

Gestion des données et accès à l'information

Introduction

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

Master Statistiques
Sorbonne Université

2024-2025

Page web

https://schwander.isir.upmc.fr/enseignement/m2stat_gdai/

Contenu du cours

Contenu

- ▶ Bases de données (relationnelles, noSQL)
- ▶ Recherche d'information et utilisation des modèles de langue
- ▶ Extraction de données (services web, pages web, emails)

Évaluation

- ▶ Mini-projet (50%)
- ▶ Examen final (50%)

Travaux pratiques

Sur vos machines

- ▶ Prêt d'ordinateur possible

Si besoin: machine virtuelle

- ▶ Fichier fourni sur le site du cours
- ▶ Machine Linux avec tout installé dessus

Pour un statisticien et machine-learner

Chaîne de traitement complète

- ▶ *Acquérir les données*
- ▶ *Stocker les données*
- ▶ *Mettre à jour les données*
- ▶ Traiter les données
- ▶ Visualiser, faire des rapports

Objectifs

- ▶ Maîtriser la totalité de la chaîne
- ▶ Comprendre la conception des systèmes
- ▶ Connaître les outils

Bases de données

Stockage structuré des données

Logiciels pour le stockage

- ▶ Un serveur accessible à travers un réseau (souvent)
- ▶ Permettant de lire et d'écrire
- ▶ Garantissant des propriétés intéressantes

Langage de requête

- ▶ Exprimer des demandes
- ▶ Filtrer pour n'obtenir que les résultats intéressants
- ▶ Efficacement
- ▶ Standardisé (SQL) ou pas (noSQL)

Accès à l'information

Requêtes

- ▶ Langage de requêtes
- ▶ Mots-clés
- ▶ Langue naturelle

Indexation

- ▶ Passage de la données brutes à quelque chose qu'on peut retrouver
- ▶ Structure d'une base de données
- ▶ Représentation de documents non-structurés

Extraction des données

Comment récupérer les données

- ▶ Depuis une base de données
- ▶ Depuis une interface documentée
- ▶ Depuis des fichiers dans un format structuré
- ▶ Depuis des fichiers dans un format non-structuré
- ▶ Depuis une interface non-documentée
- ▶ Depuis une interface non-coopérative, non-documentée, qui peut changer

Quelques bibliothèques, techniques et outils

- ▶ Pour des services web
- ▶ Pour des pages web
- ▶ Pour des emails