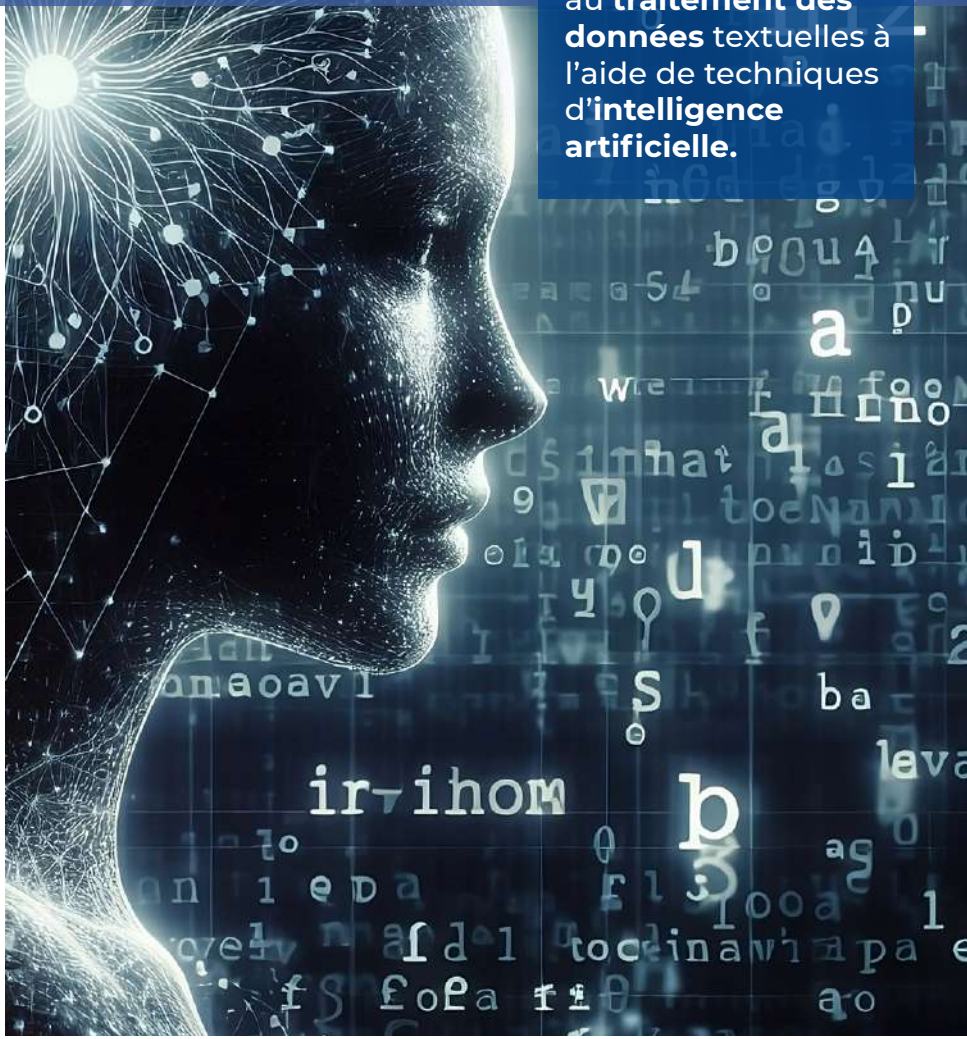
École d'été

Text mining en Python

Du 8 au 12 juillet 2024

Formation intensive au **traitement des données** textuelles à l'aide de techniques d'**intelligence artificielle.**



UCLouvain
Faculté des sciences

FORMATION CONTINUE

CENTAL
centre de traitement automatique du langage

LIDAM
SMCS | Louvain Institute of Data Analysis and Modeling in economics and statistics

L'ÉCOLE D'ÉTÉ

PRÉSENTATION DE LA FORMATION

L'UCLouvain, par l'intermédiaire de sa plateforme de Support en Méthodologie et Calcul Statistique (SMCS), offre à la communauté scientifique et aux entreprises la possibilité de se former à plusieurs techniques d'analyse de données. Cet été, le SMCS et le Centre de Traitement automatique du Langage (CENTAL) vous proposent une nouvelle école d'été pour vous permettre d'acquérir des compétences pratiques et relativement avancées en ***text mining*** à l'aide du langage de programmation Python.

Le traitement automatique de la langue (TAL ; en anglais *natural language processing* – NLP) est une discipline qui se situe à l'intersection de la **linguistique**, de **l'informatique** et de **l'intelligence artificielle**. Il utilise un ensemble de technologies qui permettent aux ordinateurs de « comprendre », d'analyser et de générer du **langage humain**. Le moteur de recherche de Google, l'application Siri d'Apple ou chatGPT d'openAI montrent à quel point les résultats des recherches dans cette discipline font **partie intégrante de notre quotidien.** Qu'il s'agisse d'extraire des dates dans des courriels pour planifier un évènement ou d'identifier les actions à effectuer à la suite d'une commande vocale, **la structuration de l'information et l'analyse du contenu** sont essentielles.

La formation qui vous est proposée est idéale pour toute personne ayant déjà une **base assez solide en programmation Python** et souhaitant élargir ses compétences dans ce domaine en pleine croissance.

Les **domaines d'application** du TAL sont **variés** et vont de l'extraction d'information dans des procès-verbaux de réunion à la structuration des documents médicaux, en passant par le suivi de la notoriété d'une marque sur les réseaux sociaux et l'analyse des contrats juridiques. Contrairement aux éditions précédentes, la formation de cette année n'est pas consacrée à une thématique spécifique. Elle explorera diverses facettes du *text mining* au travers de quelques exemples d'applications phares, favorisant ainsi une **compréhension globale et polyvalente** de ce domaine en constante évolution.

LE CENTAL

Le CENTAL est une plateforme technologique de l'UCLouvain impliquée à la fois dans des activités de recherche et d'enseignement. Il collabore également aux projets de plusieurs centres de recherche de l'UCLouvain auxquels il apporte son expertise en matière de **traitement informatique des données textuelles**. Les contacts spécifiques avec les entreprises sont également nombreux et prennent la forme de prestations ponctuelles, d'activités de conseil (guidance scientifique de plus ou moins longue durée) ou encore de projets de recherche et développement.

LES FORMATEURS

DAMIEN DE MEYERE est linguiste-informaticien au CENTAL. Diplômé en 2014 du master en linguistique à finalité spécialisée en traitement automatique du langage de l'UCLouvain, il a collaboré à divers projets de recherche, en développant à la fois des outils d'extraction d'information appliqués à des données textuelles de nature diverse (textes médicaux, administratifs, juridiques ou issus de réseaux sociaux) et des modalités d'interrogation et de visualisation des connaissances générées.

THOMAS FRANÇOIS, professeur de linguistique appliquée à l'UCLouvain, travaille depuis plus de 15 ans dans le domaine du traitement automatique du langage. Son expertise porte sur l'application de techniques d'apprentissage automatisé à des problématiques linguistiques. Ses travaux se sont vus récompensés par le prix de la meilleure thèse de l'ATALA en 2012, par le prix du meilleur article à la conférence TALN en 2016 et sont régulièrement publiés dans les meilleures conférences du domaine (ACL, EMNLP, Coling, etc.).

HUBERT NAETS a travaillé pendant quatre ans au Commissariat à l'énergie atomique (Fontenay-aux-Roses, Paris), avant de rejoindre le CENTAL de l'UCLouvain en 2007 en tant que linguiste-informaticien. Son travail et ses recherches portent sur des domaines très variés liés au traitement automatique des langues, dont notamment le traitement de corpus en langues modernes et anciennes, l'acquisition de langue seconde ou encore l'extraction d'informations sémantiques.

PATRICK WATRIN est devenu responsable opérationnel du CENTAL, après avoir réalisé un doctorat à l'UCLouvain, un post-doctorat à l'Institut Gaspard Monge et avoir créé une Spin-Off, Earlytracks, active dans le secteur de l'information médicale. Pendant plus de quinze ans, il a mené des recherches dans le domaine de l'extraction et de la structuration d'information et se consacre maintenant à l'étude des réseaux de neurones, à leur application sur des données réelles et aux questions relatives à leur interprétabilité.

CONTENU DE LA FORMATION

Cette école d'été a pour objectif de vous familiariser au *text mining*, c'est-à-dire un ensemble de techniques qui reposent sur des algorithmes, modèles statistiques et ressources linguistiques permettant de traiter les données textuelles stockées sur un support informatique. Vous en découvrirez de multiples usages au travers de différentes techniques issues de l'apprentissage automatique (similarité, clustering, machine learning, deep learning, larges modèles de langue). Le programme est divisé en modules qui mêlent à la fois exposés théoriques et mises en situation pratiques. L'ensemble offre une chaîne de traitement complète de *text mining* utilisant le langage de programmation Python.

ACQUISITION ET PRÉTRAITEMENT DE DONNÉES

Les dernières avancées en TAL s'appuient largement sur des techniques de deep learning qui exigent des volumes considérables de données textuelles pour optimiser leur efficacité. Cette première journée vous initiera à la problématique de l'acquisition et de la préparation de ces données pour une utilisation au sein de réseaux de neurones. Vous aurez la possibilité de collecter et de normaliser des documents, d'identifier les éléments de contenu pertinents et de leur associer différentes métadonnées afin d'obtenir un corpus d'entraînement propre et documenté. Par ailleurs, vous découvrirez des outils utiles pour l'exploration et la visualisation de ces données. Enfin, vous appréhenderez les concepts de base de réseaux de neurones et des techniques essentielles pour le chargement efficace et l'utilisation optimale des données.

SIMILARITÉ ET PARTITIONNEMENT (CLUSTERING) DE DOCUMENTS

Déterminer si deux documents sont similaires ou proches est une tâche fondamentale en *text mining*. On peut vouloir identifier des textes qui partagent les mêmes mots ou certaines séquences de mots (détection de plagiat, par exemple) ou établir si des textes sont sémantiquement proches (recherche d'information, suggestion de documents proches). Au terme de ce module, vous serez en mesure de comparer des textes en utilisant des approches basées sur les fréquences des mots, la modélisation thématique (topic modeling), ou encore des réseaux de neurones (word2vec, transformers).

Dans certains cas, on veut diviser un ensemble de documents en plusieurs sous-ensembles qui partagent des caractéristiques communes. On parle alors de clustering. Vous serez en mesure d'utiliser plusieurs techniques de partitionnement de documents à la fin du module.

CLASSIFICATION DE DOCUMENTS

Vous serez confronté aux enjeux et défis des méthodes de classification pour la détection automatisée de caractéristiques des textes (ex. : catégorisation thématique de documents, évaluation automatisée d'essais, analyse des sentiments ou des opinions, etc.) ou de caractéristiques de leurs auteurs (identification de hobbies à des fins de marketing, etc.). Vous prendrez conscience des difficultés liées à la définition et à l'annotation en classes, ainsi que des questions relatives à la création et à la manipulation d'un grand nombre de variables linguistiques. Le module présentera d'abord la méthodologie de conception et d'évaluation de modèles de classification spécifiques au TAL et au deep learning. Ensuite, seront abordés les principaux algorithmes de machine learning (arbres de décision, SVM) et de *deep learning* non récurrents (perceptron multi-couches, transformer et mécanisme d'attention).

ÉTIQUETAGE AUTOMATIQUE

Vous découvrirez les techniques essentielles d'étiquetage automatique de textes, c'est-à-dire l'attribution de catégories linguistiques ou de classes sémantiques aux (séquences de) mots d'un document. Ce module permettra d'aborder les réseaux de neurones récurrents (LSTM, GRU…), de même que les stratégies d'analyse séquentielle et conduira à développer un système de reconnaissance d'entités nommées (ex. : noms de personnes, de lieux, d'entreprises…).

MODÈLES GÉNÉRATIFS

Dans le cadre d'une communication, Didier Schwab, professeur à l'Université de Grenoble et coordinateur du groupe de recherches à l'origine du modèle de langue Flaubert, présentera les dernières avancées en matière de modélisation générative. Il exposera les récents développements en matière de grands modèles de langues (LLM) pour le français et, en particulier, les travaux qu'il supervise dans le cadre du projet Pantagruel. Pantagruel vise à développer des modèles autosupervisés librement accessibles pour le français, comprenant une à trois des modalités (écrit, oral et pictogrammes) pour les domaines généraux et cliniques.

PRÉREQUIS

La formation s'adresse à des personnes ayant une base correcte en programmation en Python. Une mise à niveau en Python est proposée en amont de la formation afin de la rendre accessible au plus grand nombre : Initiation au langage Python

INFORMATIONS PRATIQUES

PLAGE HORAIRE DE LA FORMATION

Du 08 au 12 juillet 2024, de 09:00 à 17:30

LIEU

Maison des langues
Voie du Roman Pays 3A
1348 Louvain-la-Neuve

TARIFS ET INSCRIPTIONS

Jusqu'au 30 avril 2024 :

- Membre UCLouvain ou chercheur d'une autre Université : 500€
- Personne externe à l'UCLouvain : 1250€

À partir du 30 avril 2024 :

- Membre UCLouvain ou chercheur d'une autre Université : 750€
- Personne externe à l'UCLouvain : 1500€

L'inscription se fait directement **via le site du SMCS**

Les frais d'inscription sont dus dès l'inscription et seront facturés dans leur intégralité en cas d'annulation par le participant après le 30 avril 2024. Si l'annulation est demandée par écrit avant cette date, seule une retenue de 150€ pour frais administratifs sera opérée.

Le maintien de l'école d'été est conditionné par l'inscription d'un nombre suffisant de participants. Les participants seront tenus informés d'une éventuelle annulation au plus tard le 31 mai. En cas d'annulation, le montant payé pour l'inscription sera reversé dans son intégralité.

ACTIVITÉS SOCIALES

Le lundi 8 juillet 2024 : verre d'accueil à partir de 17h30
Le jeudi 11 juillet 2024 : activité à 17h30

CONFÉRENCIER INVITÉ

Didier Schwab
« Modèles génératifs »

POINTS CPD

Cette formation est reconnue par l'IABE et donne lieu à des points CPD

COMITÉ ORGANISATEUR

Plateforme de Support en Méthodologie et Calcul Statistique (**SMCS**)
Centre de traitement automatique du langage (**Cental**)

CONTACT

POUR TOUTE QUESTION, N'HÉSITEZ PAS À NOUS CONTACTER !

smcs-stat@uclouvain.be / +32(0)10/47.94.07

www.uclouvain.be/smcs

https://www.linkedin.com/company/smcs-uclouvain

LE SMCS

Le SMCS est une plateforme technologique de l'UCLouvain dont l'objectif est d'apporter une expertise et un accompagnement dans l'utilisation de **méthodes et logiciels de statistique**. Les services offerts incluent des formations, de la consultance personnalisée, de l'aide à la réalisation d'enquêtes et des collaborations dans des projets de recherche. Le SMCS délivre ses services à ses membres, mais également à l'extérieur de l'institution.

UCLouvain
Faculté des sciences

FORMATION CONTINUE

CENTAL centre de traitement automatique du langage

SMCS | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics