T7 – Science des données et apprentissage automatique

- http://devlog.cnrs.fr/jdev2017
- http://devlog.cnrs.fr/jdev2017/t7

Objectifs

Les technologies et les outils, les méthodes et les techniques d'analyse pour faire parler les masses de données.

Description

La science des données et l'apprentissage automatique sont au coeur de la recherche par les données. L'essor des données, l'évolution des technologies d'acquisition, de stockage, d'accès, de traitement et d'analyse ouvrent de nouveaux horizons.

Nous nous intéresserons aux compétences et savoir-faire nécessaires en informatique et en analyse pour concevoir, mettre en oeuvre les solutions pour faire émerger du sens aux données.

Aujourd'hui, le scientifique de la donnée se doit de maîtriser toute la chaîne, de l'acquisition (qualité, contrôle, intégration, …) à la prédiction ou à la prise de décision (restitution, visualisation, …) en passant par le traitement et l'analyse (modélisation statistique, machine learning, …).

Planning

Agenda à checker avec http://www.jdev2017.fr

- Mardi Matin: AP01, AP02, AP04, AP05, GT05
- Mercredi Matin: AP03, A04, GT01/GT02
- Mercredi après-midi: T7.P présentations plénières et T7.GT06 à 17h35-18h45
- jeudi Matin: A01, A02, A05,
- jeudi après-midi: A06, A07, GT03/GT04, GT07/GT08

(A) Atelier , (GT) Groupe de Travail, (P) les présentations.

Agenda générale

Mots clés

- IA (Intelligence Artificielle) (AI)
- Big data
- data scientist
- Outils informatiques

- Outils statistiques
- Apprentissage supervisé et non-supervisé
- Apprentissage automatique (Machine learning)
- deep learning
- Prédiction et prise de décision
- Algorithmiques et implémentation
- Architectures distribuées (Hadoop, Spark, Mapreduce)
- Structure des données (arbres, forêt, pile, file d'attente,..)
- Optimisation et complexité
- Langages (python, julia, R, …)
- ETL et alignement de données
- Valorisation des données
- Business Intelligence

Public

- réseaux Devlog, calcul, rdbb, …
- GDRs GPL, MADICS, …
- data scientist, data engineer, software engineer

Présentations

- **14h00-14h40**: Historique, évolutions, tendances, ressources et outils de l'apprentissage. Synthèse des techniques en fonction des domaine et problème (ouverture sur le deep learning) - **Liva Ralaivola (équipe QARMA@LIF, Marseille)** et **Hachem Kadri (équipe QARMA@LIF, Marseille)**
- **14h40-15h20** : Apprentissage statistique pour la recherche par les données - **Sébastien Dejean, Laurent Risser (IMT, Toulouse)**
- **15h20-15h40** : Pause
- **15h40-16h20** : Le centre de données scientifique de Paris-Saclay, l'éco-système pour la recherche par les données Science des données et recherche par les données - retour d'expérience du Paris-Saclay CDS (focus sur l'inférence de connaissance) - **Balazs Kegl** (IN2P3,Paris-Saclay)
- **16h20-17h00** : Tutoriel sur le Deep Learning - **Benoit Favre (équipe TALEP@LIF, Marseille)** et **Thierry Artières (équipe QARMA@LIF, Marseille)**
- **17h00-17h40** : La visualisation des données et python (dataviz, pandas)" - **Romain Vuillemot (Ecole Centrale Lyon, laboratoire LIRIS)**. Principes de base en visualisation; panorama des bibliothèques existantes (matplotlib, seaborn, bokeh, plot.ly); utilisation de Notebooks.

Ateliers préparatoires

- T7.AP01 : Initiation python - **Laurent Risser** (IMT)
- T7.AP02: Manipulation de données pour débutants avec R - **Sébastien Dejean** (IMT, Toulouse)
- T7.AP03 : Les bases du calcul scientifique avec Python. Python pour le calcul scientifique" (numpy, scipy, scikitlearn, networkx, sympy, etc. - **Tristan Colombo** (Éditions Diamond, GNU/Linux Magazine France)
- T7.AP04 Manipulation de données pour débutants en Julia - **Frederic Pont** (CRCT/oncopole/Inserm)
- T7.AP05 : Initiation au DeepLearning (DIGITS/Caffe) - **Gunter Roth** (NVIDIA)

Ateliers

- T7.A01 : Python pour l'apprentissage - **Laurent Risser** (IMT)

Notebooks (1, 2) (Machine learning avec ScikitLearn?)

- T7.A02 : R pour l'apprentissage - **Sébastien Dejean**(IMT Toulouse) (Comparaison d'algorithme de machines learning en R)
- T7.A04 : Prototypage rapide de modèles prédictifs dans l'environnement collaboratif du centre de données scientifiques de Paris-Saclay (RAMP): RAMP (Rapid Analytics and Model Prototying) - **reconnaissance d'image d'insecte (donc deep learning)** - **Balazs Kegl** (IN2P3,Paris-Saclay)
- T7.A05 : Prototypage rapide de modèles prédictifs dans l'environnement collaboratif du centre de données scientifiques de Paris-Saclay (RAMP): RAMP (Rapid Analytics and Model Prototyping) - **séries temporelles (Sea ice, El Nino)** - **Balazs Kegl** (IN2P3,Paris-Saclay)
- T7.A06 : RAMP (Rapid Analytics and Model Prototyping) - RAMP cas complexe (drug spectra ou HEP anomaly) - **Balazs Kegl** (IN2P3,Paris-Saclay)
- T7.A07 : Intro aux outils pour les DataScience avec Go - **Sébastien Binet** (CERN)

Groupes de travail

- T7.GT01 : Préparation des données pour l'analyse statistique et le machine learning (mise en oeuvre avec R) - **Sébastien Dejean** (IMT, Toulouse)

Réference Tidy data (bien ranger les données) de Hadley Wickham

- T7.GT02 : Retour d'experience sur l'utilisation de perl, R, Julia, GO, Python, OpenCL en sciences médicales. - **Laurent Risser** (IMT) et **Frederic Pont** (CRCT/oncopole/Inserm)
- T7.GT03 : Préparation des données avec Pandas, Python Data Analysis Library. **Laurent Risser** (IMT) et **Yves Auda** (GET/OBS-MIP)
- T7.GT04 : Un outil pour la transcription et la recommandation temps-réel durant une vidéo-conférence. Reconnaissance de la parole et machine learning : OpenPaaS solution de PaaS

pour service collaboratif avec videoconf en mode rtc avec reconnaissance de la parole. LINAGORA en relation avec la T5. HUBL.IN - **Tom Jorquera (LINAGORA)**

- T7.GT05 : architectures récentes de réseaux de neurones profonds, notion de saillance - **Thomas Pellegrini** (IRIT/UPS)
- T7.GT06 : Services d'analyse et machine learning à portée de tous sur AWS. Quand les dashboards ne suffisent plus, la plate-forme AWS permet de déployer des solution simples et innovantes pour utiliser les méga-données au quotidien pour tout un chacun. **Moshir Mikael/AWS.**
- T7.GT07 : Rex sur Les librairies et toolbox de Deep Learning sur GPU: Digits, Caffee, Magma, Torch, … - **Gunter Roth** (NVIDIA)
- T7.GT08 : ANNULE : Retours d'expérience par NVIDIA en deep learning : cas d'usage smartcity, telco, industry, robotic. REX de 5 à 10 minutes - **Francois Courteille** (NVIDIA).