Recueil et structuration de corpus

ECTS
2 crédits

Présentation

DESCRIPTION

SL4BY050

De la détection d'encodage à la normalisation de données en passant par l'extraction de contenus Web, les étudiants aborderont au cours de ce module les étapes préalables à la constitution d'un corpus textuel en vue de son exploitation par les outils état de l'art. Ils tireront partie de méthodes de nettoyage et de structuration automatisées (python, perl) pour produire des documents dans des formats variés (TXT, CSV, XML/TEI).