Extrakce příznaků a klasifikace komentářů

Jakub Drdák

Zadání

- Vytvořit **lokální kopii** databáze s přístupem přes **webový portál**
- Extrahovat příznaky
- **Klasifikovat** komentáře do 4 tříd
  - *mimo kvalifikaci*
  - *střet zájmů*
  - *nedostupné materiály*
  - *ostatní*

Klasifikace komentářů

- **Střed zájmů**
  - 'I have a conflict of interest.'
  - 'Knihu nemohu hodnotit, neboť jsem jejím spoluatorem a editorem.'
- **Mimo kvalifikaci**
  - 'Publikace je mimo okruh mé specializace, nehodnotím.'
- **Nedostupné materiály**
  - 'Hodnocení nebylo možno provést vzhledem k nedostatku poskytnutých podkladů. '
- **Ostatní**
  - 'A below-median IF journal (out of the 140 sociology journals in the WoS).'
  - '43 social psychology -- ok, B. '
    'good B. weak journal. almost no citations. however, author provides original theory of agency/structure. '

Technologie

- Python
- Extrakce příznaků: regex
- Klasifikace: 1-hot encoding (word2vec), logistická regrese/SVM
- Výsledky
  - **Accuracy**:
    - log. regrese: 89%
    - SVM 80%
  - **Příznaky**:
    - Mimo kvalifikaci Střet zájmů Nedostupné
      1. mimo jsem available
      2. nemám interest není
      3. kvalifikaci conflict not

Děkuju za pozornost