Co-funded by the
Erasmus+ Programme
of the European Union

i. Name of the course

Directed Study in Linguistics (NLP strand)

ii. Level of the course

BA

iii. Workload

4 ECTS

iv. Institution

University of Malta

v. Course instructor(s)

Marc Tanti

vi. Brief course description

This course is aimed at students who wish to deepen their knowledge of some area of linguistics, by carrying out an assigned research or analysis task under supervision. It is on offer to 2nd and 3rd year undergraduate students, as well students in the MA Preparatory Course, which is designed to allow graduates of non-Linguistics-related degrees to undertake an MA in Linguistics.

In the specific iteration run for UPSKILLS, the directed study is in the area of Computational Linguistics. Students will be scraping and parsing recipe webpages in order to automatically extract a CookLang formatted recipe.

With the help of an assigned tutor, the students will:

- Write a Python script that downloads the HTML of a handful of assigned recipe webpages.
- Use BeautifulSoup and regular expressions in order to find the title, ingredients, and procedure of the recipe.
- Parse the ingredients to separate the item, quantity, and unit of each ingredient.
- Determine the different ways that the ingredients are referred to in the recipe procedure. An important distinction of the CookLang recipe format is that

ingredients are highlighted within the recipe procedure rather than given as a list at the top, so some minor natural language processing needs to be applied in order to determine different ways how an ingredient is mentioned, such as olive oil being referred to as 'the oil'.

- Output a text file containing the recipe formatted in CookLang.
- Compare the automatically generated text files with manually written gold text files in order to evaluate the program.
- Write a report about how the program works, how it performed, what was learned, and how the program can be improved.

vii. Research related subject

Natural Language Processing, Information extraction

viii. Tools and data the students work with

Python, HTML, regular expressions, shallow parsers

ix. Topics

The topics to be covered are:

- Natural language processing (without machine learning).
- Information extraction.
- Shallow parsing.
- Text distance measures (for comparing the automatically generated output to the gold output).

x. Learning outcomes

The students will learn:

- How to scrape websites.
- How to identify patterns that can be exploited to make a small program work on different websites.
- How to extract reliable bits of information from text.
- How to convert data between different formats.
- How to evaluate text generation.

xi. Evaluation

Research report evaluation, following the criteria outlined in the UPSKILLS guidelines.

xii. Further information on the course that the instructor considers relevant (assessment, background, reading materials, detailed weekly plan, career paths etc.)

Given that this course comprises an extensive practical component, the following sources will be used to familiarise the students with the task they need to complete:

- https://www.geeksforgeeks.org/python-web-scraping-tutorial/
- https://beautiful-soup-4.readthedocs.io/en/latest/
- https://docs.python.org/3/howto/regex.html
- https://www.nltk.org/api/nltk.chunk.html