ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Bachelor's or Master's Thesis Proposal

Machine learning assisted crawler for legal compliance analysis of registration forms

Supervisors: Karel Kubíček, Carlos Cotrini
Professor: Prof. David Basin

Issue Date: March 30, 2021

1 Motivation

Have you ever wondered whether the marketing emails you receive are legally allowed? Have you ever been curious about who gave your data to the companies that send you the emails? And have you asked yourself what could be done against it and whether you have potentially a legal claim?

We want to create a crawler that scans websites, detects privacy violations regarding sending marketing emails, and reports those to regulatory services. The ePrivacy Directive of the EU holds that the use of email for direct marketing purposes requires – as a general rule – the prior consent of subscribers (Art. 13). What constitutes consent is defined in the General Data Protection Regulation (GDPR). While these are European Union laws, they are also very relevant for any Swiss firm dealing with European customers, and Switzerland is currently revising its data protection rules based on the GDPR. According to the GDPR, non-compliant companies may face fines of up to 20 million EUR or 4% of worldwide turnover and damage claims from persons violated in their rights.

In this thesis, your task is to extend an existing web crawler. The current crawler can register for up to 4% of websites, but we believe it should be feasible to register for at least 10%. For that, we need to use machine learning methods for mimicking human skills to interact with various types of registration forms.

2 Methodology

For analyzing compliance of received emails, we need to register for websites. To scale such analysis, we implemented a crawler that can currently register for up to 4% of websites. The crawler's current version relies solely on keyword detection for both navigation and interaction with the form. This requires a complicated system for keyword ordering, which is with an increasing number of keywords inevitably prone to mistakes. Given the large amount of data we operate, we can train a machine learning model to classify the page type (e.g., distinguishing the login and registration form) and the form content.

As the whole crawler already exists, the student will extend it with machine learning and compare the updated crawler's results with the former version. For this, the crawler comes with a framework supporting distributed execution and the result evaluation.

3 Tasks

We scale the complexity of the thesis by different requirements on what models the student should implement. A Bachelor's student is expected to fulfill at least the following tasks.

1. Page classification: the distinction between login page, registration page, terms and conditions before registration, and successful/unsuccessful registration response.
2. HTML subtree classification: many pages contain both login and registration forms next to each other, so these have to be classified as well.
3. For each of the previous tasks, we expect the student to:
   a) collect training dataset,
   b) train a classifier (supervisors can bootstrap this effort by providing their feature selection and training algorithm),
   c) implement the usage of model in the crawler, and
   d) evaluate the progress by rerunning a large crawl.

A Master's student is also expected to work on the following tasks.

1. Subclassify the registration form fields.
2. Support multilingual crawling using automated translation.
3. Interact with multi-step registration forms.
4. Detect wrongly-inserted fields of a failed registration attempt.
5. Classify any pop-ups as Cookie consent that blocks interaction with the registration form.

Note that several computer scientists and lawyers actively develop this project, so you should be a proactive team player to join this work.

4 Requirements

The student interested in this thesis should satisfy the following requirements.

- Passed Introduction to ML (for BSc students) and Advanced ML (for MSc students) courses.
- Experience with common NLP processing, supervised learning models, and their evaluation.
- Programming fluency in Python, JavaScript is an advantage.

5 Deliverables

- At the end of the second week, the project's schedule must be shared with the supervisor.
- At the end of the project, a presentation of 20 minutes must be given during an InfSec group seminar. It should give an overview as well as the most important details of the work.
- Software, configuration scripts, and any supplementary material must be delivered to the supervisors.
- A final report consisting of an introduction, a discussion on the related work, an overview of the procedure and implementation, data from experiments, and the evaluation of common violations.

References

[1] Steven Englehardt, Jeffrey Han, and Arvind Narayanan. I never signed up for this! Privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies*, 2018(1):109–126, 2018.

[2] Hugo Jonker, Stefan Karsch, Benjamin Krumnow, and Marc Sleegers. Shepherd: A generic approach to automating website login. 2020.