Využití metod strojového učení pro určování autorství českých textů z přelomu 19. a 20. století

Jan Pokorný a František Válek

Národní knihovna ČR

Problém určování autorství

Určování autorství anonymních nebo sporných textů je dlouhodobě významným tématem v literární vědě.

- historické texty bez jasné atribuce
- neúplné bibliografické záznamy v digitálních archivech
- potřeba ověření atribuce starých děl
- detekce plagiátů a neoprávněných autorství

Přesné určení autorství umožňuje nejen korekci nesprávné atribuce textů, ale i hlubší analýzu stylového vývoje jednotlivých autorů a lepší pochopení literárních dějin.

Zvolená metoda

Předpoklady:

- Každý autor má unikátní jazykový styl.
- Styl zůstává stabilní i přes různé témata.
- Strojovým učením lze natrénovat tyto vzory.

⇒ **Analýza statistických charakteristik textu** (jazykový styl, nikoli obsah)

Technické řešení: korpus

- vybrané knihy: v 1. fázi 23 českých autorů, v 2. fázi 325 autorů, počet se dále rozrůstá
- období: přelom 19. a 20. století
- zdroj: Národní digitální knihovna
- autoři: Čapek, Neruda, Jirásek, Heyduk atd.

Technické řešení: příprava dat

- digitalizace: skenování a OCR vybraných knih
- čištění autorských textů od jiných částí knih: odstranění poznámek, obsahu, editorských poznámek
- normalizace: jednotné kódování UTF-8
- segmentace: rozdělení na segmenty po 1000, 500, 200 a 100 slovech

Technické řešení: zpracování

Delexikace:

- nahrazení významových slov (autosémantika – podstatná jména, přídavná jména, slovesa, příslovce) slovním druhem (POS tagy), např. „pes běžel rychle“ se převede na tagy: NOUN, VERB, ADV,
- ponechána funkční slova (synsémantika – předložky, spojky, zájmena, částice), např. „v“, „a“, „který“, „že“ zůstanou, protože funkční slova se používají velmi často a jejich frekvence je velmi typická pro styl autora,
- generování n-gramů: kombinace slov pro detekci vzorů

Klasifikace textu pomocí metody Support Vector Machine (LinearSVC klasifikátor):

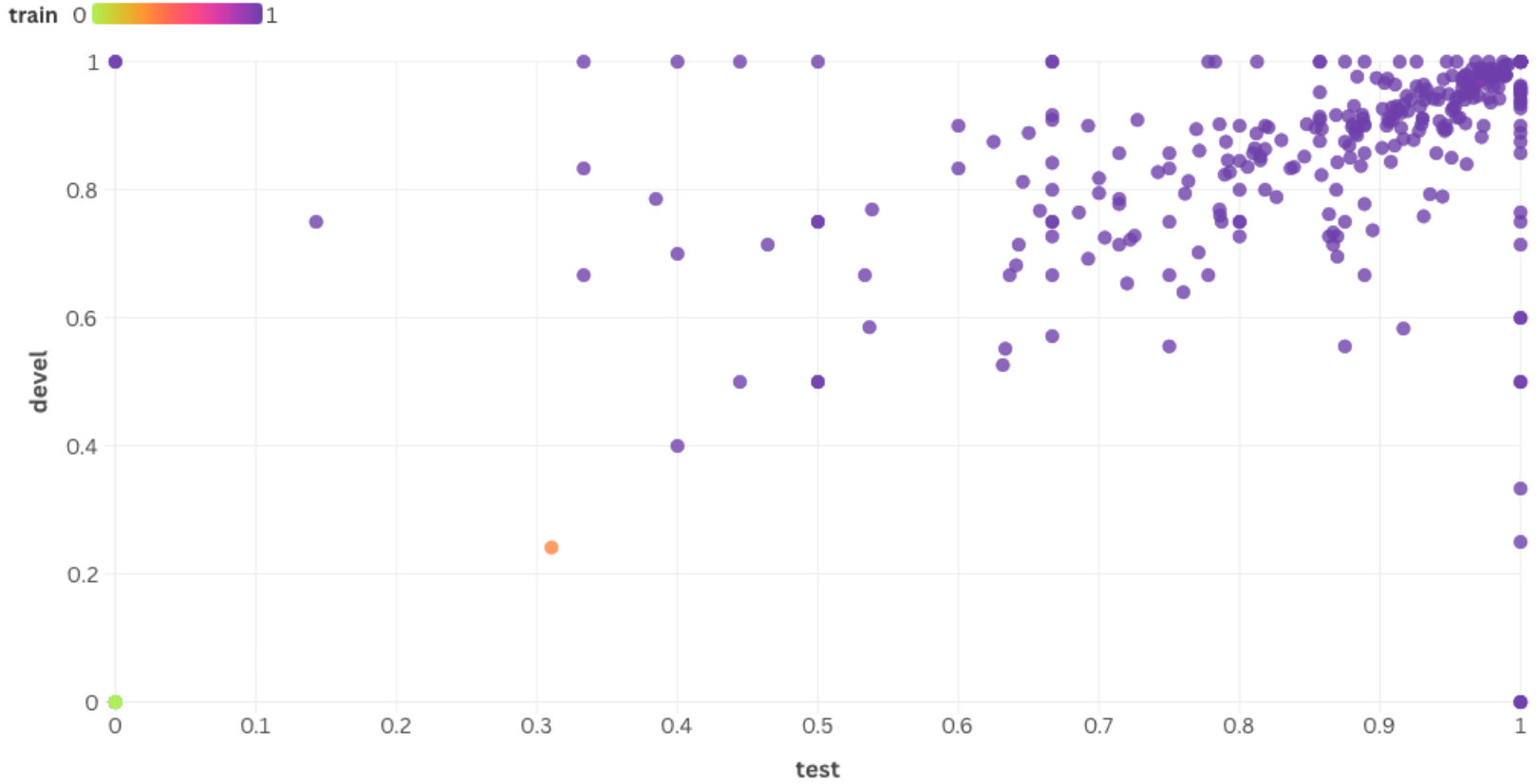
- trénování: 60% textu daného titulu
- testování: zbylých 40% textu (20% ve 2 nezávislých sadách)
- vyhodnocení: % přesnost rozpoznání autorů

Výsledky a zjištění

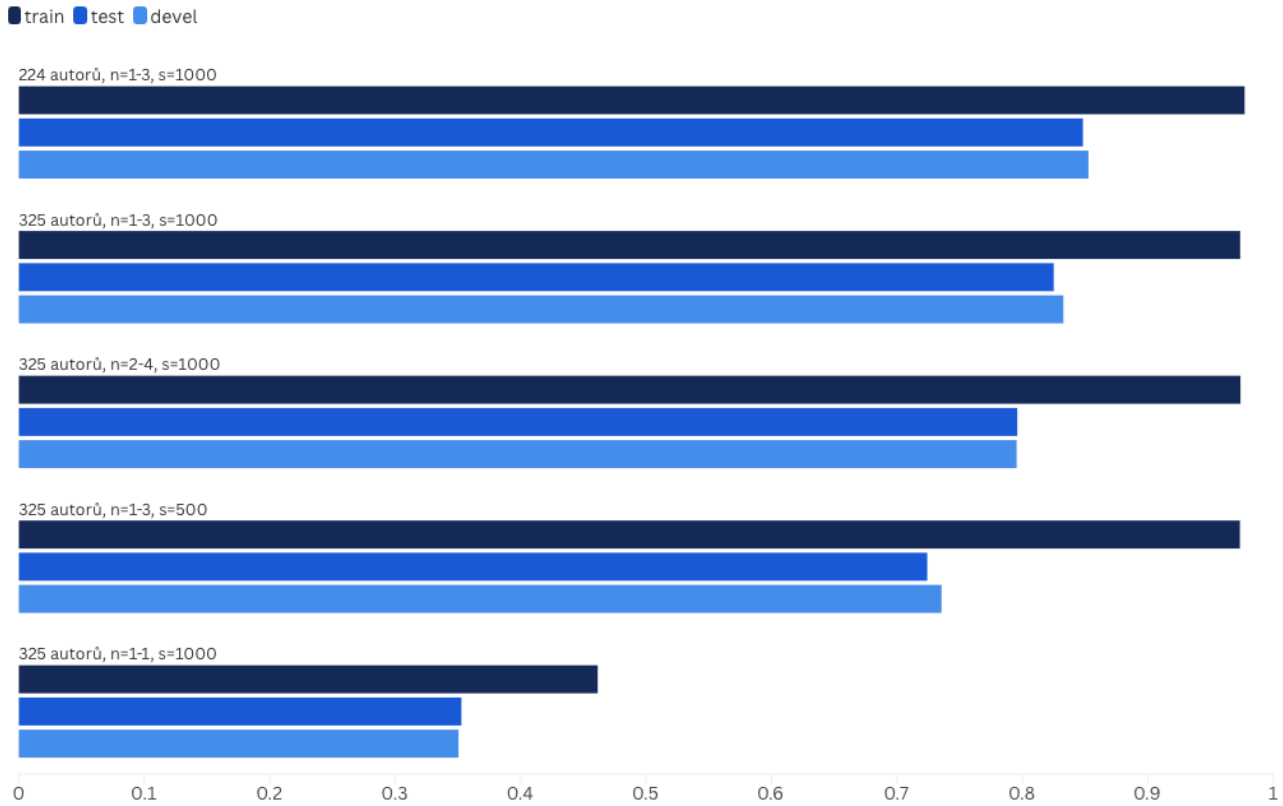
- přesnost je dána délkou textového originálu
- při 1000 slovech dosahuje přesnost 96% u 23 autorů, 82% u 325 autorů
- autoři s konzistentním stylem jsou lépe rozpoznatelní
- u kratších textů s 50 slovy klesá přesnost na 42%

Někteří autoři mají značné rozdíly mezi jednotlivými díly, tj. mají nízkou konzistenci stylu ⇒ ztížené určování autorství

Výsledky testování pro jednotlivé autory



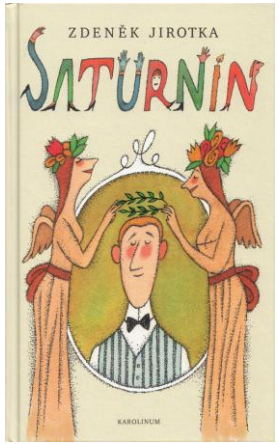
Výsledky testování na sadách train, test a devel



Originál vs. pastiš

Lze rozlišit skutečného autora od imitátora?

- test: Zdeněk Jirotka vs. Miroslav Macek (pokračovatel)
- výsledek: 100% přesnost u delších textů
- závěr: model rozpozná i stylové napodobení



Praktická aplikace

- Detekce plagiátů: najít neoprávněné kopie
- Verifikace metadat: ověření správnosti autorství v digitálních knihovnách
- Identifikace anonymních textů

Produkt AuthorGuesser