Clifford Chance
Data Scientist – Synthetic Data Generation for Legal Domain

About the Organisation

Synthetic data is used as a substitute when real world data is inconvenient to get, expensive or constrained by regulation. However, there is an open question on its robustness and how well it reflects real-world data characteristics. Gartner estimates that by 2030, synthetic data will completely overshadow real data in AI models. This assignment is about to investigate its impact on the legal domain and how it can be used to support business operations.

Clifford Chance LLP (CC) is an international law firm headquartered in London, United Kingdom. It is a member of the "Magic Circle", a group of the most prestigious London-based multinational law firms. It ranks as one of top ten largest law firms in the world measured both by number of lawyers and revenue. Being a thought leader within the legal domain, the firm invests considerable effort in using the latest and greatest technology.

Our Data Science Lab was established 6 years ago and has been working with several cutting-edge technologies (Natural Language Processing, Spark, DevOps, MLOps, Cloud Technologies, AutoML, explainable AI…). Use cases are ranging from document discovery, natural language understanding, question answering, cost optimization, classification of time recording narratives. Respecting the principles of ethical AI is one of our values.

The team consists of 11 individuals with various backgrounds and covering the whole spectrum of the trade: data engineering, data science and ML engineering, data visualization and operationalization.

Role Description and Responsibilities

The person appointed to this role will work on the following:

- Understanding the state of the art in synthetic data with special focus on legal domain use cases
- Cleaning, aggregating, and interpreting datasets like time reporting narratives and legal contracts
- Partnering with the members of the Data Science Lab to build synthetic datasets that can be used as open-source benchmarks for different machine learning tasks
- Understanding how synthetic data can be generated without disclosing sensitive information
- Gaining experience with latest and greatest technologies including Databricks, Spark, Delta Lake, cloud services, cutting-edge NLP approaches

The goal of the project is to generate synthetic datasets that can be used for machine learning tasks whilst preserving sensitive business information. Ideally such datasets can be open sourced to facilitate the development of AI in the legal domain.

Majority of the effort shall be spent on developing algorithms to reach this goal.

Expected Outcomes

The expected outcomes are as follows:

- Project report containing the summary of the literature search
- Synthetic data generation algorithm together with the Python code that produces the new dataset
- Final presentation summarising the results
- Potentially a research artefact (article or conference submission) in case results satisfy the required quality of both parties. This project is highly relevant to the scope of the 19th International Conference on Artificial Intelligence and Law.

Supervision and Mentorship

This role will be supervised by, and report to the Head of the Data Science Lab at Clifford Chance.

Person Specification

Our ideal intern has the following attributes:

- Extensive experience in natural language processing (large language models, information extraction, embeddings, synthetic text generation…)
- Fluent in Python
- Knowledge of machine learning modelling techniques and how to fine-tune corresponding models eg. Deep Neural Networks, Transformers, boosted tree ensembles.
- Experience using one or more of the following specialized machine learning libraries eg. Fastai, Keras, Tensorflow, Pytorch, sci-kit learn, huggingface, spacy
- Must demonstrate the capacity of reading, understanding and implementing new techniques
- Highly organised and self-motivated able to coordinate and drive projects forward with limited supervision
- Strong verbal/written communication and data presentation skills

Internship Logistics

Salary: £30,000 p.a pro rata.

Internship length: 6 months, 3-5 days a week for 1 student or 20hrs/week/student in case of 2 applicants

Location: Remote with occasional (~once a month) travel to London to work alongside colleagues or for training

A background check will need to be completed before the successful candidate(s) can be onboarded.