Knowledge Representation Workshop 6

CITS3005

September 14, 2023

The Project: Information Extraction

The unit project is released at `https://teaching.csse.uwa.edu.au/units/CITS3005/` and requires you to build a knowledge graph and ontology represnting a fragment of the information contained in the UWA Handbook.

The particular information you should focus on is the undergraduate units and undergraduate majors.

1. Examine the handbook and discuss the nature of the knowledge it contains. What information is structured, what information is unstructured, what information is implicit?

2. Open the HTML page course for the handbook and look at the way the information is informatted. Is the information consistently formatted. What techniques can we use to extract the information from the handbook.

3. An example web crawler is provide that uses a mixture of beautifulsoup and regular expressions to extract information on all the units in UWA. Examine this code, and compare it to the provided units.json dump. How can it be improved.

4. Consider the problem of extracting the information regarding Majors at UWA. What information is contained; how can we extract it; what should we ignore?

5. Discuss possible schemas for a handbook knowledge graph.