You are applying for a Data Scientist position at Hyperlex and have been selected to pass the technical exercise! Congrats!

NLP applied to company statutes

You'll have access to a bunch of French company statutes and you'll have to process them with several NLP/ML techniques.

1. Report

You are free to hand in your results in any format (report, notebook, ...). The same applies for your code, pick the language you're the most comfortable with. In the report, please elaborate how you investigate and solve the problems.
[1]

If you don't have time to explore some techniques (and that's totally fine!), do not hesitate to explain what you would have tried and/or link to some paper/pointer that explains the method. We are also very interested in knowing all the problems you encountered!

2. Data

In the `dataset` directory, you'll find 647 subdirectories containing an original statute of a company in PDF format along with a json file with the following structure:

```
{
    "filename": "Statuts 10.PDF",
    "clauses": [
        "Article 1. This is the first clause",
        "Article 2. This is the second clause",
        ...
    ]
}
```

[1] Ce dataset est une propriété d'Hyperlex, il est donc strictement interdit de l'utiliser/exploiter en dehors de l'exercice prévu. Ce dataset doit être obligatoirement supprimé à la fin du test technique.

A clause can be defined as a structural unit in a legal document. In general, one clause deals with one specific contractual topic: confidentiality, designation... You'll use these clauses as the basis for your analysis.
You don't have to use the PDF files, they are mainly here to help you better understand the data and the tasks.

3. Your tasks

a. Extract different categories of clauses

After reading some statutes, you'll understand that all the contracts basically follow the same structure and have similar clauses (e.g. "Capital social", "Siège Social").

Q1. Retrieve, in an unsupervised fashion, the different categories of clauses that can be found in the corpus. You can use any clustering techniques. Try to assess your results with some qualitative analysis.

b. Extract variations of a clause

Even if the clusters previously extracted are quite clean, you'll probably find out that inside a single cluster, a lot of variations can happen (e.g. "Clause Objet").

Q2. How would you extract these variations? Try to use some semantic similarity scoring here.

c. (optional) Train a clause classifier

Let say your client gives you another bunch of statutes. He/She wants you to automatically classify the different clauses in these new statutes.

Q3. Based on the clusters extracted in Q1., train a clause classifier. Validate your results with some quantitative analysis.