Strojové učení v knihovnách

Petr Žabička a Alžbeta Zavřelová
Moravská zemská knihovna v Brně

Konference ESK 2021, 18.-19.5.2021

Umělá inteligence (AI)

Neexistuje jednoznačná oficiální definice

- Není to to co známe ze sci-fi literatury
- Těžko lze na první pohled poznat co je pro současnou umělou inteligenci snadné a co je složité

Autonomie - schopnost provádět úkoly ve složitých prostředích bez řízení uživatelem.

Adaptabilita - schopnost zlepšit se prostřednictvím učení se ze zkušeností

Nejedná se o učení, porozumění a inteligenci v jak ji definujeme u lidí.

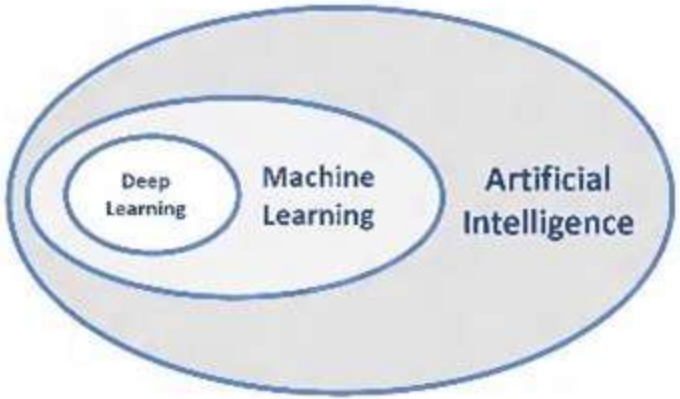
Slabá AI - na úzce definovaných úlohách, dosahuje lepší výsledky než lidé

Silná AI - obecné úlohy, zatím neexistuje a nikdo neví jak je to daleko

Strojové učení (ML)

Strojové učení (machine learning)

- podoblast umělé inteligence
- systémy, zlepšující svůj výkon v dané úloze na základě narůstajících zkušeností nebo objemu dat
- statistika, pravděpodobnost (četnost)



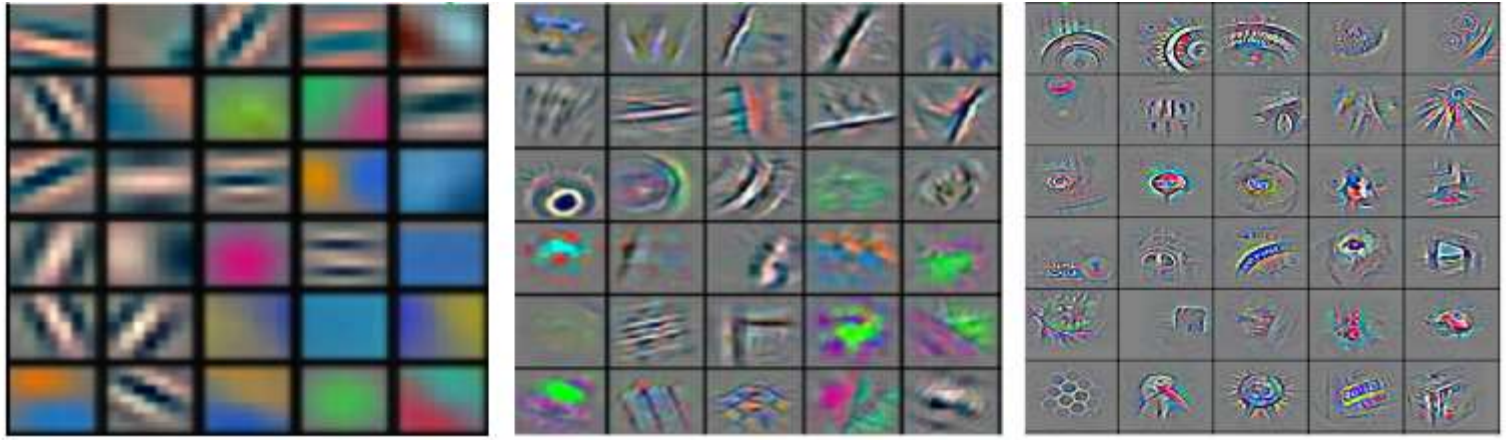
Hluboké učení (deep learning)

- podoblast strojového učení, princip neuronových sítí
- komplexnější matematické modely přinášející novou kvalitu
- data jsou (paralelně) analyzována množstvím jednoduchých jednotek, tj. rozdělení problémů na řadu menších, dobře vymezených úloh
- s výhodou lze použít výpočtů na grafických kartách

Hluboké učení

Hluboké učení - konvoluční neuronové sítě

- strojové zpracování obrazu v několika úrovních detailu
- v každé úrovni se postupně vyhodnocují části obrazu počínaje jednoduchými body nebo hranami až po abstraktnější prvky



Generativní adversariální sítě

- dva systémy postavené proti sobě
- jeden generuje obrázky na základě naučených modelů
- druhý se snaží rozlišit skutečné od strojem generovaných obrázků
- postupným zlepšováním se obou sítí vznikají obrázky nerozeznatelné od skutečnosti

Způsoby učení

- Učení s učitelem
  - součástí trénovacích dat jsou i informace o tom co má být výsledkem
- Učení bez učitele
  - kategorizace, není informace o tom co má být výsledkem
- Zpětnovazební učení
  - stroj hledá optimální řešení a dostává ohodnocení výsledku

Vždy jsou potřeba velké sady dat pro učení

Využití strojového učení

- klasifikace dat
- doporučovací systémy
- predikce budoucích výsledků
- řízení systémů

Strojové učení v knihovnách

Možnosti využití

- automatická klasifikace dat
- tvorba nebo obohacování metadat
- zefektivnění a zjednodušení vyhledávání
- doporučovací systémy
- analýza obrazu - anotace obrazu, OCR

digitalizáty + přepisy textů → učící modely

- data trénovací - učí se z nich (ukázkové data)
- data testovací - testuje/generuje naučené
- částečná chybovost, nutné i manuální korekce, korigované dokumenty mohou rozšířit trénovací datovou sadu

Komunita AI4LAM

AI4LAM (AI pro knihovny, archivy a muzea)

- platforma pro spolupráce paměťových organizací v oblasti AI
- zlepšuje povědomí a sdílí zkušenosti ve využití nástrojů umělé inteligence

- inovativace a inspirace - AI může změnit způsob, jakým LAM plní své poslání
  - obohacení globální praxe
- etické a transparentní nasazování AI
- řešení výzev v rámci komunity LAM
- otevřená data, sdílené modely a opakovatelné výsledky
- spolupráce jednotlivců a institucí (sdílené hodnoty)
- otevřená mezinárodní komunita napříč organizacemi

Komunita AI4LAM

Pravidelné setkání

- Zoom online jednou za ca. 6 týdnů, obvykle v 17 nebo 18 hod. našeho času, k dispozici zápisy a nahrávky jednání
- globální + pracovní skupiny + od 2020 neformální skupina koordinátorů pro AU a NZ

Členové

- správci sbírek GLAMR (galleries, libraries, archives, and museums, records management) a výzkumníci v Digital humanities/digitálních humanitních vědách
- péče o kulturu, etiku, gramotnost a roli AI v těchto souvislostech
- hlavní koordinátoři a dobrovolníci

Komunita AI4LAM

Pracovní skupiny v rámci komunity - koncepty:

- projektová dokumentace - školení/workshopy - soutěže/výzvy - etika
- sdílené modely - sdílené datové sady - produkční AI - uživatelské rozhraní
- periodika/noviny, ...

Komunikační kanály:

- Google groups: https://groups.google.com/g/ai4lam
- Slack: https://ai4lam.slack.com/
- GitHub: https://github.com/AI4LAM
- Twitter: https://twitter.com/ai4lam

Akce AI4LAM

Bezplatný online workshop

- 25.5.2021: AI-enabled GLAMR practice: the technical landscape, https://www.eventbrite.com/e/ai-enabled-glamr-practice-the-technical-landscape-tickets-152622685561

Online + fyzická konference

- **9.-10-12.2021: Fantastic Futures, 3rd International Conference on Artificial Intelligence for Libraries, Archives and Museums** - hlavní konference
  - výukové programy nebo interaktivní workshopy: úvod do aplikace, případy užití a implementace AI v GLAM nebo nástroje použitelné na data a sbírky GLAM
  - prezentace: zkušenosti s implementací AI v GLAM, výuka AI, etika AI v GLAM, vytváření a sdílení datových sad a tréninkových modelů v rámci komunity
  - https://easychair.org/cfp/FantasticFutures21

Visual Geometry Group (Oxford University)

- datasety paměťových institucí, využívají kvalitní metadata z knihoven
- počítačové vidění - zejm. rozpoznávání obrazu/textu:
- klasifikace, porovnávání a vizuální variace
  - tiskařské a typografické chyby, porovnávání ilustrací/exemplářů, zakřivené stránky
  - vizuální vyhledávání tiskařských štočků, motivů
- řeší výzkumné otázky z humanitních věd
- nástroje pro výzkum i zábavu

https://www.robots.ox.ac.uk/~vgg/
https://www.youtube.com/watch?v=Ku7IZ1lNICk

Bodleian Ballads Search

Query Image

name: MS. Wood E 25(95)

Search Results 1 to 10

MS. Wood E 25(95)

Detailed matches

MS. Wood E 25(40)

Detailed matches

MS. Wood E 25(29)

Detailed matches

Douce Ballads 2(260a)

Detailed matches

MS. Wood E 25(43)

Detailed matches

4o Rawl. 566(55)

Detailed matches

4o Rawl. 566(76)

Detailed matches

Douce Ballads 2(262a)

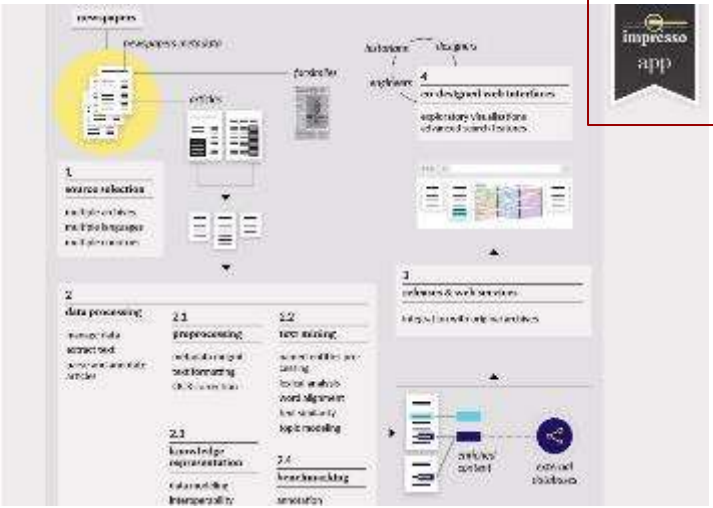
Detailed matches

Projekt impresso

impresso
Media
Monitoring
of the Past

Impresso - mediální monitorování minulosti: vytěžování 200 let historických novin

- text mining sbírek periodik - vyhledávání relevantních zdrojů
- technologický rámec pro extrakci, zpracování, propojení a prozkoumávání dat z historických tiskovin
- interdisciplinarita - spolupráce výpočetních lingvistů, digitálních humanistů, designérů, historiků, knihovníků a archivářů
- strojově čitelný obsah OCR + automatické prozkoumávání obsahu novin



https://impresso-project.ch/project/objectives/
https://www.youtube.com/watch?v=fuGYc_svLXg

Projekt impresso

impresso
Media
Monitoring
of the Past

1) Sada nástrojů pro sledování historických médií

- vícejazyčné a časově specifikované techniky text miningu
  - obsah pro sémanticky indexovaná, strukturovaná a propojená data
- řada komponent pro zpracování přirozeného jazyka (NLP) pro postkorekci OCR, indexování n-gramů, distribuční sémantické indexování, zpracování pojmenovaných entit a kategorizaci textu a shlukování

2) Vizualizační rozhraní a vizuální analýza pro průzkum obsahu

- aktivní průzkum a kritická analýza novinových korpusů - složitých historických dat
  - vyhledávací funkce - klíčová slova a fazety

3) Digitální historie - doplněk cílů

- implementace a dopad vyvinutých nástrojů (vyzdvihování národních a kulturních specifik)

Knihovny.cz a strojové učení

V rámci projektu NAKI byly vytvořeny nástroje využívající sémantické technologie a strojové učení a jejich výstupy byly integrovány do portálu

- otestování využitelnosti těchto nástrojů při zpřístupnění knihovních katalogů
- obohacení záznamů pro účely vyhledávání
- sémantické nástroje byly zaměřeny na obohacování bibliografických záznamů o entity
- nástroje byly aplikovány vždy jen na určitou část záznamů (tam, kde to bylo relevantní)
- pracuje se nejen s bibliografickými záznamy ale i s obsahy nebo plným textem dokumentů
- identifikované entity (osoby, místa apod.) byly v indexu označeny tak, aby při jejich hledání měl daný dokument větší relevanci
- samostatně byly identifikovány osobní jména v textech obsahů dokumentů (pravděpodobní autoři článků nebo kapitol)

KNIHOVNY.CZ

Knihovny.cz a strojové učení

Automatický klasifikátor

- první experimenty - rozdělení na beletrii a odbornou literaturu
  - problém byly např. čítanky
- přiřazení třídy konspektu
  - bylo identifikováno několik nejednoznačných tříd = tříd, u kterých váhá i knihovník
  - došlo k redukci (zjednoznačnění) tříd konspektu pro tento účel
  - bylo nutné opravit v trénovacích datech neplatná MDT a nesprávně přidělené znaky konspektu
- pro publikace v češtině
- původně s využitím fulltextu, později jen na základě bibliografického záznamu
- pro některé třídy konspektu chybí dostatek trénovacích dat
- velmi přesné tam kde bylo v záznamu MDT
- nakonec obohaceno konspektem 294 459 záznamů monografií
- využití zejména při filtrování (faseta Obor = konspekt)

Knihovny.cz a strojové učení

Nejčastější záměny při klasifikaci s využitím fulltextu (na základě testovací datové sady)

Originální kategorie	Kategorie přiřazená klasifikátorem
Dějiny zemí střední Evropy	Dějiny Česka a Slovenska
Malířství	Výtvarné umění
Řízení a správa podniku	Management. Řízení
Literatura. Literární život	Česká literatura (o ní)
Výtvarné umění	Malířství
Geografie Česka a Slovenska, reálie, cestování	Dějiny Česka a Slovenska
Umění	Výtvarné umění
Biografie	Vnitropolitický vývoj, politický život
Biografie	Film. Cirkus. Lidová zábava
Dějiny Česka a Slovenska	Dějiny zemí střední Evropy

ObalkyKnih.cz - doporučování literatury

- **Vychází z dat o výpůjčkách uživatelů z několika knihoven (anonymizované údaje o uživatelích a jejich výpůjčkách)**
  - sbírají se základní údaje o dokumentu plus hash identifikátoru uživatele a volitelně další údaje o uživateli (pohlaví, věk, typ čtenáře, PSČ) - nyní z JVK, MZK, KVKLI, KJM, SVKKL
  - využívá se i anotací knih
  - na základě těchto údajů vzniká model, který se použije pro doporučování knih
- **Doporučení na základě titulu**
  - nejpoužívanější
  - není potřeba zasílat data o uživatelích
- **Doporučení na základě čtenáře**
  - přesnější, pracuje s historií výpůjček přihlášeného uživatele
- **Doporučení na základě preference**
  - doporučení dokumentu ze zvolených kategorií konspektu, případně v kombinaci s historií výpůjček čtenáře

Projekt PERO

PERO - Pokročilá extrakce a rozpoznávání obsahu tištěných a rukou psaných digitalizátů pro zvýšení jejich přístupnosti a využitelnosti (2018–2022)

- automatické kontroly kvality a zlepšování kvality digitalizátů (mikrofilmy)
- automatické rozpoznání textu OCR u starých tištěných dokumentů (fraktura i antikva)
- polo-automatický přepis ručně psaných dokumentů

Nástroje a technologie pro zpřístupnění obsahu digitalizovaných historických dokumentů

- zvýšení čitelnosti dokumentů, snazší vyhledávání a využití obsahu
- možnosti strojové indexace a fulltextové vyhledávání, analýza obsahu

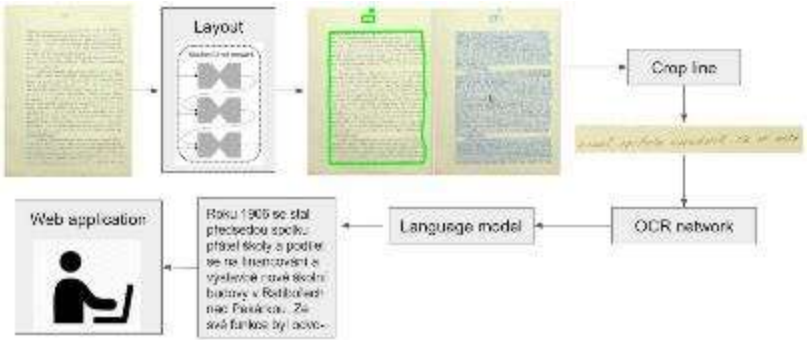
Projekt PERO

Metody: počítačové vidění, strojové učení a jazykové modelování

- rozpoznávání tištěného/psaného textu, opravy obrazu (neuronové sítě)

- analýza rozvržení stránky
- detekce řádků
- automatický přepis řádků
- jazykové modely:
  - přiřazování pravděpodobnosti větám
  - umožňuje vybrat nejlepší možnost

  (+dotrénovává na starší jazyk)

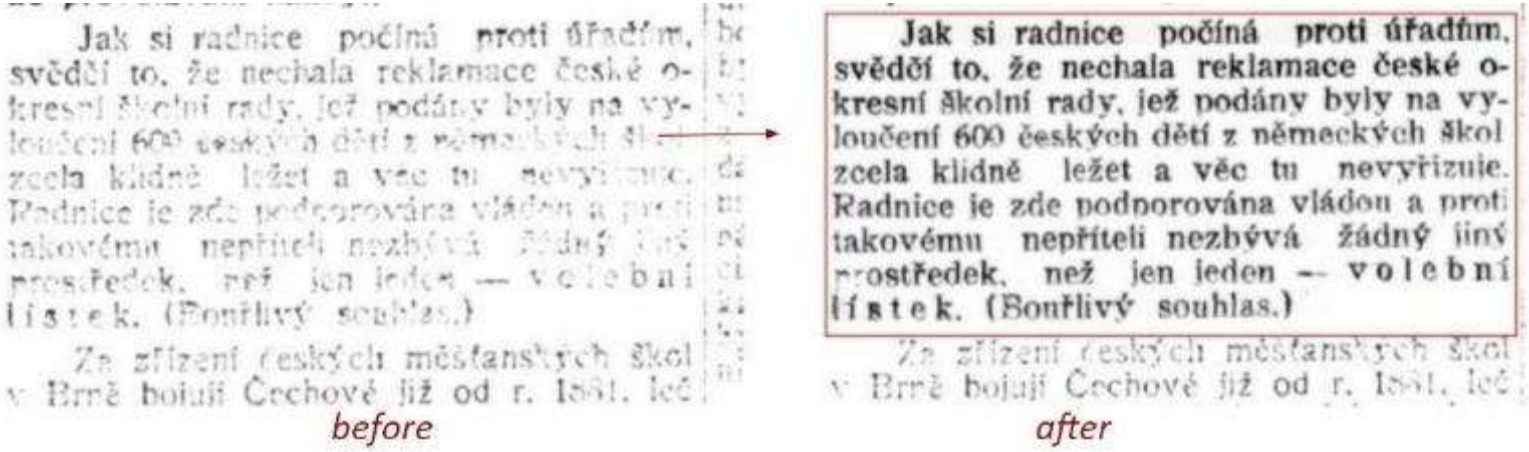


z hlediska jazyka

Projekt PERO

Zlepšování kvality digitalizátů

- zejména digitalizované mikrofilmy
- nízká kvalita skenování, vysoké riziko poškození materiálu
- opravy obrazu pro optimalizaci čitelnosti (konvoluční neuronové sítě)



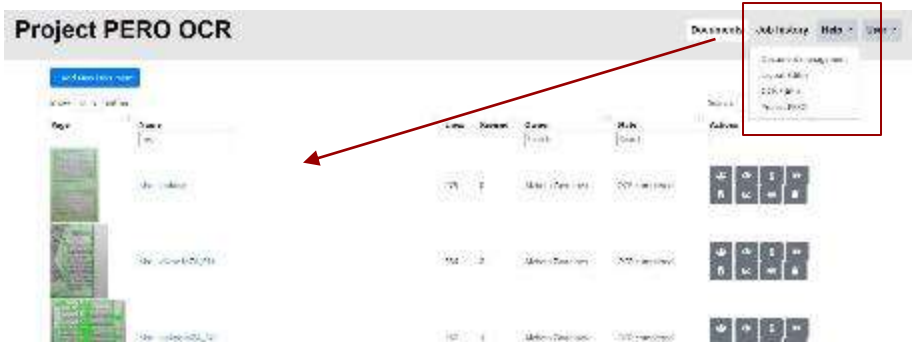
PERO-OCR aplikace

OCR online aplikace

- aplikace umožňuje uživatelům práci s vlastními dokumenty
- manuální opravy - korekce dle řádků předlohy, problematické části zvýrazněné
  - ruční opravy nutné k opakovanému trénování modelů
- formát TXT, ALTO, ALTO XML

Budoucnost:

- nasazení do digitalizační linky prostřednictvím API (nyní testovací provoz s ProArcem)



PERO-OCR aplikace

- ukázky OCR výstupů v online aplikaci (bez zásahů)
  - tiskoviny, staré tisky
  - rukopisy moderní (i histor.)

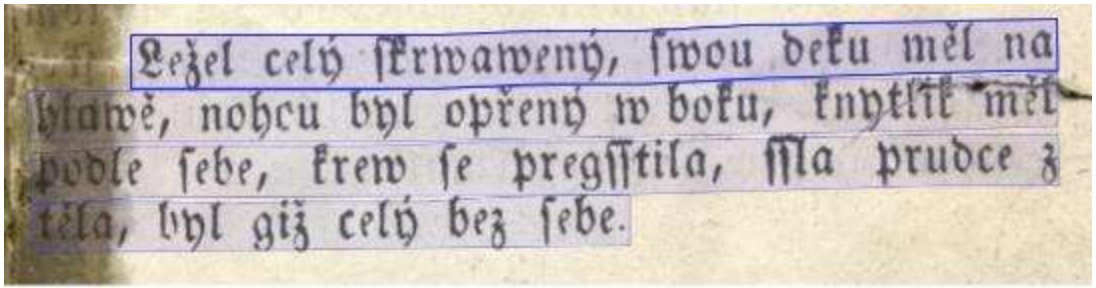

LIDOVÉ NOVIN

ROČ. 33. ČÍS. 105. V SOBOTU 28. ÚNORA 1925 RÁNO

Německo zdrojem evropského nebezpečí.

Úsilí o dohodu v koalici.

Před návratem Švehlovým žádné změny. — Lidovci doufají ve společnou frontu katolickou.


Leżel celý ſkrwawený, ſwou deku měl na hlawě, nohcu byl opřený w boku, knytlik měl podle ſebe, krew ſe pregſſtila, ſſla prudce z těla, byl giž celý bez ſebe.

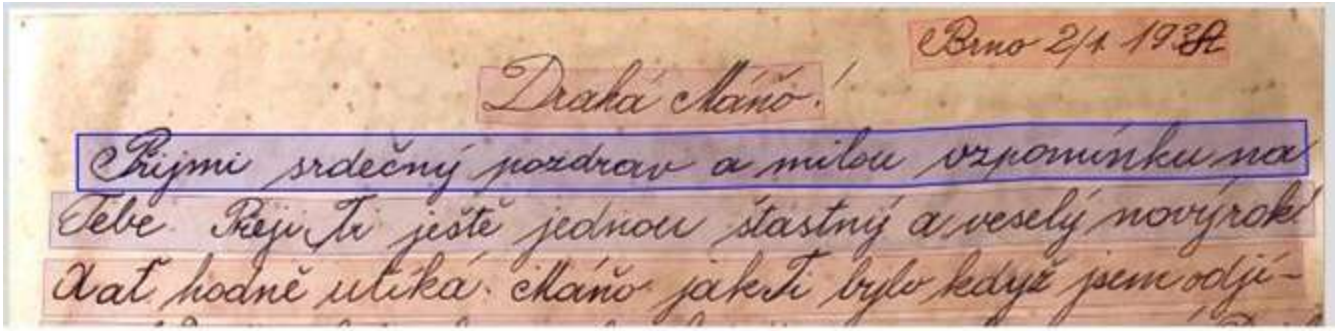
Ležel celý ſkrwawený, ſwou deku měl na hlawě, nohou byl opřený w boku, knytlik měl podle ſebe, krew ſe pregſſtila, ſſla prudce z těla, byl giž celý bez ſebe.

8

J - Praha 27. února.
Včerejší usnesení lidoveckého klubu znamená vítězství umírněného směru Šrámkova ve straně, který se zračí i v tom, že do ohlášeného

rozpoznávání tištěných textů

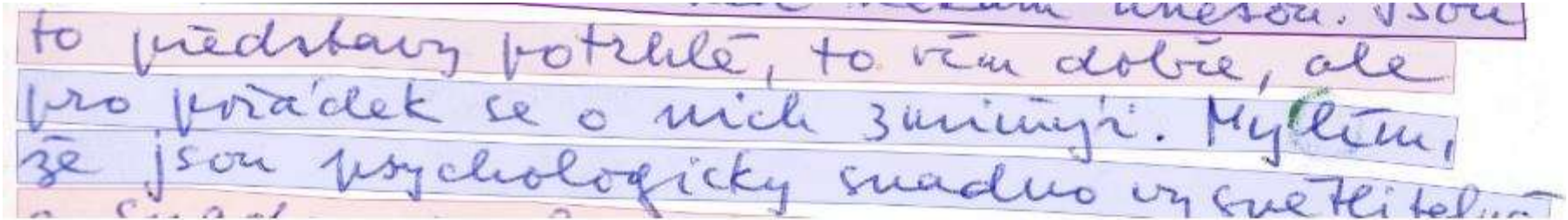
- digitalizované noviny
- staré tisky, kramářské písně (fraktura, antikva)



Brno 2/1. 1939
Drahá Máňo!
Přijmi srdečný pozdrav a milou vzpomínku na
Tebe. Přeji Ti ještě jednou šťastný a veselý nový rok.
Ať hodně utíká! Máňo jak Ti bylo když jsem odjí-

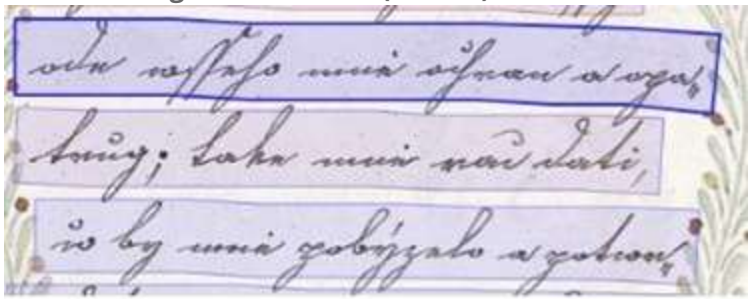
rozpoznávání rukopisných textů (20.st)

- korespondence 1938 a V. Havel



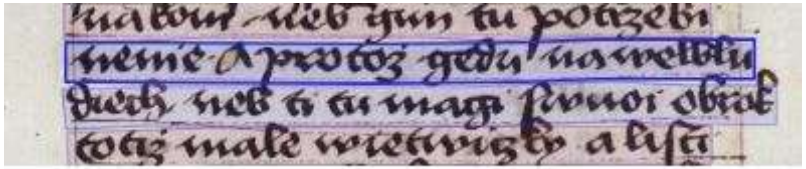
to představy potřísně, to vím dobře, ale
pro pořádek se o nich zmiňuji. Myslím,
že jsou psychologicky snadno vysvětlitelné

česká novogotická kurzíva (kurent)

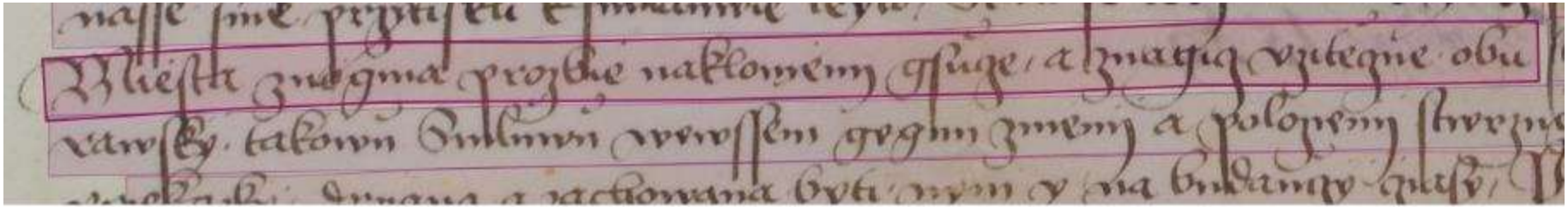


ode wſſeho mně ochran a opa„
trug; take mně rač dati,
co by mně pobýzelo a potwr„

experimentální rozpoznávání historických rukopisů



nenie A protož gedu na welblu
diech neb ti tu magi swuoi obrok
totiž male wietwiczky a listi



Mieſta znagma prozlie naklowieny gſuge, a nagio vtegne obu
u Omlunu wewſſem gegen, zmeni a Polonen ſtwrzngem a vpewnugem Kterazto Cmluwli gia odnas y bida
gmnych wſſech ledy wſſelwakt

Na co si musíme dát pozor

Nedostatečné množství dat k učení

- systém se musí mít na čem učit

Nepředpokládaně nesprávná interpretace trénovacích dat

- systém funguje na testovacích datech správně, ale identifikuje jinou vlastnost než je očekáváno

Hranice mezi zlepšením a podvrhem

- oprava vs. změna nečitelného textu

Předpojatost (BIAS)

- trénovací data obsahují předpojatost, ta se přenáší i do výsledků
- kdo rozhoduje co je předpojatost a co ne? Realita vs. politické rozhodnutí.

Děkujeme za pozornost

<Petr.Zabicka@mzk.cz>
<Alzbeta.Zavrelova@mzk.cz>