mondaq
Connecting knowledge & people

About Mondaq

Mondaq is a market leading information, technology, and data analytics business for professional services firms. Its model allows client firms to showcase and market their expertise by distributing their content to a global audience on Mondaq.com, through leading business intelligence services (Bloomberg, Lexis etc.) and many specialist websites. On the other side are consumers of the content in the form of business users seeking expert advice.

Mondaq.com is the fifth-most visited website for legal information globally, and the leading marketing site for law firms (over 14m visitors, 80m+ pageviews a year).

For inquiries regarding any of the projects mentioned below, please get in touch with Rūta Petraitytė, Sr Data Scientist at Mondaq (ruta.petra@mondaq.com).

Summer Dissertation Projects

Below are 3 projects that could be carried out by MSc students as part of their dissertation. For each project we explain its purpose, data availability and requirements.

Project 1

<table>
<tr><th colspan="2">Title</th></tr>
<tr><td colspan="2">Named Entity Recognition: Entity Identification in Mondaq Articles</td></tr>
<tr><th colspan="2">Description</th></tr>
<tr><td colspan="2">Build a Named Entity Recognition (NER) model, that would be able to extract entities like *company names*, *locations*, and *dates* from Mondaq articles.
Please note that part of the project will require to spend some time creating the training data set (i.e. annotating entities in Mondaq articles using Brat). This is a great opportunity for a student to be involved in a project that requires to not just build a model, but also build the training data set, which are not always readily available.</td></tr>
<tr><th>Data</th><th>Requirements</th></tr>
<tr><td>• 435,000 legal articles (in English)
• A sample of annotated legal articles (some annotation work is required here)
Sample size – flexible (to be discussed at the start of the project).</td><td>- Good grasp of Natural Language Processing (NLP) techniques
- Experience in building and evaluating ML/DL models using textual features as input
- Good grasp of Python or R (in context of building ML models and processing textual data)
- Good attention to detail when annotating data</td></tr>
</table>

Project 2

<table>
<tr><th colspan="2">Title</th></tr>
<tr><td colspan="2">Predicting User Engagement based only on Article Titles</td></tr>
<tr><th colspan="2">Description</th></tr>
<tr><td colspan="2">When searching for content on Google or directly on Mondaq.com website – the first thing that users see are the article titles. Only after inspecting the articles title user makes the decision to click on it (or not), thus there is an assumption that the way article titles are formulated highly contribute to the number of clicks. The task here is to investigate this assumption and try to build a regression (or classification) model that would try to predict levels of user engagement based only on the articles title.

Note: while a strong predictive model is of interest, here it is more important to understand *why* some type of titles tend to get more clicks than others, thus some knowledge in textual feature engineering and linguistics is required.</td></tr>
<tr><th>Data</th><th>Requirements</th></tr>
<tr><td>• 435,000 legal article titles (in English)
• Total clicks in the first 7 days since the article was published (clicks from Google)

Sample size – flexible (to be discussed at the start of the project).</td><td>- Good grasp of Natural Language Processing (NLP) techniques
- Experience in building and evaluating ML/DL models using textual features as input
- Good grasp of Python or R (in context of building ML models and processing textual data)</td></tr>
</table>

Project 3

<table>
<tr><th colspan="2">Title</th></tr>
<tr><td colspan="2">Legal Article Classification</td></tr>
<tr><th colspan="2">Description</th></tr>
<tr><td colspan="2">Build a classification model, that would learn to predict topic(-s) for a given Mondaq article.

Due to the available 2-level taxonomy, the project could be approached in two ways:
• Predict a single topic for each article
• Predict multiple topics for each article (preferred approach)</td></tr>
<tr><th>Data</th><th>Requirements</th></tr>
<tr><td>• 435,000 legal articles (in English)
• 2-level Mondaq in-house taxonomy (33 parent and 276 children topics)

Sample size – flexible (to be discussed at the start of the project).</td><td>- Good grasp of Natural Language Processing (NLP) techniques
- Experience in building and evaluating ML/DL models using textual features as input
- Good grasp of Python or R (in context of building ML models and processing textual data)</td></tr>
</table>