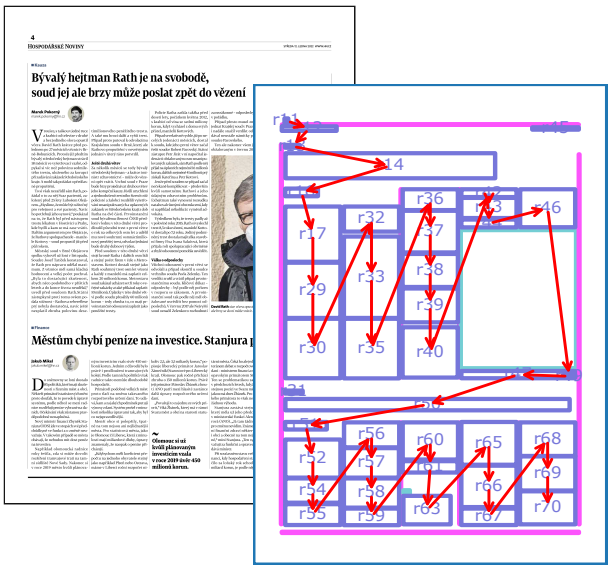
Uspořádání zpřeházených řádků s pomocí jazykového modelu

autor: Michael Holubec
vedoucí práce: Ing. Karel Beneš

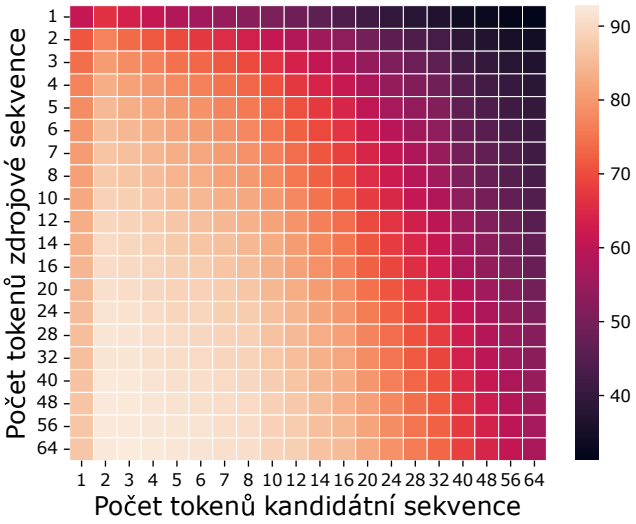
23

Reading order

- Posloupnost čtení (Reading order) definuje pořadí textových prvků na stránce
- Správná identifikace je klíčová při digitalizaci dokumentů a rekonstrukci toku textu
- Různé druhy rozložení stránek, kniha vs novinový článek
- Implementovány tři přístupy
  - Jazyková analýza (text)
  - Prostorová analýza (geometrie)
  - Kombinovaná analýza (geometrie i text)
- Vyhodnocení na 13 novinových stránkách, Recall
  K: 88,9 %, P: 84,9 %, J: 16,1 %



Obrázek 1. Znázornění identifikace čtení na novinové stránce, výsledek kombinované analýzy



Obrázek 2. Heatmapa jednotlivých konfiguracích a úspešnosti výběru kandidáta, každá dlaždice 10 000 testů, výběr ze 4 kandidátů, v procentech

Jazykový model

- Model schopný určit pravděpodobnost sekvence o dané skladbě slov
- LSTM architektura
- Experimenty pro zjištění vhodné konfigurace
- Čím více tokenů zdrojové sekvence, tím lepší odhad pravděpodobnosti návaznosti kandidátní sekvence
- Jazykový model použit u jazykové a kombinované analýzy