UNIVERSITY OF LIVERPOOL

RESEARCH ASSISTANTSHIP

Webscraping university training with GenAI, for engage@liverpool, Methods North West and the North West Social Science Doctoral Training Partnership

Table of Contents

Project Outline

This proposed project places an NWSSDTP PGR with Prof Michael Mair, Director of engage@liverpool, for Methods North West and the North West Social Science Doctoral Training Partnership in order to undertake data science work on:

(a) how Generative AI (GenAI) tools recover information on advanced research training from university websites through webscraping;

(b) whether GenAI webscraping can be utilised by universities themselves to gather together, auto-curate and auto-publicise relevant activities (via, for instance, RSS feeds), and;

(c) offer conclusions on concrete ways in which universities can make their training offers more visible and accessible by optimising their webpages to be read by GenAI, including via (b).

Background and rationale

Internet searches are changing. While machine learning algorithms have long been used to access, parse and rank results to produce the returns of web searches, the introduction of 'AI mode' web searches, via LLMs or as options within search engines, has had an impact on how different users engage with the web. Rather than work through the list of web pages returned by a standard search, users now receive generated summaries, often truncated to the most relevant returns, which parse the results for them. That change has consequences. Existing websites are organised around fairly well-established principles of Search Engine Optimisation (SEO), i.e., making pages available to be read by a search engine. How GenAI technologies read and return information from webpages is much less well understood as is, by corollary,

how 'GenAI search optimisation' should be accommodated by organisations like universities.

With more and more entities seeking to scrape university websites from the outside for various purposes using GenAI, understanding what those engagements are actually returning is becoming increasingly important. Not only is it important for understanding how Gen AI tools parse, summarise and thus shape perceptions of what is happening within universities and what they do ((a) above), it can also be used for universities to more effectively organise content to be read by GenAI and through that make the training they offer more visible and thus accessible ((b) and (c) above). As a pilot for other Methods North West and NWSSDTP institutions, this project will focus on what GenAI webscraping recovers in terms of advanced training at the University of Liverpool (UoL).

Placement Timeline and Key Objectives

Phase 1: first two weeks (or four weeks PT)

- initial discussions with project team and working through the objectives, approach and milestones.

Phase 2: first month (or first two months PT)

- begin trials with GenAI retrieval and webscraping the UoL website to recover methods and methodology related activity under different parameters.

Phase 3: second month (or months 2-4 PT)

- assess results of the automated returns through discussions with relevant staff to determine how complete and accurate the results are.
- explore auto-curation and auto-publication in light of those discussions.

Phase 4: third month (or month 5-6 PT)

- produce a project report and write a protocol with codebase based on the work to-date to enable other teams to extend the project to other universities across the region.

Required and Desired Skills

The project will suit a student from any pathway with advanced data science training, including high level coding skills and an understanding of webscraping, who is interested in and has a working knowledge of GenAI, including prompting and vibe coding, and wants to both develop those skills further as well as extend their industry experience in a sector where new AI tools have already had a great deal of impact.

Timeframe

This is a 13-week placement worked either full-time (1.0 FTE) over 13 weeks or part-time (0.5 FTE) over 26 weeks.

Delivery of project

The project would have a hybrid mode of delivery. Much of the project can be completed off-site given the project is focused on the UoL website. Start date to be discussed with project team but with a view to the project beginning after the end of the Easter break, w/c 13th April. However, some time on campus would be expected to facilitate conversations with relevant parties about the project. At the same time, the student would be welcome to embed themselves with local teams if preferred.

Induction process and support you will receive

The project will be supervised by Prof Mair, who will have weekly meetings with the student (if full-time, or every other week if part-time) about progress.

Administratively, the student will be embedded with the PGR group in Sociology, Social Policy and Criminology in the School of Law and Social Justice in the UoL. They will have hotdesking space with that group when on site and will receive the standard induction given to new PGRs in SLSJ covering all of the essential topics listed above: namely, health and safety protocols, any risk assessments or security checks that will need to take place, and any additional resources or points of contact to support their integration and well-being during the placement.

Ethics Approval

None required.

Equality and Diversity commitments

The project team aligns with the NWSSDTP's Equity, Diversity and Inclusion commitments and are willing to make reasonable adjustments for disabled students or students facing barriers to inclusion for other reasons.

How to apply

To apply for this role, please complete the 'Expression of Interest form' available on the NWSSDTP website, using job title mentioned in the NWSSDTP advertisement.

Once completed and discussed with your supervisor, return the completed form to Amy Connolly (ripps@liverpool.ac.uk).