Týmový projekt BiMat
2016/2017

MU

IBA

EVROPSKÁ UNIE

esf

MINISTERSTVO ŠKOLSTVÍ MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Týmový projekt

?

Týmový projekt – rozdělení úkolů

Týmový projekt - team-leadership

Týmový projekt – dvě obhajoby

(1) TEORIE:
obhajoba zvolených metod a postupů

(2) PRAXE:
obhajoba celého projektu včetně realizace algoritmu a výsledků

Týmový projekt – hackathon

Týmový projekt - zápočet

Týmový projekt – tři zadání

(i) Segmentace obrazů z optické mikroskopie
(Daniel Schwarz)

(ii) Strojové učení z obrazových dat
(Roman Vyškovský)

(iii) Dolování z textových dat
(Martin Komenda a Matěj Karolyi)

Týmový projekt – tři skupiny

AUTOENCODERY

- Kratochvílová M.
- Bezděková M.
- Zouharová S.

SEGMENTACE

- Bučková B.
- Prelecová V.

KLÍČOVÁ SLOVA

- Ježová K.
- Rakušanová S.

Týmový projekt – termíny

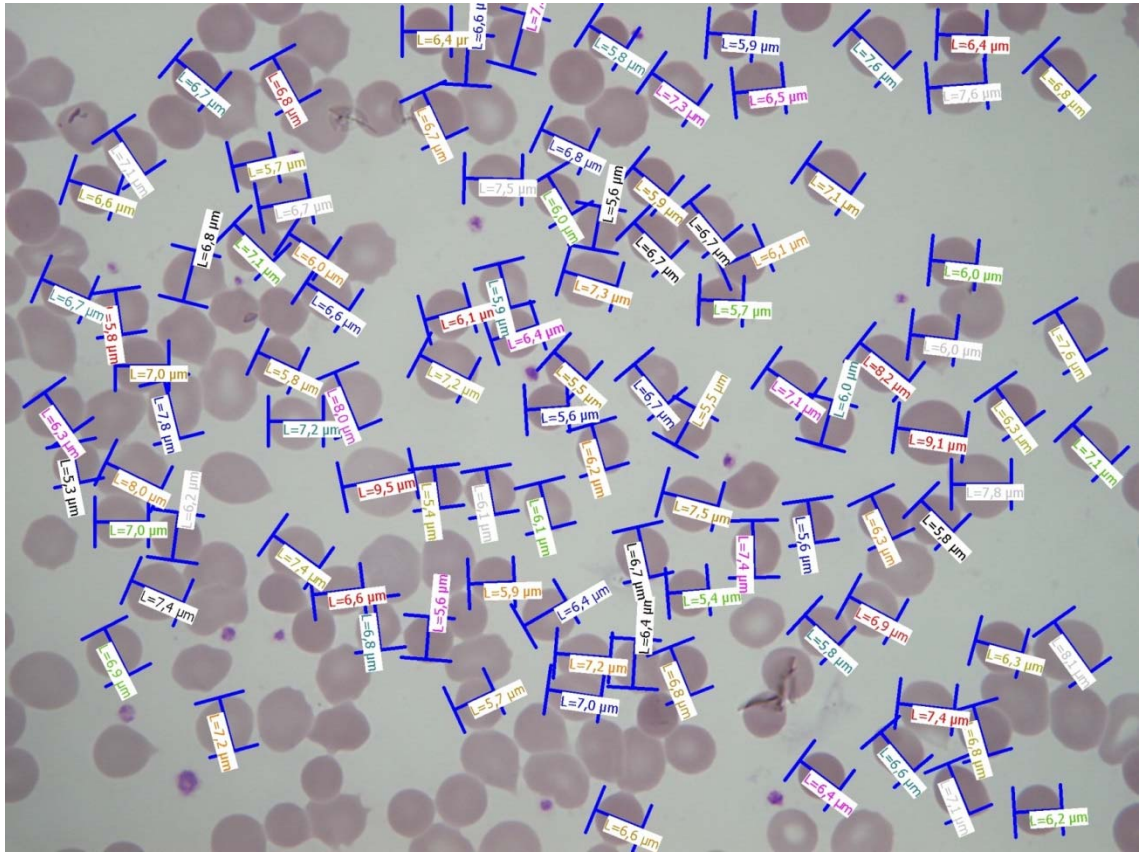
12. října 2016
dle rozvrhu

9. listopadu 2016
dle rozvrhu

14. prosince 2016
dle rozvrhu

SEGMENTACE

Segmentace erytrocytů v digitálních obrazech hematologických nátěrů



Laboratorní diagnostika, morfometrické analýzy

Segmentace erytrocytů v digitálních obrazech hematologických nátěrů

Úloha: Z digitálních obrazů získaných optickou mikroskopií při laboratorní diagnostice chorob krve sestavte XLS sešit se seznamem erytrocytů a přiložte obrázek s legendou

Návrh pracovních balíků:

W1) Charakteristické rysy obrazů z optické mikroskopie, způsob vzniku obrazu, artefakty, zkreslení apod. Laboratorní diagnostika chorob krve a kde jsou možnosti pro automatické zpracování obrazů – morfometrické analýzy apod.

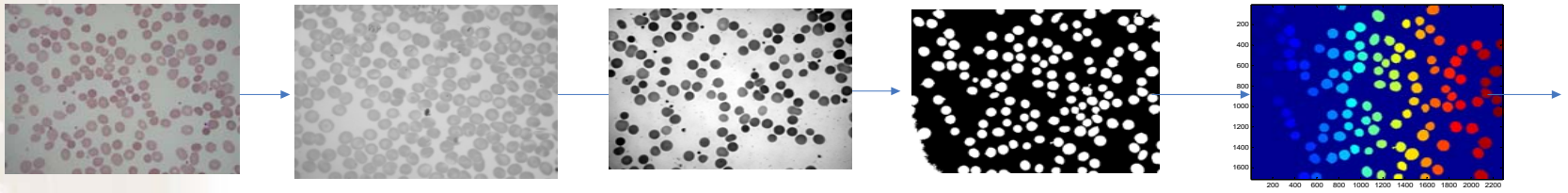
W2) Segmentační metody – základní přístupy a rozdíly mezi nimi. Rozvaha a volba jednoho z mnoha přístupů. Podle zvolené metody volit potom techniky předzpracování…

W3) Samotná práce s obrazovými daty. Předzpracování. Selekce objektů. Práce s RGB obrazy v MATLABu/Rku, vykreslování legendy do obrazu (GUI?). Práce s XLS sešity v Matlabu/Rku. Výpočet vybraného morfometrického parametru (např. průměr, sféricita apod.)

W4) Prezentace výsledků

SEGMENTACE

Segmentace erytrocytů v digitálních obrazech hematologických nátěrů



6.3836
7.0361
7.2651
4.5126
6.8845
6.2507
6.1848
6.9645
5.5038
6.3679
6.4793
5.9986
7.3981
6.8407
5.6103
6.1653
6.2048
5.7739
6.8631
6.0626
5.1926
6.8909
6.8820
6.2709
7.6576
5.9130
5.7892
6.2346
6.0806
5.8936
5.4011
⋮

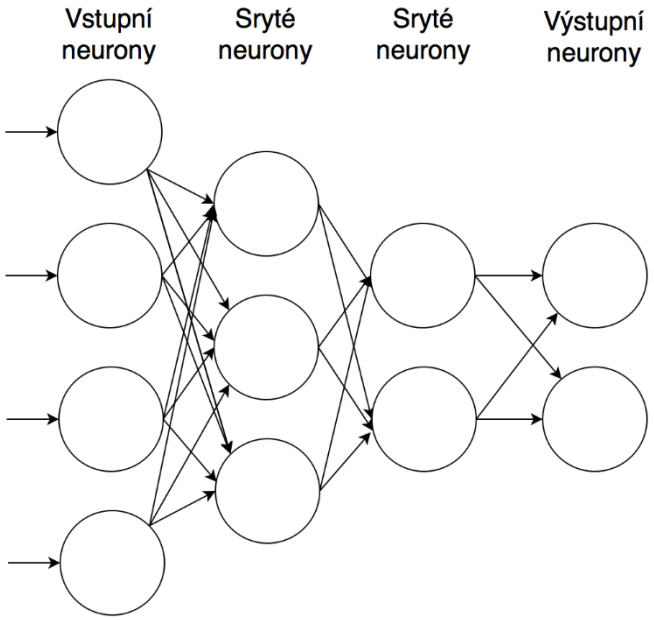
Laboratorní diagnostika, morfometrické analýzy

Autoenkodéry

Vedoucí: Roman Vyškovský

Motivace

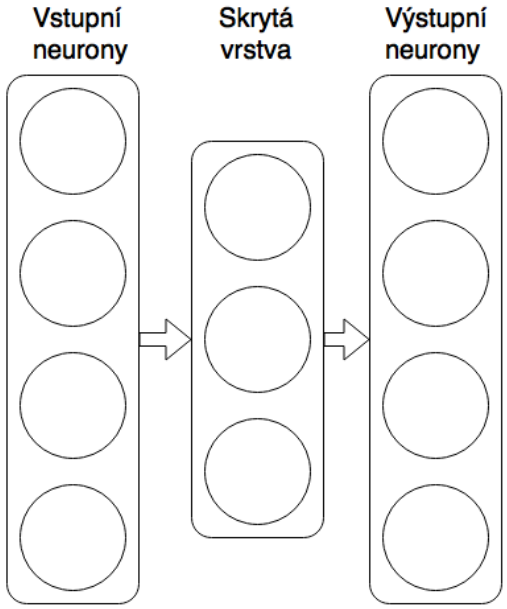
- Hluboká neuronová síť má schopnost zachytit složité závislosti v obrazových datech
- Algoritmus zpětného šíření chyby není pro hluboké sítě efektivní (pomalá optimalizace vah nižších vrstev)
- Síť s mnoha vrstvami a neurony je často přeučená



Neuronová síť

Autoenkodér

- Cíl: Naučit vrstvu neuronů (často s menší dimenzionalitou než vstupní obraz) tak, aby tento vstupní obraz rekonstruovala
- Jde vlastně o extrakci příznaků
- Lze tímto způsobem předučit váhy hluboké neuronové sítě pro klasifikaci



Autoenkodér

Úkoly

- 1. Najít soubor s obrazovými daty a lékařskou/environmentální tématikou vhodný pro klasifikaci
- 2. Naučit hlubokou neuronovou síť s využitím autoenkodérů
- 3. Otestovat na nezávislých datech
- 4. Srovnat výsledek s klasickou neuronovou sítí stejné architektury
- 5. Je tento typ neuronových sítí vhodný pro malé datové soubory?

INSTITUT
BIOSTATISTIKY
A ANALÝZ
Masarykova univerzita

Bi4012 Projekt z Matematické biologie

Zpracování dat v praxi: Redukce klíčových slov

Martin Komenda, Matěj Karolyi

Motivace a použití v praxi

- Klíčové slovo
  - identifikátor při značkování a následnému třídění a vyhledávání obsahu
  - výraz, který se nejčastěji opakuje v textu

- 791 000 000

Motivace a použití v praxi

- Praktické použití napříč různými doménami lidského poznání
  - Marketing, knihovnictví, webdesign, …
  - Zdravotnictví
    - Získání relevantních informací ze záznamů pacienta (volný text)

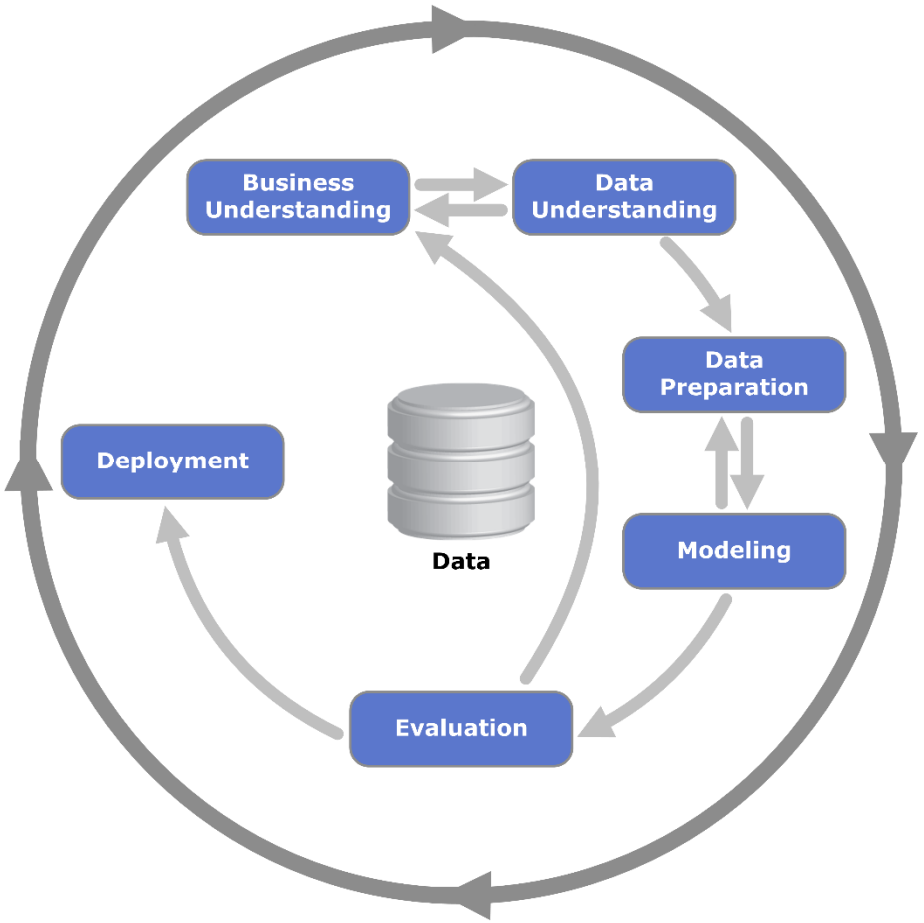

Righteous
Human Rights
Conscience
Gratitude
Forgiveness
Tolerance
Reconciliation
Civil Courage
Sympaty
Responsability
Moral Resistance
Dialogue
Hope
Good
Dignity
Solidarity
Truth
Memory
Freedom
Future

INSTITUT
BIOSTATISTIKY
A ANALÝZ
Masarykova univerzita

Cíle

- Projekt si klade za cíl osvojit si:
  - zpracování velkoobjemových dat
  - znalost vybrané metodiky pro úlohy z oblasti vytěžování dat
  - týmovou spolupráci

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

INSTITUT BIOSTATISTIKY A ANALÝZ
Masarykova univerzita

Porozumění problematice

- Fáze zaměřená na pochopení cílů projektu a požadavků na řešení včetně formulace výzkumných otázek.

- Student
  - Pochopí zadání projektu.
  - Navrhne teoretický postup pro řešení.

INSTITUT BIOSTATISTIKY A ANALÝZ
Masarykova univerzita

Porozumění datům

- Fáze začíná prvotním sběrem dat a následují činnosti, které umožní získat základní představu o datech samotných.

- Student
  - Porozumí vstupní datové sadě (lokální uložení datových souborů a seznámení se s jejich strukturou).

INSTITUT BIOSTATISTIKY A ANALÝZ
Masarykova univerzita

Příprava dat

- Fáze zahrnuje činnosti vedoucí k vytvoření datového souboru, který bude následně dále zpracováván.

- Student
  - Navrhne algoritmus pro zpracování dat včetně eliminace nežádoucích slov (stop-word list).

Modelování

- Fáze zahrnuje algoritmy pro dobývání znalostí.

- Student
  - Aplikuje navržený algoritmus pro vygenerování finálního datového souboru.
  - Vizualizuje výsledky v určené grafové podobě.

INSTITUT BIOSTATISTIKY A ANALÝZ
Masarykova univerzita

Vyhodnocení výsledků

- Ve této fázi se dosažené výsledky vyhodnocují z pohledu splnění cílů formulované na počátku projektu.

- Student
  - Ověří dosažené výsledky s očekávanými výstupy.

Využití výsledků

- Finální fáze projektu, která zahrnuje sepsání závěrečného reportu.

- Student
  - Představí projekt (průběh řešení a výsledky).

Zadání

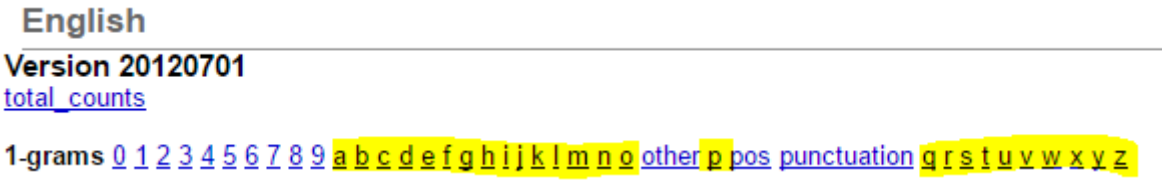
- V souladu s metodikou CRISP-DM najděte a vizualizujte nad vybraným velkoobjemových korpusem dat (Google Books databáze v řádu jednotek GB) skupinu 10 nejčastěji vyskytujících se klíčových slov.

- Výsledná klíčová slova **nesmí obsahovat slova kratší než 4 znaky** a současně nesmí obsahovat slova ze seznamu nežádoucích výrazů (stop-word list – **přiloženo v souboru google-10000-english.txt**).

INSTITUT
BIOSTATISTIKY
A ANALÝZ
Masarykova univerzita

Vstupní data

- Korpus
  - Google Books - Datová sada English Version 20120701
  - 1-gramy (pouze A – Z)

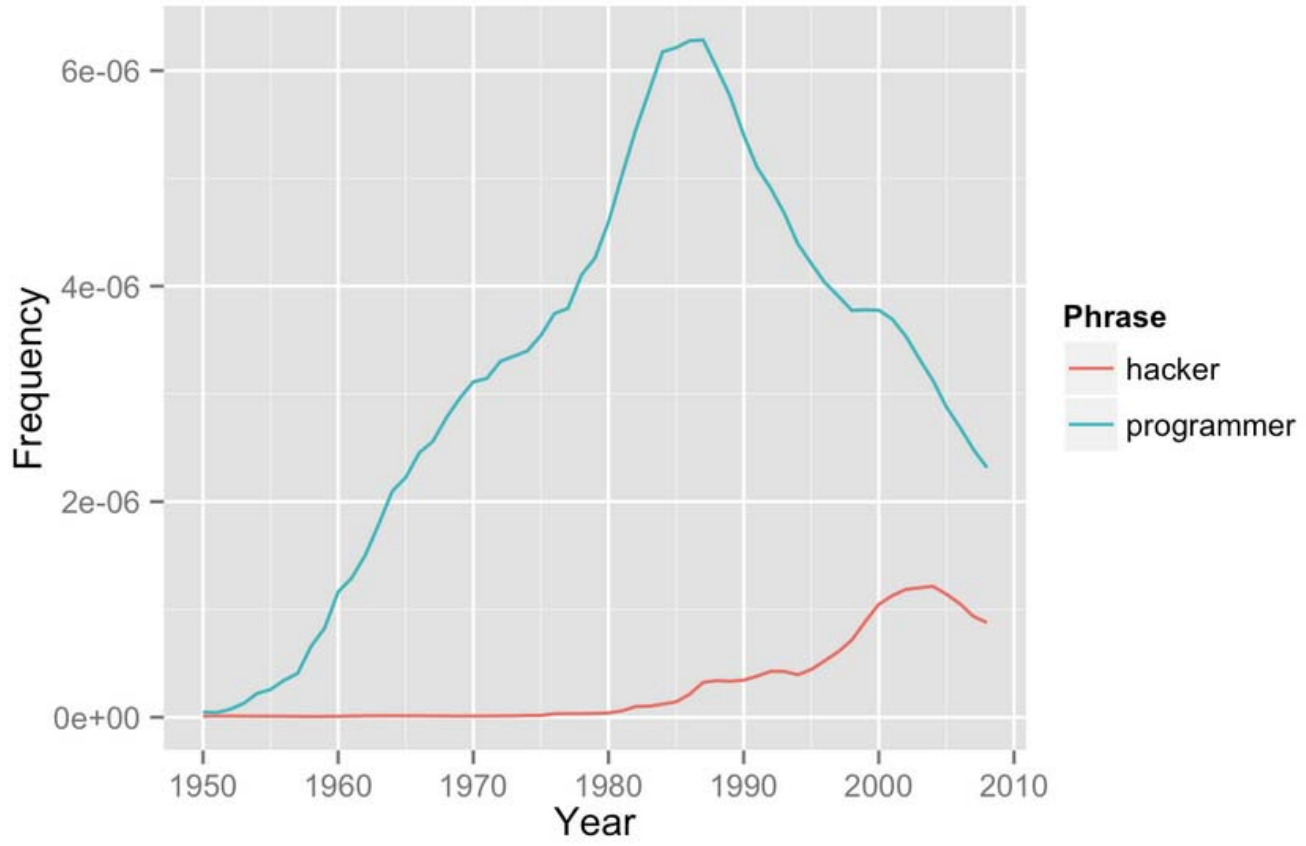

English

Version 20120701
total_counts

1-grams 0 1 2 3 4 5 6 7 8 9 a b c d e f g h i j k l m n o other p pos punctuation q r s t u v w x y z

- Ukázka datové struktury
  - circumvallate 1978 335 91
  - circumvallate 1979 261 91

Požadovaný výstup



INSTITUT
BIOSTATISTIKY
A ANALÝZ
Masarykova univerzita

Dotazy sem:

komenda@iba.muni.cz
karolyi@iba.muni.cz