Leveraging weak-supervision for automated NER dataset generation

DESCRIPTION

Our EY Data Science team is looking for creative postgraduate students (Master and PhD students in computer science, engineering or related fields) who are passionate about emerging topics of artificial intelligence and machine learning to join us for a 3-6 month internship.

As an intern, you'll partner with other data scientists and machine learning engineers to build one or several Machine Learning assets to serve real-world applications in the industry.

These last years, there has been an explosion of interest in machine-learning-based systems across industry. A main driver has been the advent of deep learning techniques, which can learn task-specific representations of input data, obviating what used to be the most time-consuming development task: feature engineering. However, deep learning has a major upfront cost: these methods need massive training sets of labelled examples to learn from often tens of thousands to millions to reach peak predictive performance. Such training sets are enormously expensive to create and labelling training data is increasingly becoming the largest bottleneck in building/deploying ML systems.

The goal of this internship is to explore and leverage the data programming paradigm and weak supervision for automated data annotation (e.g. Stanford's Snorkel) in order to design and build an end-to-end solution which automates the generation of the dataset required to train a custom NER (Named Entity Recognition) model. NER is the most common and challenging task we encounter within the context of unstructured documents intelligence use cases.

SUPERVISION:

- Works independently and collaborates effectively as part of a team of Data Scientists, Machine Learning Engineers and Academic Partners

QUALIFICATIONS

- On-track for a master degree in Computer Science, Engineering, Statistics or Mathematics
- Strong technical knowledge in algorithms and data structures
- Good understanding and strong personal interest in machine learning
- Strong technical knowledge in programming/scripting languages; experience in Python is a plus
- Strong software development skills
- Enthusiasm for applying machine learning to real-world problems
- Ability to present your ideas clearly
- Strong personal interest in learning, researching, and creating new technologies with high business impact