**Classification des textes**

Le but de ce TP est d’implémenter un classificateur de texte en s’appuyant sur plusieurs modélisations par exemple naïve de Bayes, SVM etc. On dispose d’un corpus de 1000 textes positifs et de 1000 textes négatifs :

* Diviser aléatoirement ce corpus en **corpus** **d’entrainement** et **corpus de test** (en premier temps opter pour une division aléatoire de 80% de textes pour l’entrainement et 20% pour la phase des tests). Les textes sont disponibles sur le réseau.
* Prétraiter les données en ne gardant que les noms, verbes, adjectifs et adverbes.
* Charger le corpus d’entrainement et de test avec *sklearn.datasets.load\_files*
* Vectoriser les données avec *sklearn.feature\_extraction.text.CountVectorizer*
* Indexer les données avec *sklearn.feature\_extraction.text.TfidfTransformer*
* Créer vos modèles d’entrainement (naif Bayes, SVM, …)
* Remplacer les opérations suivantes par un pipeline *sklearn.pipeline.Pipeline*
* Évaluer votre modèle de classification en l’appliquant les métriques proposées dans *sklearn.metrics*

Durée de travail 3 semaines.

À rendre le 21/4/2017