Coles Captcha Issue

Introduction

This document outlines the challenges and solutions encountered while dealing with high-level CAPTCHAs during a data collection project from the Coles supermarket website. These mechanisms are designed to differentiate human users from bots, posing significant hurdles for automated systems.

The aim of this document is to provide a comprehensive overview of the methodologies adopted, the research conducted, the solutions implemented, and the challenges faced. This will serve as a resource for similar future endeavors, particularly under constraints of limited funding.

CAPTCHA systems are crucial in maintaining the security and integrity of online services. They prevent automated software from performing tasks that could potentially lead to service disruption or unauthorized data access.

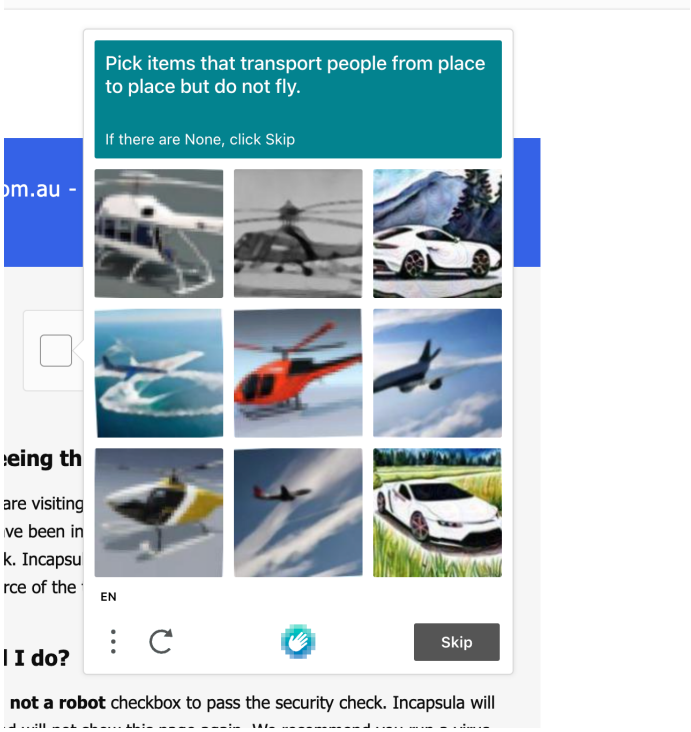
Background

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a type of challenge-response test used in computing to determine whether or not the user is human. These are typically designed to be easy for humans but difficult for computers.

In the context of web scraping, CAPTCHAs serve as a primary defense against the automation of data extraction processes. They ensure that users are physically present and are not using automated tools to access or harvest data, which can sometimes violate terms of service or lead to data privacy concerns.

Problem Statement

During the project, we encountered several forms of high-level CAPTCHAs including image recognition tasks, distorted text, and puzzles. These CAPTCHAs were dynamically generated and tailored to increase in complexity upon repeated access attempts, making them particularly challenging to bypass.



Example captcha

The CAPTCHAs significantly slowed or even prohibited the progress of the data collection process, necessitating a deeper exploration into advanced solving techniques and imposing additional resource requirements on an already limited budget.

Research Phase

Our initial research involved gathering information from a variety of sources including scholarly articles, online forums, and existing documentation on CAPTCHA technology. We focused on understanding the different types of CAPTCHA and their specific vulnerabilities to various solving techniques.

We evaluated a range of tools including traditional OCR (Optical Character Recognition) software, AI-based image recognition systems, and third-party CAPTCHA solving services like 2Captcha and Anti-CAPTCHA. Each tool was assessed for accuracy, integration complexity, and cost-effectiveness.

The research phase also included a review of case studies where similar CAPTCHA challenges had been encountered and overcome. Insights from these studies informed our approach and highlighted potential pitfalls.

Technical Exploration

We explored several commercial CAPTCHA-solving services that utilize human labor and machine learning techniques to decode CAPTCHAs. These services provide APIs which were integrated into our scraping scripts to facilitate real-time CAPTCHA resolution.

Considering the project's budget constraints, we also investigated free, open-source machine learning models that could potentially be trained to solve CAPTCHA challenges. However, the adaptability and complexity of the CAPTCHA mechanisms used by Coles required more advanced solutions.

Open-source tools such as Tesseract OCR were initially tested but found inadequate for the complex CAPTCHA forms we encountered, which often required contextual understanding beyond the capabilities of basic OCR technology.

Results

The financial implications of using paid services were substantial, given the high volume of CAPTCHAs encountered. This placed considerable strain on our limited budget, highlighting the need for a more sustainable long-term solution.

The additional steps required to resolve CAPTCHAs introduced delays in the data collection process, impacting overall project timelines and efficiency.