ÉCOLE D'ÉTÉ

Text mining et analyse des réseaux sociaux en Python

Du 11 au 15 juillet 2022
pour l'école d'été

Les 7 et 8 juillet 2022
pour l'initiation à Python

Louvain-la-Neuve, Belgique

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

Sponsorisé par Business & Decision

UCLouvain LSBA sMCs CENTAL LIDAM

L'école d'été

Le volume de données numériques générées par les activités humaines croît chaque année ; cela rend plus complexe, mais aussi plus intéressant l'usage de ces données dont les applications sont multiples (tels que le système de questions-réponses Watson développé par IBM ou encore les assistants personnels virtuels Siri d'Apple ou Alexa d'Amazon). 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 ces données. Cette 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 sous une forme numérique.

Suite au succès de la première édition (École d'été 2019) portant sur le *text mining* en Python, et consacrée à l'analyse de documents financiers, le SMCS et le Centre de traitement automatique du langage (CENTAL) vous proposent une **nouvelle édition**, axée sur l'**analyse textuelle** de productions issues des **réseaux sociaux**. Cette nouvelle formule peut s'envisager de manière **indépendante ou complémentaire de l'édition précédente**, puisque la formation sera adaptée aux **problématiques spécifiques** des réseaux sociaux, et notamment à l'acquisition et au traitement textuel des données, souvent bruitées, propres à ces réseaux.

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 l'assistant personnel intelligent Alexa d'Amazon 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 aux membres de l'UCLouvain 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.

Contenu de la formation

Cette école d'été a pour objectif d'initier les participant·es au *text mining* dans un cadre de **veille sur les réseaux sociaux**. Des **techniques** issues du *machine learning*, mais aussi des approches plus linguistiques seront abordées au travers d'**objectifs concrets**, tels que le profilage, l'extraction d'**opinions** et la **détection de polarités**. Ces différentes tâches seront réalisées à l'aide du langage de programmation **Python**, qui est 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.*

- Introduction générale et contexte
  - Découvrir la nature des données générées au travers de l'utilisation de différentes plateformes sociales, ainsi que plusieurs stratégies pour récupérer ces données.
  - Prendre conscience des usages possibles de l'analyse textuelle de productions issues des réseaux sociaux et découvrir ainsi différents outils et ressources utilisables dans une démarche d'analyse de réseaux sociaux.

- Collecte et prétraitement des données
  - Comprendre les grandes étapes de collecte, de nettoyage et d'exploitation de données.
  - Apprendre comment acquérir et préparer des données textuelles issues des réseaux sociaux.
  - Pouvoir normaliser de larges volumes de données, identifier les éléments de contenu pertinents et leur associer différents types de métadonnées afin d'obtenir un corpus de travail exploitable.

- Étiquetage automatique de données bruitées
  - Découvrir 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. noms d'organisation, de lieu, de personne, dates, etc.) pour la recherche d'information.
  - Prendre conscience de certains défis liés à l'analyse de données, comme la gestion des formes non normées et du bruit linguistique de façon plus générale, très fréquents dans les textes provenant des réseaux sociaux.

- Similarité textuelle
  - Apprendre comment mesurer la similarité entre documents à l'aide d'une représentation vectorielle (*vector space model*).
  - Découvrir des techniques plus avancées, qui incluent notamment des méthodes à base de réseaux de neurones (ex. Word2Vec) – ces techniques seront notamment utilisées à des fins de profilage ou de recommandation de contenu sur la base des messages envoyés par les utilisateurs et les utilisatrices.

- Classification
  - Se confronter aux enjeux et défis des méthodes de classification appliquées à des données textuelles pour l'analyse d'opinion et de sentiments ou pour l'identification de caractéristiques des utilisateurs et des utilisatrices.
  - Prendre conscience 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 spécifiques au TAL.

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, notamment au CENTAL et au Social Media Lab de l'UCLouvain. Ses intérêts de recherche combinent la recherche en TAL médical ainsi que l'extraction et l'analyse automatisée de productions textuelles issues de plateformes socionumériques.

Thomas François

Thomas François, professeur de linguistique appliquée à l'UCLouvain, travaille depuis plus de 12 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 vu récompenser par le prix de la thèse ATALA en 2012 et par le prix du meilleur article à la conférence TALN2016.

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.

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 7-8/07) afin de la rendre accessible au plus grand nombre. Les informations sont disponibles sur notre site : https://sites.uclouvain.be/training/smcs/view.php?id=436&l=fr

Tarifs

- Membre d'université : 500€
- Autre : 1250€

Inscription

INFORMATIONS via notre site :
https://sites.uclouvain.be/training/smcs/view.php?id=344&l=fr

Lien pour l'INSCRIPTION :
https://limesurvey.uclouvain.be/limesurvey319/index.php/174753

Reconnue par l'IABE, l'école d'été donne lieu à des points CPD.

Activités sociales

- Le lundi 11 juillet, verre d'accueil à partir de 17h30
- Le jeudi 14 juillet, activité à 17h30 puis repas à 19h

L'inscription à ces activités se fait lors de l'inscription à l'école d'été. Suivant l'évolution de la situation sanitaire, elles pourront être annulées ou modifiées. Quoi qu'il en soit, il vous sera demandé de confirmer votre participation d'ici quelques mois. Vous ne vous engagez donc pas encore définitivement à participer à ces activités mais en vous y inscrivant, vous nous permettez d'évaluer le nombre de participants pour chacune.

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 avril. Si l'annulation est demandée par écrit avant cette date, seule une retenue de 75€ 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 11 mai. En cas d'annulation, le montant payé pour l'inscription sera reversé dans son intégralité.

Lieu

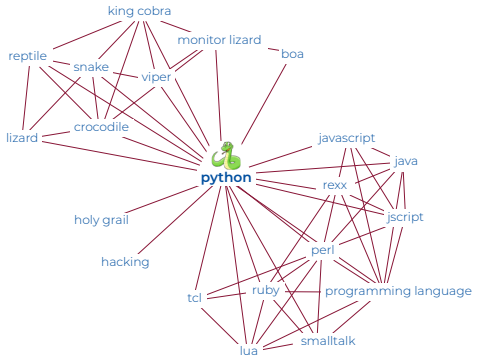
Louvain-la-Neuve, Belgique / Auditoire Socrate 031-032
Place du Cardinal Mercier, n°10-12, 1348 Louvain-la-Neuve.

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.



smcs-stat@uclouvain.be

Contact

Pour toute question, n'hésitez pas à nous contacter !
smcs-stat@uclouvain.be / +32(0)10/47.94.07

UCLouvain LSBA sMCs Plate-forme Technologique de Support en Méthodologie et Calcul Statistique CENTAL Centre de traitement automatique du langage LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics