DATA SCIENCE 344: Assessment Further (AF) 2024

You are required to form groups of three students for presentations scheduled near the end of the fourth term. A SUNLearn link will be provided for the purpose of forming groups.

Dates for the final presentations will be announced.

Your group needs to investigate, and then present practical applications of the material covered in the module. A key requirement is that both the two main topics of the module (NLP & Deep Learning) need to be represented. It is not necessary or even encouraged for groups to present on every subsection of the two topics. Although not a requirement, further practical exploration into cloud computing or any other aspect of the course is strongly encouraged. A 4-page executive summary needs to be handed in the week preceding the presentations. This summary should explain all the steps and results from your application.

The exercises described in the various sections of the module can serve as points of departure, but this project should explore beyond the limitations of these exercises.

As an example for NLP, you could identify some websites where you can extract textual (and other) information which you can then shape and wrangle into a form that makes it possible to do exploratory descriptive work (eg. plotting most used words or phrases). You could then identify relevant topics from the data and even develop a statistical learning application based on the topics identified.

Regarding the Deep Learning section, you should apply methods related to NLP in deep learning. You can use the same data used for the first part or use a different dataset. Possible topics include:

- Text classification, sentiment analysis, topic classification, spam detection
- Text generation
- Question answering
- Text summarisation
- Machine translation
- Speech to text / text to speech

These topics are merely examples. Students are urged to formalize their own applications.

Regarding the deep learning section, one of two broad approaches can be followed. You can either apply your own custom model on a simpler problem or you can use a pre-trained model for more complex tasks. If the latter approach is followed, a technical summary of the pre-trained model must be given too. For example, if you use a pre-trained transformer (like BERT or GPT), it is required to give an overview of the architecture, where the weights were obtained, and how the model was fine-tuned. Those interested in this approach should read about the “transformers” Python package from HuggingFace.

The code to reproduce the results must be submitted with the summary. You can either upload the code during the submission period on SUNLearn or provide a link where it can be downloaded (e.g. on GitHub).

MARKING GRID

	Criterion	Mark
A	The two main sections of the module are both fully covered: NLP & Deep Learning.	/20
B	The presentation demonstrates a good grasp and adequate depth of the material covered in the module.	/20
C	Presentation as a team effort.	/15
D	Quality of the presentation: Structure, flow and clarity.	/15
E	Quality of the executive summary document. A brief (max 4 pages) document clearly describing the problems and headline findings.	/20
F	Presentation of work that extends beyond what was offered in class.	/10
	TOTAL	**/100**