Ekstrakcija podataka iz novozelandskih pravnih dokumenata

Jelena Matković

Novi Sad, 2024.

Zadaci

- ekstrakcija metapodataka
- pronalaženje citiranih precedenata i regulativa
- povezivanje subsekventnih odluka/presuda
- određivanje zakonskih regulativa na koje se presude odnose

Ekstrakcija metapodataka

- skrejpovanje presuda i zakona sa sajtova
- *PDF* čitači i *OCR* za preuzimanje tekstualnog sadržaja
- kombinacija regularnih izraza sa ekstrakcijom pomoću pozicije unutar dokumenta
- ispravka grešaka *OCR-a*

Pronalaženje citata i povezivanje dokumenata

- regularni izrazi za identifikacione oznake dokumenata i neutralne citate
- određivanje da li se citati odnose na precedente ili subsekventne presude pomoću poređenja imena učesnika

Određivanje regulativa

- ekstrakcija naziva regulativa iz teksta pomoću regularnih izraza
- neuronske mreže za klasifikaciju teksta
- prva dva paragrafa presuda korišćena kao podaci za treniranje mreže
- nazivi regulativa korišćeni kao labele

Rezultati i diskusija

- Nedostupnost određenog broja dokumenata onemogućava povezivanje po vremenskoj liniji
- Preuzeto preko 80% metapodataka
- Mali broj dokumenata sadrži nazive regulativa u samom tekstu, pa samim tim nema dovoljno podataka za treniranje mreža
- Nekonzistentnost iziskuje veliki broj šablona kako bi se preuzele informacije od značaja