Wavelength Capstone Project 1

Data Extraction & Web Scraping from bailii.org

This first part of the capstone project involved creating a script to extract various pieces of information from 50 planning court cases on www.bailii.org based on a provided excel template

A python library called Beautiful Soup was selected as the main tool for this information extraction exercise. It is designed to scrape information from HTML webpages which are structured as nested groups of tagged elements containing textual and other forms of data. The Beautiful Soup library, together with an html parser, can convert the base html web page into an easily navigated, nested data structure, and it provides a wide range of methods to search, extract and manipulate the information stored within tagged elements

Selenium is another widely used library for web scraping, providing sophisticated methods to navigate through linked web pages. After prototyping with the Beautiful Soup library, Selenium was investigated as a method of navigating through the various query result pages, as only 10 results were shown per page. This issue was subsequently resolved by adapting the initially provided query URL to show 50 results per page, this proved to be simpler than using a second library to navigate the various results pages from www.bailii.org .

Manual inspection of the elements on the query results page showed the case name and case URL fields were stored in an HTML list on the summary query results page. This information was scraped with Beautiful Soup for all 50 cases via the list item tags and stored in a pandas dataframe. It was then possible to extract the case date using a regular expression from the retrieved case name and to then derive the Week field from the case date.

Further extraction required the creation of an individual BeautifulSoup object for each case page from the retrieved Case URL. For most cases the required information was stored inside elements with accessible tags such as citation, panel and casenum, there appeared to be two main formats of HTML and so the meta tags of the web page were used as a condition to assess which tags would be work for a given document. There was one case from the High Court of Ireland that did not match either format and appeared to be an incorrectly returned result given that it didn't match the query criteria provided, due to this inconsistency the additional information for this case was not retrieved.

For the extraction of information on relevant local authorities and key cases additional code was required to strip irrelevant data from the retrieved elements and to include link references in the export. A python Counter collection was used to keep a count of the number of times each case tag appeared.

For the more complex problem of retrieving relevant Acts and Sections the Spacy library and a library based on Spacy called Blackstone was used. Initially regular expressions and Beautiful Soup were applied as with the other extracted fields but with limited results due to the wide variety of formats. Blackstone seemed like an intriguing solution as it fine-tunes Spacy's statistical models using a large corpus of English legal texts and provides Named Entity Recognition (NER) for several legal terms including Acts (INSTRUMENTS) and Sections (PROVISIONS).

An interesting feature of Blackstone is its entity linking model which uses a knowledgebase built from a large corpus of English legal texts. This entity linking model can be used to link PROVISION entities to their relevant INSTRUMENT entities and retrieve links to the relevant legislation from the legislation.gov.uk website. This was tested towards the end of the capstone and the results are included as an extra field in the column linked legislation.