Projet NLP

# Pretrained LLM

Prétraitement des données :

Parties importantes : <section class="dsr-section" data-number="1.7" data-title="ABROGATION DE DISPOSITIONS ANTERIEURES" data-type="BodySection.CHAPTER">

Transformer html en json

# Regex

Patterns pour l’identification du document :

**Avec dates:**

- ARRÊTÉ du 20 AVR. 2020

- Arrêté du 22 février 2023

**Avec nombres:**

- ARRÊTÉ N° 02/IC/566

- ARRÊTÉ … N° 07/IC/297

- Arrêté complémentaire n° 5103 - 11-45

- ARRETE N° 5103-2013-48

- ARRÊTÉ D'AUTORISATION PRÉFECTORAL N°1122-21-20-026

- ARRÊTÉ PRÉFECTORAL COMPLÉMENTAIRE N°1122-20-20028

- Arrêté préfectoral complémentaire n° M21-21-20-064

- Arrêté préfectoral complémentaire N°122-22-20-032

- Arrêté n° 2350-22-00082

- Arrêté n° 2021-14-338

- Arrêté n° D1-B1-17-143

- ARRETE n° 2008201PC(064.00001)

- ARRETE PREFECTORAL n°2016-58

- Arrêté n°2020-334-PC

- Arrêté n°2023-58-MED

- Arrêté n°2023-323-URG

- Arrêté n° 2015015-0009

- ARRÊTÉ N° 2015/110059

- ARRÊTÉ N° 20151005-2

🡪 Liste de regex

Autre:

- ARRÊTÉ DE MISE EN DEMEURE

- ARRÊTÉ PRÉFECTORAL DE MISE EN DEMEURE

- ARRÊTÉ D’AUTORISATION

- Arrêté de prescriptions complémentaires

- ARRÊTÉ PRÉFECTORAL COMPLÉMENTAIRE

- ARRÊTÉ PRÉFECTORAL

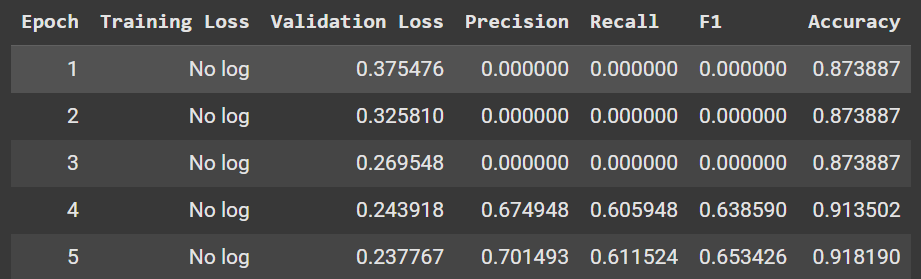
- ARRÊTÉ

Problèmes :

* N° devient Ndeg
* <a>ARRETE PREFECTORAL COMPLEMENTAIRE modifiant l'arrete prefectoral ndeg994156 du 21 decembre 1999</a>

Avec la nouvelle base de données, on peut entrainer notre LLM !

# Pretrained LLM fined-tuned on files with regex

Il arrive enfin à prédire autre chose que zéro

Test :

<header class="dsr-header"> <div class="dsr-entity"> <div> Préfecture </div> <div> Direction des Collectivités Locales et des Produits Publics </div> <div> Bureau des Foyers Publics et Installations Classées </div> </div> <div class="dsr-identification"> **<h1> <a>A</a>RRÊ<a>TÉ du 20 AVR</a>. 2020\*\* portant** autorisation d'exploiter une unité de valorisation énergétique de combustibles solides de récupération (CSR), de déchets d'activité économique (DAE) et d'ordures ménagères (OM) sur le territoire de la commune de Bantzenheim à la société B+T ÉNERGIE France Sas en référence au titre VIII du livre I et au titre I° du livre V du code de l'environnement </h1> </div> <div class="dsr-visa"> VU le code de

Ce n’est pas mal !

class="dsr-visa"> VU l'arrêté ministériel du 23 mai 2016 relatif aux installations de production de chaleur et/ou d'électricité à partir de déchets non dangereux préparés sous forme de combustibles solides de récupération dans des installations prévues à cet effet associées ou non à un autre combustible et relevant de la rubrique 2971 de la nomenclature des installations classées pour la protection de l'environnement ; </div> <div class="dsr-visa"> **VU l'arrêté ministériel du <a>20</a> septembre 2002 relatif** aux installations d'incinération et de co-incinération de déchets non dangereux et aux installations incinérant des déchets d'activités de soins à risques infectieux ; </div> <div class="dsr-visa">

Mais moins bien que les regex

Sources :

* [Fine-Tuning Large Language Models For Named Entity Recognition | by Nitin Kumar | Medium](https://medium.com/@nk94.nitinkumar/fine-tuning-large-language-models-for-named-entity-recognition-da97c38df742)
* Hugging face : https://huggingface.co/docs/transformers/tasks/token\_classification