Accurate: Modeling Password Guessability Using Neural Networks".

Neuroniniai tinklai

Generatyviniai besivaržantys tinklai - Generative adversarial networks (GAN).

PassGAN yra du generatyviniai besivaržantys tinklai – vienas generacinis tinklas ‘mokosi’ ir bando sužinoti statistinę jam pateikiamų duomenų struktūrą, siekiant sudaryti naujus, statistiškai panašius pavyzdžius. Antras tinklas naudojamas testavimui (klaidų aptikimui), kuris bando aptikti, kurie pavyzdžiai yra sugeneruoti generacinio tinklo, o kurie yra originalūs (pirminiai). Mokymas baigiamas kai testavimo tinklas negali atskirti reikšmės kilmės.

Kuo tinklams pateikiamas realių slaptažodžių skaičius didesnis tuo šis metodas yra patikimesnis.

Rezultatų palyginimas

Lentelė Nr. 1: Mokymui ir testavimui naudotas RockYou nutekintų slaptažodžių rinkinys (viso 32,503,388 slapt.), slaptažodžių ilgis iki 10 simbolių (imtinai), testavimui naudotas 1,978,367 – unikalių slaptažodių rinkinys.

Slaptažodžių parinkimo metodas	Unikalių slaptažodžių skaičius	Sutapę slaptažodžiai testavimo rinkinyje	Slaptažodžių skaičius sugeneruotas PassGAN	Sutapę PassGAN slaptažodžiai testavimo rinkinyje
PCFG	10^9	486416 (24.59%)	2.1×10^9	511453 (25.85%)
Markovo 4 eilės	4.9×10^8	532961 (26.93%)	2.47×10^9	532962 (26.93%)
RNN	7.4×10^8	652585 (32.99%)	6×10^9	653978 (33.06%)

Šaltinis: https://arxiv.org/pdf/1709.00440.pdf, "PassGAN: A Deep Learning Approach for Password Guessing".

Vilniaus universitetas

Rezultatų palyginimas

Lentelė Nr. 2: Testavimui naudotas LinkedIn nutekintų slaptažodžių rinkinys (viso 60,065,486 slapt.), slaptažodžių ilgis iki 10 simbolių (imtinai), testavimui naudotas 40,593,536 – unikalių slaptažodių rinkinys.

Slaptažodžių parinkimo metodas	Unikalių slaptažodžių skaičius	Sutapę slaptažodžiai testavimo rinkinyje	Slaptažodžių skaičius sugeneruotas PassGAN	Sutapę PassGAN slaptažodžiai testavimo rinkinyje
PCFG	10^9	7288553 (17.95%)	3.6×10^9	7419248 (18.27%)
Markovo 4 eilės	4.9×10^8	5829786 (14.36%)	1.6×10^9	5829916 (14.36%)
RNN	7.4×10^8	8290173 (20.42%)	6×10^9	8519060 (21.00%)

Šaltinis: https://arxiv.org/pdf/1709.00440.pdf, "PassGAN: A Deep Learning Approach for Password Guessing".

Vilniaus universitetas

Tarpinės išvados

1. Aukščiau pateikti rezultatai rodo realias praktines galimybes naudoti mašininio mokymosi algoritmus slaptažodžių parinkime. Slaptažodžių atspėjimas varijuoja nuo 14 iki 33 %, kas įrodo, kad taikant aukščiau pateiktus metodus galima atspėti slaptažodžius, pasinaudojus skirtingomis vartotojų slaptažodžių sukūrimo taisyklėmis.

2. Neuroninių tinklų panaudojimas lenkia klasikinių algoritmų galimybes.

Per pusmetį gauti moksliniai rezultatai

1. Sudaryta ir nuolat pildoma svarbiausių publikacijų (spausdintų leidiniuose, referuojamuose ir turinčiuose citavimo indeksą *Clarivate Analytics Web of Science* duomenų bazėje) disertacijos tematika bazė.

2. Atliekama naujausių (PassGAN, GENPass, CSNN, TG-SPSR ir kt.) ir klasikinių (žodynai, nutekintų slaptažodžių duombazės) slaptažodžių parinkimo metodų analitinė apžvalga publikacijų leidiniuose, referuojamuose ir turinčiuose citavimo indeksą *Clarivate Analytics Web of Science* duomenų bazėje, pagrindu.

3. Vykdomo tyrimo tarpiniai rezultatai, buvo pristatyti 61-oje Lietuvos matematikų draugijos konferencijoje (Informatikos sekcija). Pranešimo pavadinimas - ***Slaptažodžių parinkimo metodų grindžiamų mašininiu mokymusi apžvalga***.

4. Renkamos empirinių duomenų (iš įvairių paslaugų tiekėjų, jų tarpe ir lietuviškų, nutekintų slaptažodžių) bazės.

Kito pusmečio darbo planas.

1. Disertacijos tyrimo objekto detalizavimas ir mokslinių problemų susietų su tyrimo objektu identifikavimas;

2. Išlaikyti privalomojo dalyko „Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika“ egzaminą;

3. Esant palankiai epidemiologiniai situacijai, pateikti paraišką stažuotei į Bolonijoje esantį didžiausią Italijos skaičiavimo centrą CINECA.

Vilniaus universitetas

Ačiū už dėmesį

Andrius Chaževskas

VU DMSTI doktorantas
Andrius.Chazevskas@mif.stud.vu.lt