ÉCOLE D'ÉTÉ

Text mining et analyse de données médicales en Python

Troisième édition

Du 11 au 15 septembre 2023
pour l'école d'été

Les 7 et 8 septembre 2023
pour l'initiation à Python

www.uclouvain.be/smcs

Comité organisateur :

Plateforme de Support en Méthodologie et Calcul Statistique (SMCS), UCLouvain
Centre de traitement automatique du langage (Cental), UCLouvain
Ecole de Statistique, Biostatistique et Sciences Actuarielles (LSBA), UCLouvain

UCLouvain SMCS LIDAM LSBA CENTAL FORMATION CONTINUE

L'école d'été (troisième édition)

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. Chaque année, une école d'été est consacrée au *text mining*, c'est-à-dire à un ensemble de techniques qui reposent sur des algorithmes, des modèles statistiques et des ressources linguistiques permettant de traiter les données textuelles stockées sur un support informatique.

Suite au succès des deux premières éditions de l'École d'été portant sur le *text mining* en *Python*, consacrées respectivement à l'analyse de documents financiers et à celle de données de réseaux sociaux, le SMCS et le Centre de Traitement automatique du Langage (CENTAL) vous proposent une **nouvelle édition**, axée sur l'**analyse textuelle de données médicales**. En effet, la langue naturelle demeure le vecteur de diffusion privilégié des informations dans le monde médical (rapports médicaux, lettres de sortie d'hôpital, etc.). Les défis liés à l'exploitation informatique de ces données peu ou pas structurées rendent d'autant plus intéressantes les nombreuses applications dont elles peuvent faire l'objet (encodage des dossiers médicaux, pharmacovigilance, prédictions de risques, aide à la recherche clinique, etc.). Cette nouvelle formule peut s'envisager de manière **indépendante ou complémentaire aux éditions précédentes**, puisque la formation sera adaptée à des **problématiques spécifiques**, notamment à l'étiquetage et la normalisation de données, parfois bruitées, générées dans un contexte médical.

Le Text Mining

Le traitement automatique de la langue (TAL ; en anglais natural language processing – NLP) et plus particulièrement le *text mining* font aujourd'hui **partie intégrante de notre quotidien**. Le moteur de recherche de Google, l'application Siri d'Apple ou chatGPT d'openAI sont autant d'exemples illustrant cette omniprésence.
Pour chacun de ces programmes, la **structuration de l'information** est essentielle. Qu'il s'agisse d'extraire des dates ou des horaires dans des courriels pour planifier un évènement ou d'identifier les actions à effectuer à la suite d'une commande vocale, l'**analyse du contenu** est au cœur de ce type d'applications.

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.

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 des entreprises belges sont 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

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

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 et par le prix du meilleur article à la conférence TALN, en 2016.

Hubert Naets

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

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 s'intéresse maintenant à l'utilisation des réseaux neuronaux afin d'augmenter l'efficacité des outils d'analyse sur des données réelles.

Contenu de la formation

Cette école d'été a pour objectif d'initier les participant·es au *text mining* dans un cadre d'**analyse de données médicales** en les introduisant à différents usages possibles des techniques de TAL. Des techniques issues du *machine learning*, mais aussi des approches plus linguistiques seront abordées au travers d'**objectifs concrets**, tels que la détection et la normalisation d'entités médicales. Ces différentes tâches seront réalisées à l'aide du langage de programmation **Python**, qui est l'un des langages de programmation de référence en *data science*.

Le programme sera divisé en modules, qui mêleront à la fois théorie et pratique. L'ensemble offrira une chaîne de traitement complète en *text mining*.

• Modèles de langue et fine-tuning

Les modèles de langue encodent, de façon latente et probabiliste, un large ensemble de caractéristiques linguistiques (morphologie, relations syntaxiques et concepts sémantiques) en s'appuyant sur la distribution de mots dans un ensemble de documents. Ces modèles sont utilisés dans un nombre très important de tâches en traitement automatique des langues. Les approches neuronales génèrent et emploient de tels modèles (BERT, GPT...) qui servent, en *text mining*, par exemple à classifier des documents et des séquences de mots, ou à calculer la similarité entre des documents. Le/la participant·e sera initié·e à certains modèles de langues, à leur entraînement et à leur fine-tuning.

• Classification

Le/la participant·e sera confronté·e aux enjeux et défis des méthodes de classification pour la détection de caractéristiques dans les textes (ex. identification du domaine médical traité, diagnostic sur la base de textes, etc.). Il/elle pourra se rendre compte des difficultés liées à la définition et à l'annotation de classes, mais aussi à la création et à la manipulation d'un grand nombre de variables linguistiques. Le module présentera la méthodologie de conception et d'évaluation de modèles de classification spécifiques au TAL et au Deep learning.

• Étiquetage automatique

Le/la participant·e découvrira des techniques essentielles d'étiquetage automatique des textes, en particulier l'attribution, à chaque mot du texte, d'une partie du discours (ex. verbe, nom, adjectif, etc.) et l'identification d'entités nommées (ex. pathologies, résultats cliniques, dates, etc.) pour la recherche d'information.

• Normalisation

La multiplication des sources d'information médicale représente un défi pour les spécialistes de la santé. L'exploitation de ces données demeure en effet ardue, car une grande partie de l'information qui y est enregistrée l'est sous la forme de texte libre plus ou moins structuré. Dans ce contexte, ce module introduira les participant·es aux enjeux, défis et techniques de TAL mobilisables pour améliorer l'interopérabilité des bases de données médicales au sens large, et ainsi faciliter l'échange de connaissance entre différents services et/ou institutions de soin.

• Exploitation des données

Le/la participant·e sera sensibilisé·e à différents usages possibles des connaissances générées par les techniques vues au long des différents modules. À ce titre, il/elle découvrira différents outils et ressources utilisables dans une démarche d'analyse et d'exploration des données (représentation graphique, création de cohortes de patients...).

Prérequis

Les participants devront posséder de bonnes connaissances en programmation en général. Des connaissances de base en *Python* peuvent être utiles.

Une mise à niveau en *Python* est proposée en amont de l'école d'été (les 7 et 8 septembre 2023) afin de la rendre accessible au plus grand nombre. Les informations sont disponibles sur notre site.

Tarifs

> **Membre d'université :** 500€ avant le 15 juin 2023, 750€ à partir du 15 juin 2023.
> **Autre :** 1250€ avant le 15 juin 2023, 1500€ à partir du 15 juin 2023.

Inscription

Informations et inscription via notre site
Reconnue par l'IABE, l'école d'été donne lieu à des points CPD (32.5 points pour les 5 jours).

Activités sociales

> Le lundi 11 septembre 2023 : verre d'accueil à partir de 17h30
> Le jeudi 14 septembre 2023 : activité à 17h30

L'inscription à ces activités se fait lors de l'inscription à l'école d'été.

Conditions

Les frais d'inscription sont dus dès l'inscription et seront facturés dans leur intégralité en cas d'annulation par le participant après le 15 mai 2023. 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 30 juin 2023. En cas d'annulation, le montant payé pour l'inscription sera reversé dans son intégralité.

Lieu

Louvain-la-Neuve, Belgique / Auditoire Socrate 031-032
Place du Cardinal Mercier, n°10-12, 1348 Louvain-la-Neuve.

Accessibilité

Le campus de Louvain-la-Neuve se situe à 30 km de Bruxelles.

En train : Gare de Louvain-la-Neuve,
1348 Louvain-la-Neuve.

En bus : Gare d'autobus,
près du Parking Leclercq,
1348 Louvain-la-Neuve.

Contact

Pour toute question, n'hésitez pas à nous contacter !

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

UCLouvain SMCS LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics LSBA CENTAL Centre de traitement automatique du langage FORMATION CONTINUE