DMP title

Project Name Solving Combinatorial and Probabilistic Problems in Natural Language (FWO DMP) - DMP title

Project Identifier G066818N

Principal Investigator / **Researcher** Luc De Raedt and Pietro Totis

Project Data Contact luc.deraedt@kuleuven.be, pietro.totis@kuleuven.be

Description This project wants to develop a fully automated approach to solving exercises about combinatorics and probability that can be found in introductory textbooks on discrete mathematics. The ability to solve such problems is an important cognitive and intellectual skill as it is evaluated as part of academic admission tests such as SAT, GMAT and GRE. The combinatorics and probability questions will be formulated in natural language and the task will be to automatically answer these questions. We shall develop a two-step approach for tackling this task. In the first step, a question formulated in natural language will be analysed and transformed into a high-level model specified in a declarative language. In the second step, the high-level model will be solved solved using the inference mechanisms of for the declarative modeling language. The language and its solvers will be based on principles of probabilistic programming, is an increasingly popular programming paradigm. While the immediate goal is to solve textbook exercises, the long term goal is to contribute to the automation of probabilistic and combinatorics problem solving and to enable the modeling and programming for such problems in natural language, two goals that are highly relevant to cognitive computing and artificial intelligence.

Institution KU Leuven

1. General Information

Name applicant

Pietro Totis

FWO Project Number & Title

Project number: G066818N
Project title: Solving Combinatorial and Probabilistic Problems in Natural Language

Affiliation

- KU Leuven

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Reuse existing data

Describe the origin, type and format of the data (per dataset) and its (estimated) volume, ideally per objective or WP of the project. You might consider using the table in the guidance.

Data is composed of (1) 3040 problems collected from different textbooks about probability and combinatorics, and (2) 2175 encodings in the solver format for the problems where an encoding was possible. The examples are saved in formatted (.json) files for a total of 45 MB, while the encodings have an estimated volume of 1 MB. The example files contain the formulation of the problem in natural language, the tokenization of the text, the solution of the problem and the source (url address) from which the problem was collected. The encodings contain the formulation of the problem in the solver language. The NLP experiments are based on a dataset derived from the .json files by extracting from the formatted data the field containing the problem formulation in natural language. The encodings in the solver language are stored in plain text files. The source code is also divided between the experiments on the NLP dataset and the experiments on the solver encodings.

3. Legal & ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.

- No

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- No

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- Yes

Many examples of the dataset are protected by the copyright of the books from which they were collected.

4. Documentation & metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

A README file inside the data folder describes how the data was collected and the scripts to replicate the experiments on the dataset. A pdf document inside the

encodings folder describes the language of the solver used in the encodings of the problems. Each file is identified by a codename and each encoding contains the codename of the corresponding problem as a commented line of code.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

5. Data storage & back up during the FWO project

Where will the data be stored?

The master copy of the data and code is kept on the departmental repository server SCM.

How is back up of the data provided?

The departmental repository server has automatic backup procedures

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available then explain how this will be taken care of.

- Yes

Given the small size of the dataset capacity is not a concern for the project.

What are the expected costs for data storage and back up during the project? How will these costs be covered?

None

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The departmental repository server provides a secure access to the data.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

The data can be retained on the departmental repository server and the internal backup sytem for the publications regarding the dataset.

Where will the data be archived (= stored for the longer term)?

The data will be stored on the departmental repository server (with automatic back-up procedures) for at least 5 years, conform the KU Leuven RDM policy.

What are the expected costs for data preservation during the retention period of

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

None.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- Yes. Specify:

The data is protected by the copyright rights of the publisher of the books from which the problems were collected.

Which data will be made available after the end of the project?

Due to the copyright protecting the redistribution of the problems the dataset cannot be made pubblicly available.

The code can be made pubblicly available.

Where/how will the data be made available for reuse?

- In an Open Access repository
- Upon request by mail

The data will be only shared for research purposes upon request by mail.

The code will be made available in an Open Access repository.

When will the data be made available?

- Upon publication of the research results

The dataset will not be made available, except questions from publishers that authorize redistribution.

The code will be released upon publication of the research results by uploading it to an Open Access repository. The code is licensed under the Apache License, Version 2.0.

Who will be able to access the data and under what conditions?

The dataset can be made accessible individually for reproducibility purposes, under the condition of not violating the copyright by publicly redistributing the dataset.

What are the expected costs for data sharing? How will the costs be covered?

None

8. Responsibilities

Who will be responsible for data documentation & metadata?

Pietro Totis will be responsible for data documentation & metadata

Who will be responsible for data storage & back up during the project?

Pietro Totis will be responsible for data storage & back up during the project

Who will be responsible for ensuring data preservation and reuse ?

Pietro Totis will be responsible for ensuring data preservation and reuse

Who bears the end responsibility for updating & implementing this DMP?

Pietro Totis bears the end responsibility of updating & implementing this DMP.