**REGULATORY GENOME DEVELOPMENT LTD**
208 Mill Road, Cambridge, CB1 3NF
regulatorygenome.com
Company number: **12862325**

# Natural Language Processing Engineers

We are looking for natural language processing engineers with expertise in text extraction from PDFs, machine translation, and/or text classification to join our **remote-friendly** technical team of software engineers, machine learning experts and data scientists to help us realise our vision to transform how the world consumes regulatory information.

## Reg-Genome is *transforming how the world consumes regulatory information.*

RegGenome is a commercial spin-out from The University of Cambridge. The University's Judge Business School created the Regulatory Genome Project, part of the Centre for Alternative Finance, in 2021 to change how the world consumes regulatory information to enable innovation by unleashing the power of information locked in a human readable form on websites and in PDFs. Our proposition is to structure and convert the world's regulatory information into a machine-readable format for digital consumption; creating a repository of regulatory content that is dynamic, granular, and interoperable, enabling regulators to increase accessibility and dissemination of regulation and empowering organisations to digitise their compliance and risk management operations with confidence. Our ultimate vision is an active eco-system of users, partners, consultants, law firms, regulators, standard setting bodies, and application & infrastructure providers all convening around the data to solve the worlds regulatory problems.

## You will be a member of the RegGenome engine-room

You will own the text extraction tasks that help transform raw regulatory documents into RegGenome's structured content – and the quality assurance processes that support continued improvement.

You will have a deep understanding of or interest in the challenge of text extraction from PDFs retrieved from websites given the variability of formatting and potential for documents to be scanned. You may be experienced working with and or have previously developed web-based machine learning text mining tools to enable teams to expertly label data at scale and upload complex documents or unstructured data to generate more informed insights and deliver better customer experiences. Or you may be experienced working on machine translation of complex structured documents into English for further text mining.

You will be creating natural language processing systems to analyse unlimited amounts of text-based data in a consistent manner to extract key facts and relationships, or provide summaries.

You will be developing text mining solutions, to point to the relevant location within documents, so searchers do not have the problem of having to spend hours manually extracting the necessary data by reading through individual documents and solutions that allow for effective keyword and phrase searches across a global repository that return the relevant information from documents in a digestible manner.

You will be creating natural language processing solutions that are deployed with domain-specific ontologies. Ontologies that enable the real meaning of the text to be understood, even

when it is expressed in different ways. The NLP techniques you build will extend the power of ontologies, for example by allowing matching of terms with different spellings and by taking context into account. These ontologies will include a range of vocabularies, ontologies and related strategies to identify concepts in their correct context.

The solutions you develop will need to be enterprise grade and text extraction solutions will need to work for many formats; HTML, XML, PDF and convert text into a standard format to enable the further document processing phases to be standardised. Making the designing of processes to identify fields, classify and translate text more straightforward.

You will be required to support the development of natural language processing tools for annotators to remove the need for specialist skills, like programming expertise, command line access, scripting etc. These tools will need an intuitive graphical user interface and a web portal that enables access by non-technical users.

Our text-mining challenge stretches across millions of documents. Your natural language processing solution must therefore be able to run over tens of millions of documents, each of which may be hundreds of pages long, handle vocabularies and ontologies containing millions of terms and run on parallel cloud architectures.

You will engage machine text extraction experts, to seek guidance, benchmark models and develop longer term research initiatives.

Over time solutions you develop will need to be implemented with linguistic processing to identify the meaningful units within text such as sentences, noun and verb groups together with the relationships between them and with semantic tools that identify concepts within the text and pattern recognition to discover and identify categories of information.

## About the Technical Team

We are a team of skilled engineers and industry-leading experts in the fields of information extraction & retrieval and natural language processing. We work closely with financial regulation domain specialists who identify and annotate training data.

Culture is important to us. We value clear communication through great documentation, kind interactions, and honest feedback. We are creating a flexible and inclusive environment in which engineers can grow. We have regular cross-team knowledge sharing sessions and encourage time for self-directed learning.

We have interesting technical challenges, lots of potential long-term research projects to get involved in, that we are engaging with the University of Cambridge around.

We are a self-motivated team which deals with a lot ambiguity that comes with the territory of Natural Language Processing. We run lots of experiments to solve problems, we learn fast to succeed fast.

We aim to push models into production and automate the drive to improve accuracy through feedback.

## What must you know, or be able to do?

- Proven experience with developing, training and fine-tuning machine learning text processing models on large datasets that can operate in an enterprise grade environment
- Advanced degree in Computer Science, NLP, Cognitive Science, Human Computer-Interaction, Language Technology, Computational, Linguistics, or a closely related field.
- Strong computer science fundamentals and 3+ years of software development experience
- Experience working on text and document structure extraction from common formats, information retrieval, text classification, and/or machine translation
- Strong oral and written communication skills

## Apply if you:

- Are interested in delivering enterprise grade NLP applications
- Are familiar with machine learning toolkits for Python: TensorFlow, Keras, PyTorch, Scikit-learn
- Have experience with text, dialog, and multimodal processing tools/frameworks (such as NLTK, Praat, OpenSMILE and Kaldi).
- Enjoy working with others in the technology, commercial and across the University in solving problems iteratively.
- Are highly motivated to deliver, and able to work in a busy delivery focused environment
- Are excited about developing your skills and experience in the context of a growing company and want to be one of the first 20 employees and help shape a post seed start-up.

## What we offer

- Salary between £80,000 and £95,000
- 25 days' holiday in addition to UK Bank Holidays.
- Share options.
- A flexible remote-working environment and ample opportunity to grow with the company as we scale.
- Ample hardware budget.
- £500 annual learning and development budget to use on subscription services or conference attendance, along with 5 days a year of personal development time.

To apply, please email careers@reg-genome.com with your covering letter and CV.

We are only accepting applications from candidates able to work in the UK at this time.