DATA SCRAPING, DATA MINING, DATA LEARNING

Case Studies for the Report on the EU Copyright Implications of Training Data for Machine Learning

Dr Pinar Oruc
Research Associate, CREATe and School of Law
University of Glasgow

Introduction

- The project
  - *reCreating Europe: Rethinking digital copyright law for a culturally diverse, accessible, creative Europe* (Horizon 2020)
- Our scope
- Chosen case studies
  - Data Scraping, Natural Language Processing, Computer Vision
- Divided into stages
  - For copyright analysis purposes

www.recreating.eu

Data from copyright perspective

What is being used?

- How much of the needed data fall under the scope of subject matter of copyright law?

What is being done?

- Do these activities fall under the exclusive rights of the rightholder – and therefore require authorization?

How is this achieved?

- Is there a contract and what are the terms?
- Can it be covered by a copyright exceptions?

Data scraping

- **Data Collection**
  - Types of data available
  - Collection methods
  - Potentially affecting the website - distributing the requests
  - Terms of Service prohibiting scraping
- **Data Processing**
  - Editing, cleaning and imposing own structure
  - Enrichening own dataset with the data scraped by others
- **Data Analysis and Outputs**
  - How much of the data is visible in the outputs?
  - Non-display and non-commercial
  - Resharing datasets with fellow researchers

Natural language processing

- **Data Collection**
  - Types of works collected
  - Only targeting freely available text datasets
- **Pre-processing**
  - Converting and cleaning
  - Tasks such as tokenization and normalization
- **Training**
  - Supervised/unsupervised
  - Using pre-trained models by big companies
- **Trained Model**
  - How much of training data is visible in the trained model?
  - How to utilise the trained model? Commercial and non-commercial purposes

Computer vision for content moderation

- **Data Collection**
  - Source of images
  - Annotated datasets such as ImageNet
- **Pre-Processing**
  - Resizing, cropping, converting colour, rotating
  - Augmentation to increase datasets
- **Training**
  - Supervised
  - Unsupervised
- **Trained Model used for Content Moderation**
  - Human involvement in final decisions
  - Commercial uses for content moderation AI

Copyright concerns & next steps

Repeated problems:

- The necessary data might include copyright protected works
    - How much of it is protected?
    - Uncertainty about SGDR
    - Databases and contract law (Ryanair case)
- Reproduction right is affected
    - Not always temporary → to benefit from InfoSoc 5(1)
    - Unclear how much of the work is reproduced
- Adaptation right is affected
    - Processing stages usually include some kind of change in the main work
    - Adaptation is not harmonised in the EU
- How much of the work is in the final output? (trained model)
- Desired uses not easily covered by existing copyright exceptions
    - Even the new Text and Data Mining exception under the CDSM is insufficient – limited beneficiaries and purposes (Art 3), rightsholders opting out (Art 4).

-End of presentation-
Thank you

pinar.oruc@glasgow.ac.uk

CREATe

ReCreating Europe