RCP216 - Ingénierie de la fouille et de la visualisation de données massives

Mis à jour le 31-01-2025

Code : RCP216

Unité d'enseignement de type mixte

6 crédits

Volume horaire de référence (+/- 10%) : **50 heures**

Responsabilité nationale :
EPN05 - Informatique / 1

Contact national :

EPN05 - Informatique

2 rue Conté
33.1.9A
75003 Paris
01 58 80 87 99
Jean-mathieu Codassé
jean-mathieu.codasse@lecnam.net

Présentation

Prérequis

Bonnes connaissances mathématiques et statistiques générales, maîtrise de méthodes statistiques pour la fouille de données, connaissance de techniques de gestions de données massives faiblement structurées, connaissance de techniques de passage à l'échelle par distribution. Connaissance d'au moins un langage de programmation.
Vous êtes encouragés à évaluer votre capacité à suivre cette UE en répondant au questionnaire en ligne accessible sur https://cedric.cnam.fr/vertigo/Cours/RCP216/questionnaire.html. Vous pouvez répondre sans vous identifier, le résultat vous est donné immédiatement et n'est pas enregistré.

Objectifs pédagogiques

Cet enseignement s'intéresse à l'impact des caractéristiques des données massives (volume, variété, vélocité) sur les méthodes de fouille de données. Sont examinées les approches actuelles qui permettent de faire passer à l'échelle les méthodes de fouille, en insistant sur les spécificités des opérations de fouille en environnement distribué.
Les caractéristiques mentionnées sont ensuite considérées de façon plus spécifique pour certains problèmes fréquents dans le traitement des données massives. Sont ainsi abordés les systèmes de recommandation et la recherche efficace par similarité, la classification automatique et l'apprentissage supervisé sur une plate-forme distribuée, les opérations spécifiques au traitement des données textuelles souvent hétérogènes, les implications de la vélocité sur la fouille de flux de données, l'analyse de grands graphes et de réseaux sociaux.
L'UE s'intéresse également au rôle de la visualisation et de l'interaction, non seulement dans la présentation des résultats mais aussi dans les opérations de fouille de données.

Compétences

Réaliser la fouille de données massives en utilisant une plate-forme de calcul distribué (Spark) via JupyterHub. Mettre en place un système de recommandation. Réaliser la fouille de textes en exploitant des encodages (*word embeddings*) et des modèles de langage (*language models*) en se servant d'une bibliothèque logicielle évoluée (SparkNLP). Mettre en œuvre une visualisation pertinente des données. Traiter des données en flux. Construire des modèles descriptifs et décisionnels sur des données massives. Evaluer des critères observationnels d'équité des prédictions et modifier un modèle prédictif pour respecter des critères d'équité.

Programme

Contenu

1. Introduction : applications, typologie des données, typologie des problèmes
2. Approches : réduction de la complexité, distribution
3. Passage à l'échelle de quelques problèmes fréquents
    a. Recherche par similarité, systèmes de recommandation
    b. Classification automatique
    c. Fouille de données textuelles
    d. Fouille de flux de données
    e. Apprentissage supervisé à large échelle
    f. Fouille et visualisation de graphes et réseaux sociaux
4. Visualisation d'information : historique, applications, outils
5. Aspects éthiques dans la fouille de données

Le cours est complété par des travaux pratiques (TP) permettant de mettre en pratique des techniques présentées. Ces TP seront réalisés à l'aide de Apache Spark pour la fouille de données et de réseaux sociaux, et à l'aide de Gephi pour la visualisation de graphes. Pour les

travaux pratiques comme pour le travail sur le projet les auditeurs peuvent utiliser le JupyterHub du Cnam.
Les supports de cours et de TP, ainsi que d'autres explications concernant le déroulement de l'UE sont accessibles à partir de https://cedric.cnam.fr/vertigo/Cours/RCP216/

Modalités de validation

- Projet(s)
- Examen final

Description des modalités de validation

Note finale = ((note de projet + note d'examen) / 2).

Bibliographie

Titre	Auteur(s)
Advanced Analytics with Spark. O'Reilly.	Ryza, S., U. Laserson, S. Owen and J. Wills.
Mining Massive Datasets. Cambridge University Press, New York, NY, USA.	A. Rajaraman and J. D. Ullman.