Data Science Internship 2023

ID Ward

30th November 2022

About the Company

ID Ward is a data privacy startup based in London and Barcelona. We are creating a more private and secure web by using machine learning for data anonymisation. Digital marketing requires companies to collect data about their users and combine it with other data to identify the individual, understand their preference and deliver targeted advertising. But with the development of privacy regulations around the world, this trade in personal data is no longer lawful. ID Ward allows marketers to find consumers in a privacy-friendly way, by storing all personal data on the device and anonymising it with machine learning.

About the internship

We are looking for ambitious, confident and curious data scientists to work with us. You will be part of a young team of developers pushing the boundaries of innovation in data science. At the moment, we have one internship project available. We expect the internship to start in March 2023, and to last 4-6 months. The internship project can also be used in your dissertation, and we will support you taking some time off at the end of the project to write up the results. We offer a reimbursement of €800 per month, and we will expect you to work full time (except for time allowed spent writing the dissertation).

This internship can be fully remote, with the possibility of joining us at our Barcelona office. Choosing to work entirely remotely is also fine and will not affect your application.

About the Project - Multilingual Website Sentiment Analysis

You will be leading the development of a machine learning project to predict the sentiment of webpages for enhanced contextual understanding of web content. The project will have several challenges:

- Experiment and compare different state of the art models (GloVe, BERT, FastText) to get the best trade-off between performance and efficiency
- The model will need to be able to accurately analyse the sentiment of a limited amount of text

- It will need to label text using a tiered taxonomy: tier 1 will comprise standard sentiment analysis labels (positive, neutral, negative) while tier 2 will seek to understand the sentiment in more depth using a wider range of labels (e.g. funny, factual, angry, happy etc.) that we will define jointly
- We will initially focus on building the model for the English language, and then expand it to become a language agnostic
- You will learn how to build an end-to-end automated pipeline for the whole process using different MLOps tools (Apache Beam, Apache Airflow, MLflow).