### **1. Un projet complet**

L'archive [imdb\_smol.tar.gz](https://hub-binder.mybinder.ovh/user/loicgrobol-appr-ificiel-environ-3oeogpwd/notebooks/apprentissage-artificiel/slides/06-scikit-learn/data/imdb_smol.tar.gz) (aussi disponible [dans le dépôt](https://github.com/LoicGrobol/apprentissage-artificiel/blob/main/slides/06-scikit-learn/data/imdb_smol.tar.gz)) contient 602 critiques de films sous formes de fichiers textes, réparties en deux classes : positives et négatives (matérialisées par des sous-dossiers). Votre mission est de réaliser un script qui :

* Charge et vectorise ces données
* Entraîne et compare des classifieurs sur ce jeu de données

L'objectif est de déterminer quel type de vectorisation et de modèle semble le plus adapté et quels hyperparamètres choisir. Vous pouvez par exemple tester des SVM comme ci-dessus, [un modèle de régression logistique](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html), [un arbre de décision](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html), [un modèle bayésien naïf](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html) ou [une forêt d'arbres de décision](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html).

### **2. D'autres traits**

Essayez avec d'autres *features* : La longueur moyenne des mots, le nombre ou le type d'adjectifs, la présence d'entités nommées, …

Pour récupérer ce genre de *features*, vous pouvez regarder du côté de [spaCy](http://spacy.io/) comme prétraitement de vos données.

***(3. Et les réseaux de neurones ?***

scikit-learn permet d'utiliser un Multi-layer Perceptron (MLP). Et comme la bibliothèque ne permet pas d'utiliser un GPU pour les calculs, son utilisation est limitée à des jeux de données de taille moyenne.

scikit-learn n'est pas fait pour le *deep learning*. Il existe des bibliothèques associées qui permettent de combiner Keras ou pytorch avec scikitlearn néanmoins.

Essayez en suivant [la doc](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html)

Il y a encore plein d'autre choses marrantes à faire avec scikit-learn et on en verra, mais en attendant vous pouvez aller voir [leur exemple sur ce dataset](https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html#sphx-glr-auto-examples-text-plot-document-classification-20newsgroups-py) qui a de bien jolis graphiques.***)***