Project 2
Year prediction

The goal of the project is to provide a model able to predict for a given scientific article on NLP the year of publication. We use the same data as the first project but in a supervised way.
You are in charge to split the data in train/test/validation sets with 80/10/10 respectively for example. You can evaluate your model with a MSE error, since making a mistake of ten years or one year it's not as important.

The goal is not to reach the highest score. It's easy, you can train on all the datas and generate a model that is just a memory and achieve a perfect model, which is not very interesting.

Two mains expectations of the project :

1) propose a model to achieve this goal and explain all your modifications step by step, in order to improve your result. You can use different models, but you need at least show the evolution of one model.

2) You need to analyse the results : Which years are more difficults ? Why ? Which articles are more difficult, or which key-words are considered more difficults ?
Answering to this questions is an other way to synthetise the corpus and understand the evolution of Natural Language Processing.

You can use all tools and models seen during sessions.

You can focus on abstracts or titles (less computational effort) or select a specific subset.
Please : be aware to not use meta-data that help to predict the year ! It will improve dramatically your precision but destroy your analyse. For example, the place of the conference or edition number are a strong synonym of the year (the year of articles cited/referenced also)!

The number of articles per year are not equals. Be aware to force balanced year distribution between train/test/validation sets. For example, if year 2001 is present only in train articles and year 2010 only in evaluation articles, it's a loss of representativity in evaluation and train respectively.

Your report must explain what technics/approachs you use, how you use them and the results obtained. If an approach don't work as planned you can show and explain (It will be very appreciate).
You can work in pairs of students. Your report must contain the names of students involved.
Your report must explain the logic of your approachs and results.
You can write in English or French.
Your report must contain your link to your Colab Notebook.
Your report must be deposited on DVO before Monday **11 january 2021**.

Dataset (44Mo) : https://www.ortolang.fr/market/corpora/corpus-taln (download link at page bottom)

Good Luck