Required Assignment (max. 75%)

Imagine you are a Data Scientist who is tasked with helping scientists understand the ever-growing body of research in their field. Your clients' problem is that many new papers are published every day and a simple keyword search is unable to pinpoint all the relevant information inside these papers. In trying to understand the literature in their specific field, the scientists often ask questions like:

- Where can I find information about a particular scientific question, problem or task?
- What kind of methods, techniques and resources have been applied to this task?
- How do these methods relate to other methods I've heard about?

This assignment will guide you through the steps of prototyping a solution for your clients. You will use text analytics to help scientists discover different types of entities, how they relate to each other, and analyse the information you extract. While the clients in our scenario are scientists, the same kinds of problems apply in many other knowledge-intensive domains.

For each task:

In your report, write a section that motivates your chosen solution by briefly describing the method and its advantages/disadvantages. You should design methods that incorporate ideas from the lectures and, where suitable, discuss concepts covered in the lectures (e.g., the limitations of bag of words, the value of syntactic information, learning vs. knowledge, relational, compositional and contextual views of meaning). You may wish to include a system diagram showing your NLP pipeline or deep neural network. If you experiment with different features or hyperparameter settings on the development set, you might want to include a plot or table showing the results.

Dataset

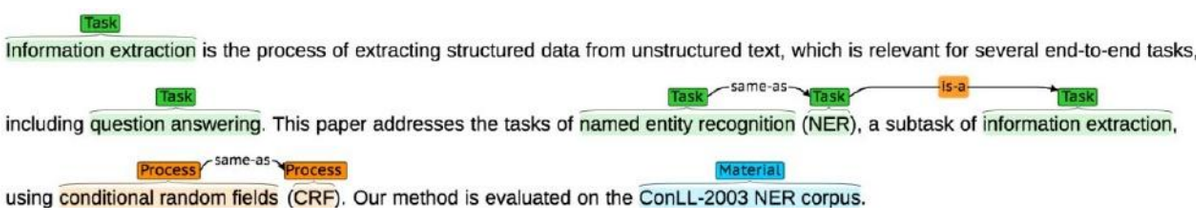
We will be working with the ScienceIE dataset, which is available for download here, which consists of abstracts of scientific articles. The dataset is already split into training, development and testing sets. The dataset directory includes the following:

- A Python module, *scienceie_loader.py,* containing a data loader, which returns a list of tokens, sequence labels and relations for each split of the dataset.
- The *scienceie_loader.py* module also contains a function that can return the raw text, with the sequence labels and relations defined by character offsets. This is needed if your proposed solution requires untokenised text or uses its own tokenizer (e.g., if using the HuggingFace Transformers library).
- *data_loader_demo.ipynb,* a notebook with an example of how to call the scieneie_loader functions.
- The original ScienceIE release of the dataset, including readme files.
- The original annotation guidelines that clearly define what each type of entity or relation means.

The data was originally released for a competition/shared task and is described in full in this paper:

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017, August). SemEval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations from Scientific Publications. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 546-555). Link to publication..

You can find out more about the dataset by reviewing the related literature, e.g., by searching for the topic on the ACL Anthology or searching for the above paper on Google Scholar. An example of the kind of annotations in the ScienceIE dataset is shown below:From https://scienceie.github.io/example.html



Task 1: Extracting Entities from Scientific Abstracts (max. 35%)

Note that this task corresponds to subtasks A and B in the original ScienceIE shared task.

The first step is to extract the relevant types of entities from the abstracts. The entity types are:

- **Task**: specific research tasks (e.g., 'dependency parsing') and broader research areas (e.g., 'machine learning').
- **Process**: methods/techniques/algorithms, physical equipment and software tools.
- **Material**: physical materials, datasets/corpora and other resources used to solve the problems in a scientific paper.

1.1. Implement and train **two methods** for extracting these three entity types from scientific abstracts. You can refer to the labs, lecture materials and textbook for this course to identify a suitable method. You can use the training set to train a model, and the development set if you need to tune hyperparameters. Choose suitable pre-processing steps for your selected methods.

The two methods you compare should employ different kinds of model (e.g., an HMM and a neural network). If the methods both use machine learning, they should also use **two different sets of feature** (e.g., POS tags + unigrams vs. word embeddings). It is not necessary to test many combinations of feature sets with models.

Explain your chosen methods and features in your report and hypothesise how the results of the two methods will differ.

1.2. Evaluate your method on the test set and include the results as a table or plot. Interpret and discuss your results. Imagine that you are building a real system for your clients and carry out this evaluation to help you and your clients decide which approach is most suitable for entity extraction. In your discussion, include the following points:

- Which performance metrics did you choose and what limitations do they have?
- How does the performance of your two methods compare?
- What is the effect of the different features?
- Show a few representative examples of errors from each method. Does either method make any common types of errors?
- How could you improve the methods?

Task 2: Extracting Semantic Relations(max. 25%)

Note that this task corresponds to subtask C in the original ScienceIE shared task.

Now that we have a way to identify entities of interest, we would like to understand how they relate to one another. The ScienceIE dataset defines two types of semantic relation:

- **Synonym-of:** the two entity phrases refer to the same thing. If this relation holds, <entity A> is the same as <entity B>.
- **Hyponym-of:** the first entity is a more specific word then the second. The second entity is a category or more general term to which the first belongs. If this relation holds, <entity A> is an <entity B>, but not vice versa.

2.1. Implement and train **one method** for predicting these two types of semantic relation between pairs of entities that are annotated in the dataset. It is up to you to choose a suitable method and features based on the labs and lectures. In the report, explain the method and how you chose it.

For the entity annotations, you may use either your best entity extraction method from task 1 or the gold standard entities in the test set. In your report, explain which source of entity annotations you use and what effect you think that may have on the results.

2.2. Evaluate your results on the test set and include the results as a table or plot. Discuss and interpret your results and include the following points:

- Show a confusion matrix of errors.

- Show a few representative examples of common errors.
- Can you identify any particular types of mistake the method makes?
- How could you improve your method's performance?

Task 3: Exploring the Data (15%)

Now that you have developed methods for entity and relation extraction, the task is to use this information to explore the dataset. The plots you will create in this section should demonstrate how your text analytics system could help your imaginary clients to understand how entities relate to one another.

Choose **one** example entity from your predictions on the test set to use as queries. Then, for each query entity:

3.1. Obtain the list of relations that include the query entity. Some relations may occur more than once: count the number of occurrences of each unique relation. Print the top ten relations alongside its number of occurrences **OR** plot a graph of the top 10 relations using Plotly or a similar package. Include the list or graph in your report, then write a few sentences stating what the relations tell you about the query entity.

3.2. Obtain a list of entities that occur the same abstract as the query entity (co-occur). Count the number of times these entities co-occur with the query entity. Print the top ten entities with highest co-occurrence, including their types **OR** visualise the entities in a word cloud using a package like Wordle. Write a few sentences describing what the list or word cloud tells you about the query entity.

Open Assignment (max. 25%)

Extend your work on the ScienceIE dataset by developing a second, improved method for task 2, extracting semantic relations. Write a brief report that addresses the same points as task 2 but additionally:

- Explain the technical challenge that your second method tries to address, i.e., your hypothesis for what will improve performance.
- Include a critical discussion of the methods and your experiments.

Credit will be given for more advanced methods or extensive experimentation, but any complexity should be justified in your report. Simple but informative experiments with a concise write-up can also receive high marks on this assignment.

Implementation

We recommend using Python 3 with the libraries used in the labs, namely NLTK, scikit-learn, Matplotlib, Gensim and Pytorch. You may use other libraries, but we cannot provide support for them. You may use either Jupyter notebooks or standard Python files. We encourage you to refer to the lab notebooks for ideas on how to implement your solutions.

Assessment Criteria

Your coursework will be evaluated based on your submitted report containing the presentation of methods, results and discussions for each task. To gain high marks your report will need to demonstrate a thorough understanding of the tasks and the methods used, backed up by a clear explanation (including figures) of your results and error analysis. The exact structure of the report and what is included in it is your decision and you should aim to write it in a professional and objective manner. The report will be assessed based on how well it addresses each of the tasks in the required and open assignments, with the percentage of marks available for each task shown above. Marks will be awarded for appropriately including concepts and techniques from the lectures.

Avoiding Academic Offences

Academic offences include submission of work that is not your own, falsification of data/evidence or the use of materials without appropriate referencing. Note that sharing your report with others is also not allowed. These offences are all taken very seriously by the University. Suspected offences will be dealt with in accordance with the University's policies and procedures. If an academic offence is suspected in your work, you will be asked to attend an interview with senior members of the school, where you will be given the

opportunity to defend your work. The plagiarism panel can apply a range of penalties, depending on the severity of the offence. These include a requirement to resubmit work, capping of grades and the award of no mark for an element of assessment.

Marking Guidelines

Outstanding project (80+)

- Completion of the required assignment with no major flaws or omissions.
- Excellent work on the open assignment.
- Proficiency in all aspects, including advanced methods.
- Impressive and novel outcome with strong research elements – close to publication quality.
- An original synthesis using both ideas from the unit as well as from the literature.
- Excellent presentation of work, with clear structure and descriptions, which could be held up as an example of strong technical writing.
- Excellent use of plots to support the interpretation of results.
- Evidence of outstanding unique and individual contributions.

Distinction (70+)

- Completion of the required assignment with no major flaws or omissions.
- Good attempt at the open assignment.
- Excellent outcome showing a good grasp of the complete development and evaluation process.
- Evidence of deep understanding and correct application of a wide range of techniques.
- Study, originality, and synthesis clearly go beyond the minimum requirements set out in the coursework description.
- Very good presentation of work, with a clear structure and descriptions of the methods and results.
- Very good use of plots to support the interpretation of results.
- Evidence of valuable contributions or insights into the methods tested.

Merit project (60+)

- Complete solutions for the required assignment, albeit with some minor flaws or omissions.
- Very good outcome that shows good awareness of how to develop and evaluate a solution.
- Evidence of correct use and strong understanding of a range of techniques.
- Study, comprehension, and synthesis fully meet or exceed the requirements set out in the coursework description.
- Good presentation of work, with a mostly clear structure and mostly complete descriptions of the methods and results.
- The use of plots supports the interpretation of results.
- Evidence of critical analysis and judgement of the methods tested.

Pass (50+)

- A good attempt was made for most tasks in the required assignment, but not all parts completed satisfactorily.
- Good outcome showing some awareness of how to design, implement, and evaluate a solution.
- Evidence of appropriate use and understanding of standard techniques.
- Some grasp of issues and concepts underlying the techniques.
- Adequate presentation of work, including a description of the methods and results, but unclear or incomplete in places.
- Some use of plots to support the interpretations but with some notable shortcomings.
- Minimal critical analysis and judgement of the methods tested.