Plan de travail prévisionnel

Azim Makboulhoussen

*OpenClassrooms*

*Projet 8 – Parcours Data Scientist*

*16 Juillet 2018*



# Présentation de la thématique choisie

L’apprentissage machine connait un essor formidable depuis quelques années. Pour apprendre, les algorithmes de *Machine Learning* doivent être entrainés sur beaucoup de données pour obtenir des taux de réussite suffisant.

Ces données doivent être fiables, cohérentes et correctement labellisées. Leur préparation demande donc un gros travail qui va de leur récupération, mise en forme, nettoyage et enfin de l’étiquetage de celles-ci. C’est donc une tâche particulièrement fastidieuse car les quantité de données sont souvent massives.

Pour pallier à cette problématique, l’université de Stanford a mise au point un système appelée **Snorkel**, un système open-source qui adopte une nouvelle approche pour la création, la modélisation et la gestion de données d’entrainement pour les systèmes de prédictions.

Nous avons choisi d’étudier cette solution pour ce projet de veille thématique.

# Les sources biographiques

Pour ce travail, nous nous baserons sur les publications de recherches suivantes :

* ***Data Programming: Creating Large Training Sets, Quickly*** - 08/01/2017 (<https://arxiv.org/abs/1605.07723>)
* ***Learning the Structure of Generative Models without Labeled Data*** - 09/09/2017 (<https://arxiv.org/abs/1703.00854>)
* ***Snorkel: Rapid Training Data Creation with Weak Supervision*** - 28/11/2017 (<https://arxiv.org/abs/1711.10160>)

La solution **Snorkel** a été présentée à la conférence VLDB (Very Large Data Bases) en 2018. C’est l’une des plus prestigieuses conférence dans le domaine de la recherche et développement des bases de données.

# Le prototype à implémenter

Pour ce projet, nous allons nous baser sur le tutoriel mis à disposition par l’équipe en charge du développement de **Snorkel**. Celui-ci nous permettra de monter en compétence sur cette technologie. L’exercice proposé consiste à identifier dans un corpus d’articles d’informations les mentions des époux. Par exemple, si on trouve la phrase « Le président Emmanuel Macron et sa femme Brigitte reçoivent les joueurs de l’équipe de France à l’Élysée aujourd’hui », il nous faudra extraire la relation maritale : (« Emmanuel Macron », « Brigitte).

Nous verrons dans un 2ème temps, si nous pouvons appliquer cette nouvelle technique sur une des problématiques que nous avons déjà rencontrée dans le cadre de ce parcours Data Scientist.