Fouille de textes

Introduction au module X3ITM10

Solen Quiniou

solen.quiniou@univ-nantes.fr

Université de Nantes

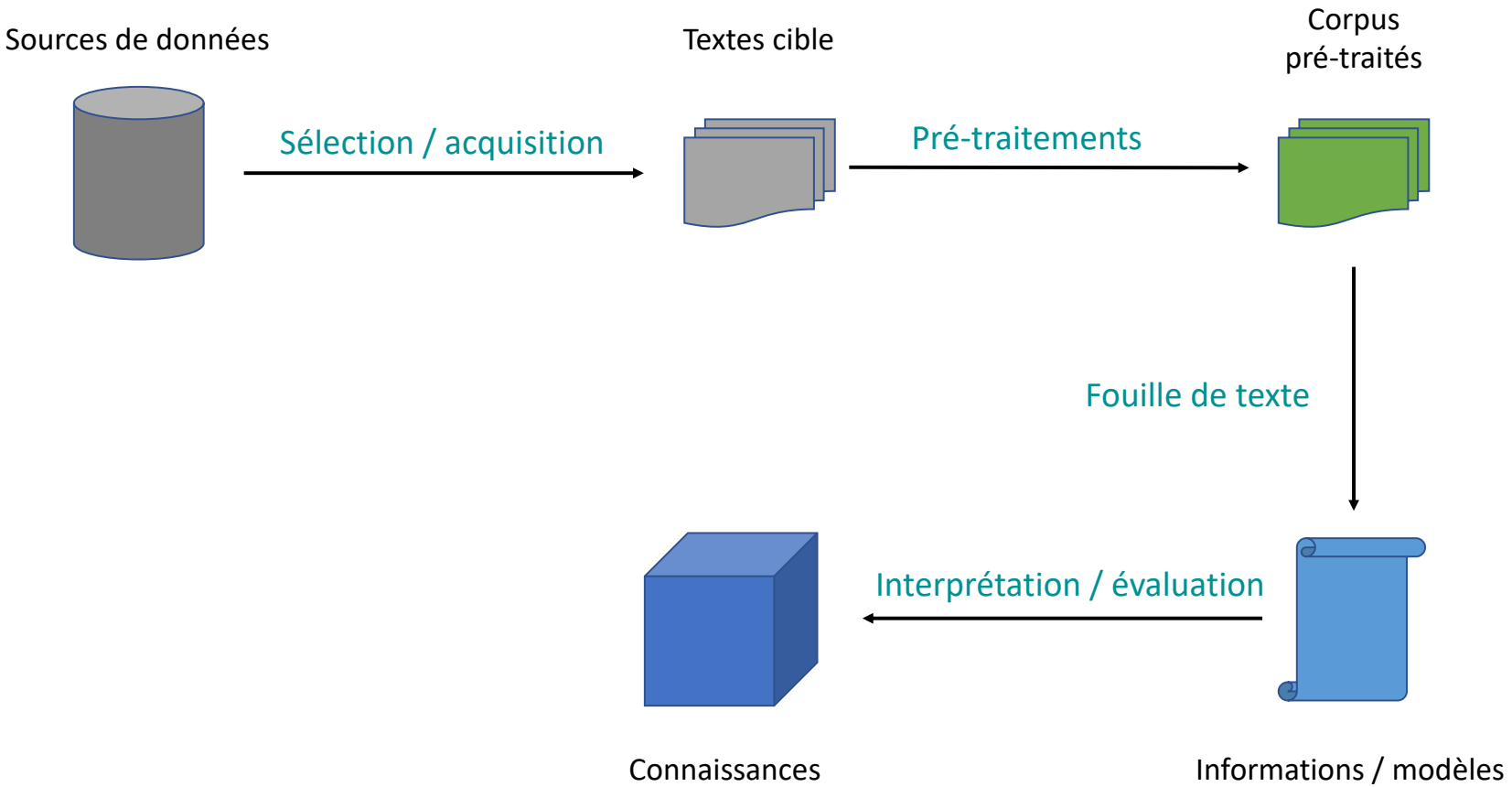
Année 2020-2021

Introduction à la fouille de données textuelles

- Fouille de données textuelles : processus d'extraction d'informations inconnues *a priori*, à partir de grands volumes de textes (appelés corpus)
  - → Les textes peuvent correspondre à de simples fichiers texte, des pages web, des emails, des sms, des tweets. . .

- Principales tâches
  - Structurer automatiquement un ensemble de textes en plusieurs groupes homogènes
  - Affecter des textes à des catégories (ou classes) prédéfinies
  - Suivre, dans une collection d'articles, l'évolution d'un sujet ou les changements de sujets
  - Faire de la veille scientifique, surveiller la concurrence
  - Extraire une sous-partie des informations contenues dans les textes
  - . . .

Schéma global d'un processus de fouille de textes



Organisation du module

- Volume horaire : 12 CM + 6 TD

- Enseignants
  - Solen Quiniou (Nantes)
  - Nathalie Camelin (Le Mans)
  - Nicolas Dugué (Le Mans)

- Notions abordées dans le cours
  - Représentation vectorielle du texte
  - Similarité syntaxique et sémantique
  - Représentation par plongements de mots
  - Catégorisation/classification de textes
  - Reconnaissance d'entités nommées
  - Fouille de motifs
  - Fouille d'opinion

- Composition de la note du module
  - Devoir écrit : 15 points
  - Présentation d'un article scientifique en groupe : 5 points