Job title	Data Scientist/Senior Data Scientist
Phase	Data Science Challenge
Submission	Via Github, Gitlab, or BitBucket

CHALLENGE

Below are two mini challenges that cover Natural Language Programming, and Computer Vision. As part of the next stage of the evaluation, attempt to solve **ANY ONE** of the challenges and share your code with the HMLR assessment team.

We are testing the following skills (from the original job description) through these challenges.

Essential Technical

- [NLP]: advanced NLP techniques, including transformers and generative models, understanding biases in NLP, proficiency in data augmentation, experience with large language models and their fine-tuning, knowledge of state-of-the-art methods, and familiarity with alternative architectures.
- [Vision] deep Neural Nets (CNNs, Vision Transformers, GANs), image processing, object detection, segmentation, 3D Vision, OCR, Geolocation and multimodal AIs.

Note:

- **You are required to only attempt one challenge**
- These tasks are challenging, and we encourage submission of solutions even if it only partially meets the requirements.
- Use any ML framework of your choice, but the code must be in Python.
- Please ensure complete documentation (including model performance statistics) and information to run your code is included
- Version control the code and commit it to an online git service (e.g. github, gitlab, bitbucket)
- Please ensure that your repository is kept private and only add the following accounts as a collaborator/developer: julian.ludlow@landregistry.gov.uk, ibad.kureshi@landregistry.gov.uk and sonia.williams@landregistry.gov.
- On successful submission of the challenge, you will be invited to an interview where you will be asked questions about your submission.

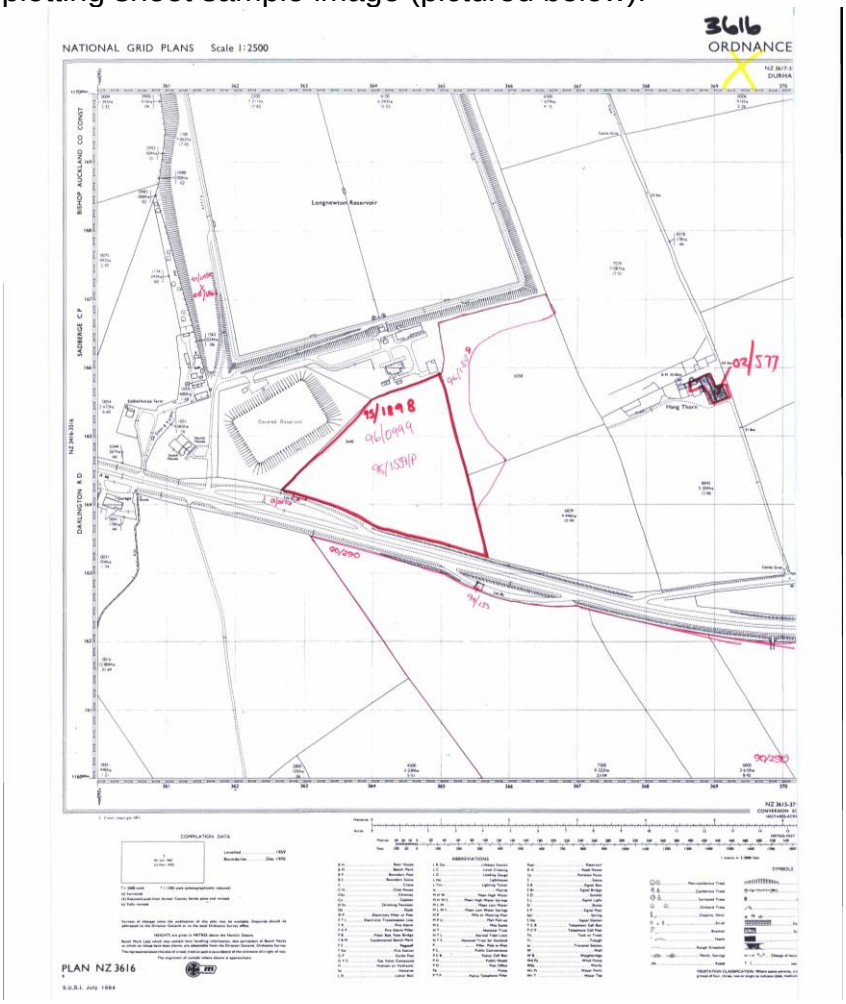
Natural Language Processing

Use the BBC dataset available from UCD (http://mlg.ucd.ie/datasets/bbc.html) and build the following solutions

Essential	Desired
Use the full text dataset and classify each existing category into sub-categories: - Breakdown 'Business' into stock market, company news, mergers and acquisitions etc. - Breakdown 'Entertainment' into cinema, theatre, music, literature, personality etc. - Breakdown 'Sports' into the type of sport: cricket, football, Olympics etc. Please create as many categories as you feel are appropriate.	- Identify documents and extract the named entities for media personalities, clearly identifying their jobs (e.g. Politicians, TV/Film Personalities, Musicians) - Extract summaries of anything that took place or is/was scheduled to take place in April.

Computer Vision

Using the attached plotting sheet sample image (pictured below):



Essential	Desired
Using a Deep/Machine learning method (not just colour segmentation), isolate/segment and extract the following sections of the image accordingly: • Red boundaries of land • Red text giving reference numbers	Convert the segmented sections into a geospatial data format (e.g. geopackage/shapefile), adding the reference numbers as meta-information and geolocating the resultant polygons.

The image attached to the email these instructions came with is the highest available resolution.