

UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ZÍSKAVANIE ŠTRUKTÚROVANÝCH DÁT O  
PACIENTOCH S OCHORENÍM COVID-19 Z  
PREPÚŠŤACÍCH SPRÁV A KRVNÝCH VÝSLEDKOV  
BAKALÁRSKA PRÁCA

2023  
MARIÁN KRAVEC



UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ZÍSKAVANIE ŠTRUKTÚROVANÝCH DÁT O  
PACIENTOCH S OCHORENÍM COVID-19 Z  
PREPÚŠŤACÍCH SPRÁV A KRVNÝCH VÝSLEDKOV  
BAKALÁRSKA PRÁCA

Študijný program:     Dátová veda  
Študijný odbor:        Informatika a Matematika  
Školiace pracovisko:   Katedra aplikovanej informatiky  
Školiteľ:               Mgr. Vladimír Boža, PhD.

Bratislava, 2023  
Marián Kravec





Univerzita Komenského v Bratislave  
Fakulta matematiky, fyziky a informatiky

---

## ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Marián Kravec  
**Študijný program:** dátová veda (Medziodborové štúdium, bakalársky I. st., denná forma)  
**Študijné odbory:** informatika  
matematika  
**Typ záverečnej práce:** bakalárska  
**Jazyk záverečnej práce:** slovenský  
**Sekundárny jazyk:** anglický

**Názov:** Získavanie štruktúrovaných dát o pacientoch s ochorením COVID-19 z prepúšťacích správ a krvných výsledkov  
*Extracting structured data about patients with COVID-19 from discharge reports and blood results*

**Anotácia:** Nemocničný informačný systém v Univerzitnej nemocnici v Bratislava obsahuje iba neštruktúrované textové dáta o pacientoch (či už prepúšťaciu správu alebo výsledky krvných testov).

Takýto formát dát neumožňuje efektívnu analýzu dát a hľadanie faktorov, ktoré ovplyvňujú prognózu liečby Covid-19.

Cieľom práce je vytvoriť softvér, ktorý tieto neštruktúrované dáta premení do tabuľkovej podoby, ktoré sa dajú následne jednoducho analyzovať.

**Vedúci:** Mgr. Vladimír Boža, PhD.  
**Katedra:** FMFI.KAI - Katedra aplikovanej informatiky  
**Vedúci katedry:** prof. Ing. Igor Farkaš, Dr.  
**Dátum zadania:** 20.10.2022

**Dátum schválenia:** 22.10.2022

doc. Mgr. Tomáš Vinař, PhD.  
garant študijného programu

---

študent

---

vedúci práce

**Pod'akovanie:** Touto cestou by som rád poďakoval svojmu školiťovi Mgr. Vladimírovi Božovi, Phd. za ochotu a rady pri písaní práce a kódu. Zároveň by som rád poďakoval pánovi doc. MUDr. Petrovi Sabakovi, PhD. za umožnenie použitia kódu vyvinutého počas projektu na účely tejto práce.

# Abstrakt

Nemocničný informačný systém v Univerzitnej nemocnici v Bratislava obsahuje iba neštruktúrované textové dáta o pacientoch (či už prepúšťaciu správu alebo výsledky krvných testov). Takýto formát dát neumožňuje efektívnu analýzu dát a hľadanie faktorov, ktoré ovplyvňujú prognózu liečby COVID-19. Cieľom práce je vytvoriť softvér, ktorý tieto neštruktúrované dáta premení do tabuľkovej podoby, ktoré sa dajú následne jednoducho analyzovať.

**Kľúčové slová:** prepúšťacia správa, regulárny výraz, tretie

# Abstract

Abstract in the English language (translation of the abstract in the Slovak language).

**Keywords:** dismissal report, regular expresion





# Obsah

<b>Úvod</b>	<b>1</b>
<b>1 Podobné práce</b>	<b>3</b>
<b>2 Štruktúra prepúšťacej správy</b>	<b>5</b>
2.1 Výzor vstupných dát . . . . .	5
2.1.1 Blok A - Osobné údaje a obdobie hospitalizácie . . . . .	5
2.1.2 Blok B - Anamnéza . . . . .	5
2.1.3 Blok C - Vyšetrenia . . . . .	6
2.1.4 Blok D - Terapia . . . . .	6
2.1.5 Blok E - Epikríza . . . . .	6
2.1.6 Blok F - Záver, odporúčania, špecifické nálezy . . . . .	7
2.1.7 Blok G - Krvné výsledky . . . . .	7
2.2 Problémy pri rozdelení dát na bloky . . . . .	7
2.2.1 Problém nenájdenných blokov . . . . .	7
2.2.2 Problém nesprávne umiestnených dát . . . . .	8
2.3 Riešenia problémov s rozdelením dát do blokov . . . . .	8
2.3.1 Homogénny text . . . . .	8
2.3.2 Menej väčších blokov . . . . .	8
<b>3 Získavanie dát</b>	<b>11</b>
<b>Záver</b>	<b>13</b>
<b>Príloha A</b>	<b>17</b>
<b>Príloha B</b>	<b>19</b>



# Zoznam obrázkov

2.1	Rozloženie správy . . . . .	6
2.2	Upravené rozloženie správy . . . . .	9



# Úvod

V nemocničnom informačnom systéme Univerzitnej nemocnice v Bratislave nie je možnosť jednoducho získať tabuľku obsahujúcu dáta jedného pacienta respektíve skupiny pacientov, tieto dát sa získavali ručne buď postupným kopírovaním jednotlivých dát nachádzajúcich sa v rôznych častiach informačného systému alebo ich hľadaním a následným prepisovaním z prepúšťacej správy.

Tento proces bolo potrebné opakovať pre každého jedného pacienta čo bolo časovo náročné a vyžadovalo si nezanedbateľné množstvo ľudskej práce.

Hlavným účelom tejto práce je vytvorenie softvéru ktorý pomôže výrazne zrýchliť a zjednodušiť získavanie týchto dát.

Tento softvér na svojom vstupe dostane prepúšťaciu správu pacienta a jeho krvné výsledky v podobe klasického textu a následne z týchto ne-štrukturalizovaných dát získava jednotlivé požadované dáta pričom v prípade, že nejakú informáciu nenájde alebo zistí, že zistená hodnota nie je v očakávaných limitoch alebo, že sa v správe nachádza viacero hodnôt pre jednu informáciu tak užívateľovi oznámi o akú informáciu ide a aký je s ňou problém.

Práca je rozdelená do štyroch kapitol. Prvá kapitola sa zameriava na štruktúru prepúšťacej správy a dáta ktoré sa z nej snažíme získať.

Druhá kapitola obsahuje problémy ktoré sa objavili pre jednotlivé získavané informácie a aké spôsoby riešenia sme sa rozhodli použiť.

V tretej sú riešené časti kódu ktoré priamo nesúvisia so získavaním dát ako je zápis dát do výslednej tabuľky všetkých pacientov, zapisovanie problematických dát do logovacieho súboru alebo grafické prostredie pre jednoduchšiu prácu so softvérom.

Posledná kapitola je... nejaká asi napríklad iné možnosti prístupu k problému alebo také niečo... asi nejaké analýzy chybovosti.



# Kapitola 1

## Podobné práce

Aj napriek moderným nemocničným informacným systémom je stále veľké množstvo nemocničných záznamov v podobe čistého alebo čiastočne štrukturalizovaného textu z ktorého je ručné získavanie dát časovo náročné. Preto sa na to využívajú automatizované systémy ktoré zväčša fungujú na jednom z dvoch princípov respektíve na kombinácii oboch. Tými prístupmi sú regulárne výrazy a metódy strojového učenia určené na spracovanie prirodzeného jazyka.

Obe tieto prístupy majú svoje výhody a nevýhody [0]. Výhodou regulárnych výrazov je ich presnosť a transparentnosť čiže možnosť vidieť a upravovať vnútorné fungovanie programu ale ich nevýhodou je, že hľadané výrazy treba ručne vytvárať a vylepšovať čo je často náročné, špecifické pre určitú oblasť a v prípade komplikovaných výrazov je aj komplikovaná údržba. Na druhej strane

Získavanie dát z textu pomocou regulárnych výrazov sa v medicínskych aplikáciach často využíva keďže, narozdiel on iných spôsobov získavania dát z textu akými sú metódy strojového učenia na spracovanie prirodzeného jazyka sú často dosť presné a zároveň ich fungovanie je transparentné takže v prípade problému so získavanými dátami alebo v prípade potreby pridania alebo odobrania získavanej informácie je úprava jednoduchá a spočíva zväčša iba v modifikácii, pridaní alebo odstránení regulárneho výrazu.

Medzi už existujúce systémy využívajúce regulárne výrazy patrí napríklad systém HEDEA [1] čiže Healthcare Data Extraction and Analysis ktorého autormi sú Anshul Aggarwal, Sunita Garhwal a Ajay Kumar a bol vyvinutý na získavanie Indických medicínskych dát. Podobne ako v našom prípade je hlavnou úlohou tohto softvéru získavať medicínske dáta z čiastočne štrukturalizovaných vstupných dát čiže lekárskeho správ a výsledkom testov. Hlavné rozdiely oproti nášmu systému sú, že systém HEDEA sa snaží získavať základné dáta o pacientovi ako sú osobné údaje, výška, váha, tlak, základné krvné výsledky a tak ďalej zatiaľ čo my získavame okrem týchto dát aj dáta špecifické pre pacientov s ochorením COVID-19 ako napríklad typ oxygenoterapie alebo výsledky



testov na prítolátky proti vírusu SARS-CoV-2, zároveň je tento model určený na spracovávanie Indických dát takže je vytvorený pre dáta písané v oficiálnom jazyku Indie v tomto prípade angličtinu zatiaľ čo náš model je vytvorený pre dáta v slovenčine.

[0] Cascaded Data Mining Methods for Text Understanding, with medical case study

[1] HEDEA: A Python Tool for Extracting and Analysing Semi-structured Information from Medical Records

## Kapitola 2

# Štruktúra prepúšťacej správy

Táto kapitola sa zameriava na pochopenie vstupných dát a približnú lokalizáciu hľadaných informácií v nich.

### 2.1 Výzor vstupných dát

Dáta z ktorých sa snažíme získať informácie o pacientovi softvér dostáva v tvare textu obsahujúceho prepúšťaciu správu a krvné výsledky. Väčšina textu v prepúšťacej správe nie je generovaná automaticky nemocničným informačným systémom ale je písaná lekárom čo spôsobuje, že každá správa je do určitej miery originálna.

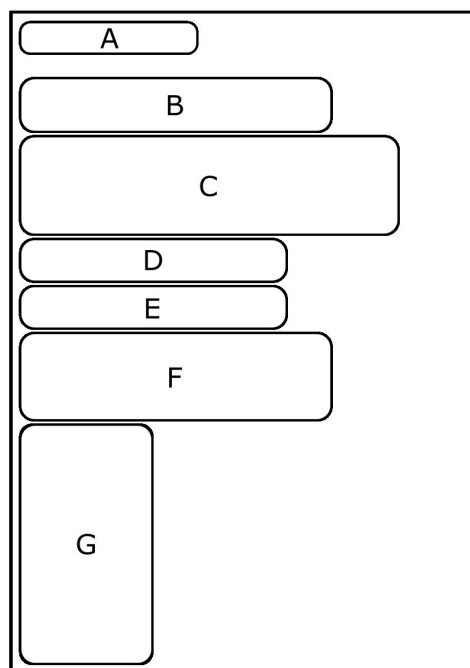
Napriek tomu existuje základná štruktúra ktorú majú spoločnú takmer všetky prepúšťacie správy vďaka ktorej je možné túto správu rozdeliť do blokov (viď obrázok 2.1) ktoré obsahujú určitý informácie. Teraz si prejdeme, čo obsahujú jednotlivé bloky a čo z nich sa mi snažíme získať.

#### 2.1.1 Blok A - Osobné údaje a obdobie hospitalizácie

Bloku A je dvojriadková hlavička v ktorej sa nachádzajú osobné údaje pacienta, čiže jeho celé meno a rodné číslo, a zároveň sa tam nachádza dátum prijatia a dátum prepustenia daného pacienta. Všetky tieto informácie sa snažíme získať.

#### 2.1.2 Blok B - Anamnéza

Tento blok obsahuje informácie o anamnéze a stave pacienta pri prijatí do nemocnice, z tohto bloku sa snažíme získavať informácie ako sú výška, váha a saturácia krvi kyslíkom pri prijatí, a informácia o dlhodobých alebo v minulosti prekonaných chorobách a problémoch ako sú cukrovka, astma, demencia, infarkt myokardu, artériová hypertenzia, fibrilácia predsiení, srdcové zlyhanie a ďalšie.



Obr. 2.1: Rozdelenie správy do špecifických blokov podľa ich obsahu

### 2.1.3 Blok C - Vyšetrenia

Blok C obsahuje informácie o vykonaných vyšetreniach, pre nás sú podstatné výsledky testov na protilátky proti vírusu SARS-CoV-2 typu IgG a IgM pri prijatí, prípadne výsledok testu na ochorenie CDI (Infekcia spôsobená *Clostridium difficile*).

Zároveň sa tu nachádzajú aj výsledky krvných testov avšak tie sa tu nemusia nachádzať úplne. Preto softvér ktorým to spracovávame ich považuje za kontrolné a samotné informácie o krvných výsledkoch sa získavajú z posledného bloku (2.1.7).

### 2.1.4 Blok D - Terapia

V tomto bloku sú informácie o terapii odtiaľto získavame informáciu o liekoch ktoré boli pacientovi podané počas hospitalizácie a o tom či pacient potreboval aj oxygenoterapia a v prípade, že áno aj o aký typ oxygenoterapie išlo.

### 2.1.5 Blok E - Epikríza

Tento blok obsahuje časť správy s názvom epikríza čiže záverečná, súhrnná správa o pacientovi, priebehu jeho choroby a hospitalizácie. Jedinou získavanou informáciou je informácie o smrti pacienta. Táto časť sa však dá zároveň využiť aj na prípadnú kontrolu iných získavaných informácii keďže môže obsahovať informáciu o iných ochoreniach pacienta, jeho liečbe, prípadne o prítomnosti protilátok proti vírusu SARS-CoV-2.

### 2.1.6 Blok F - Záver, odporúčania, špecifické nálezy

Blok F obsahuje zvyšné časti správy ako sú záver, odporúčania a špecifické nálezy. Avšak obsah tohto bloku je výrazne nekonzistentný a jeho jednotlivé časti nemusia byť vôbec prítomné v správe. Našťastie tento blok by nemal neobsahovať žiadne konkrétne získavané informácie.

### 2.1.7 Blok G - Krvné výsledky

Záverečný blok už nie je priamo prepúšťacia správa ale ide o krvné výsledky pacienta ktoré narozdiel od výsledkov v bloku C (2.1.3) sa tu nachádzajú úplné a zároveň vďaka tomu, že nie sú napísané vedľa seba ako v bloku C ale pod sebou (každý výsledok na samostatnom riadku) je práca s nimi výrazne jednoduchšia.

## 2.2 Problémy pri rozdelení dát na bloky

Pri snahe o implementáciu tohto delenia sme zistili, že aj napriek tomu, že sa pomerne veľký počet správ dal do takýchto blokov rozdeliť, tak sa objavilo niekoľko problémov či už pri samotnom delení správy ako aj pri informačnom obsahu jednotlivých častí kvôli ktorým sa ukazuje vhodnejšie takéto riešenie vôbec nepoužiť alebo použiť nejakú robustnejšiu verziu tohto delenia. Niektoré z nájdených problémov si teraz opíšeme.

### 2.2.1 Problém nenájdenných blokov

Pri kontrole správ rozdelených do blokov sme zistili, že sa občas stávalo, že náš softvér nebol schopný nájsť niektorý z blokov, väčšinu z týchto problémov vieme rozdeliť do troch skupín: chýbajúci alebo nesprávne ohraničený blok F, blok E skrytý v bloku F, chýbajúci alebo nesprávne napísaný začiatok bloku.

Prvý problém pre nás nepredstavuje veľký problém keďže v bloku F by sa nemali nenachádzať hľadané informácie.

Druhý problém je o niečo horší keďže blok E je pre nás dôležitý ale tento problém bolo jednoduché opraviť tým, že ak softvér nenájde blok E pri prvotnom delení správy ešte skontroluje či sa v bloku F náhodnou nenachádza.

Tretí problém sa ukazuje ako najproblematickejší sa ukázalo ako pomerne komplikované určiť začiatok a koniec bloku ak softvér nenájde kľúčové slová ktorými sa vo väčšine prípadov jednotlivé bloky začínajú a preto sa stávalo, že softvér niektoré bloky nenašiel a lebo ich pripojil k iným blokom. Tento problém sa stal jedným z hlavných zmeny prístupu k dátam.

### 2.2.2 Problém nesprávne umiestnených dát

Ako ďalší veľký problém sa ukazuje to, že niektoré informácie sa nenachádzajú sa očakávanom mieste. Zväčša išlo o informácie o anamnéze pacienta a výsledkoch jeho vyšetrení ktoré sme predpokladali, že nájdeme v blokoch B respektíve C avšak časť týchto informácií sa nachádzala až v bloku F.

Tento problém samotný sa dá riešiť tým, že informácie ktoré hľadáme v blokoch B a C budeme nakoniec hľadať avšak tento problém spôsobuje, že problém ktorý máme s nájdením a správnym ohraničením bloku F sa stáva relevantným a treba ho riešiť, nanešťastie tento problém je podobný problému ohraničeniu ostatných blokov a preto je jeho riešenie pomerne komplikované.

## 2.3 Riešenia problémov s rozdelením dát do blokov

Skúšaním rôznych spôsobov riešenia sa ukázalo, že najvhodnejšie je neriešiť okrajové prípady súčasného rozdelenia správy ale prerobiť samotné rozdeľovanie. Našli sme najlepšie spôsoby...

### 2.3.1 Homogénny text

Jednou možnosťou je považovať celý text ako jeden homogénny text v ktorom hľadáme všetky informácie.

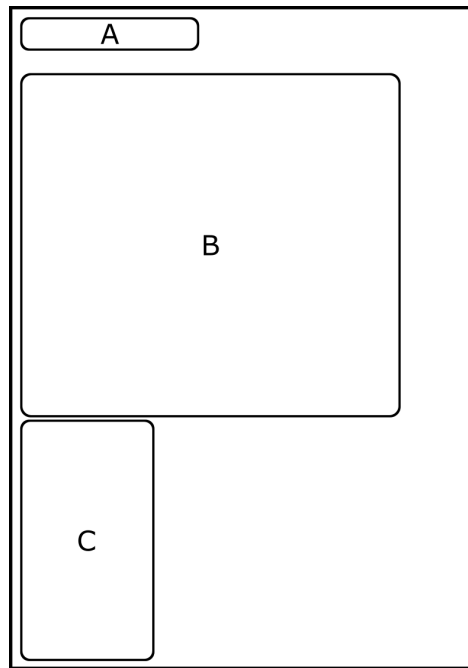
Tento prístup rieši všetky naše problémy avšak zbytočne hľadá niektoré informácie aj na miestach kde vieme, že sa nikdy nebudú nachádzať čo ho spomaľuje.

### 2.3.2 Menej väčších blokov

Tento prístup využíva rozdelenie do blokov ale namiesto pôvodných 7 blokov toto nové rozdelenie má iba 3 bloky, vďaka čomu nemusí hľadať všetky informácie v celom texte a zároveň pri správnom rozdelení vyriešime problémy ktoré sa v pôvodnom rozdelení objavili.

Výzor nového rozdelenia môžeme vidieť na obrázku 2.2. Vidíme, že jediná zmena ktorá nastala je taká, že sme bloky B až F spojili do jedného bloku s názvom B a pôvodný blok G sme premenovali na blok C.

Toto rozdelenie rieši všetky vyššie spomenuté problémy vďaka tomu, že o bloku A vieme, že vždy bude mať tvar dvojriadkovej hlavičky s pri skúšaní sme nenašli žiadnu výnimku, podobne pôvodný blok G čiže nový blok C nie je súčasťou samotnej prepúšťacej správy ale sú to krvné výsledky pacienta ktoré sú v našom vstupe vždy až za správou a majú tvar dvojíc testovaná veličina a výsledok testu (pozitivita alebo hod-



Obr. 2.2: Upravené rozdelenie správy do menšieho počtu väčších blokov

nota) vďaka čomu je jednoduché ich oddeliť od textu správy. Samotný text správy sa problematické pre ďalšie delenie preto ho nechávame pokope v bloku B.



# Kapitola 3

## Získavanie dát

V tejto kapitole si povieme niečo o jednotlivých získavaných dátach o problémoch ktorý sa pri ich získavaní vyskytli a ako sme tieto problémy riešili.





# Záver

Toto má ešte čas.



# Literatúra

- [1] X. Autor1 and Y. Autor2. *Názov knihy*. Vydavateľstvo, 1900.
- [2] X. Autor1 and Y. Autor2. Názov článku (väčšinou z konferencie). In *Názov zborníka (väčšinou názov konferencie spolu s ročníkom)*, pages 1–100. Vydavateľstvo, 1900.
- [3] X. Autor1 and Y. Autor2. Názov článku z časopisu. *Názov časopisu, ktorý článok uverejnil*, 4(3):1–100, 1900.
- [4] X. Autor1 and Y. Autor2. Názov technickej správy. Technical Report TR123/1999, Inštitút vydávajúci správu, June 1999.
- [5] Tobias Oetiker, Hubert Partl, Irene Hyna, and Elisabeth Schlegl. *Nie príliš stručný úvod do systému LaTeX2ε*. 2002. Preklad Ján Buša ml. a st.
- [6] Univerzita Komenského v Bratislave. Vnútorňý predpis č. 7/2018, Úplné znenie vnútorného predpisu č. 12/2013 Smernice rektora Univerzity Komenského v Bratislave o základných náležitostiach záverečných prác, rigorózných prác a habilitačných prác, kontrole ich originality, uchovávaní a sprístupňovaní na Univerzite Komenského v Bratislave v znení dodatku č. 1 a dodatku č. 2 smernica rektora Univerzity Komenského v Bratislave o základných náležitostiach záverečných prác, rigorózných prác a habilitačných prác, kontrole ich originality, uchovávaní a sprístupňovaní na Univerzite Komenského v Bratislave, 2013. [Citované 2020-10-19] Dostupné z [https://uniba.sk/fileadmin/ruk/legislativa/2018/Vp\\_2018\\_07.pdf](https://uniba.sk/fileadmin/ruk/legislativa/2018/Vp_2018_07.pdf).



# Príloha A: obsah elektronickej prílohy

V elektronickej prílohe priloženej k práci sa nachádza zdrojový kód programu a súbory s výsledkami experimentov. Zdrojový kód je zverejnený aj na stránke <http://mojadresa.com/>.

Ak uznáte za vhodné, môžete tu aj podrobnejšie rozpísať obsah tejto prílohy, prípadne poskytnúť návod na inštaláciu programu. Alternatívou je tieto informácie zahrnúť do samotnej prílohy, alebo ich uviesť na oboch miestach.



## Príloha B: Používateľská príručka

V tejto prílohe uvádzame používateľskú príručku k nášmu softvéru. Tu by ďalej pokračoval text príručky. V práci nie je potrebné uvádzať používateľskú príručku, pokiaľ je používanie softvéru intuitívne alebo ak výsledkom práce nie je ucelený softvér určený pre používateľov.

V prílohách môžete uviesť aj ďalšie materiály, ktoré by mohli pôsobiť rušivo v hlavnom texte, ako napríklad rozsiahle tabuľky a podobne. Materiály, ktoré sú príliš dlhé na ich tlač, odovzdajte len v electronickej prílohe.