Python pour le traitement du langage naturel (NLP)
Natural Language Processing avec les outils et bibliothèques Python

Cours Pratique de 3 jours - 21h
Réf : PTS - Prix 2024 : 1 860€ HT

Ce cours enseigne l'utilisation de Python pour le traitement du langage naturel : la préparation des données, la représentation de textes et leur modélisation. Le participant utilise des outils et bibliothèques Python pour effectuer des tâches courantes de NLP, met en œuvre et applique des modèles de NLP.

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Utiliser python pour traiter des données textuelles

Choisir les outils et bibliothèques Python nécessaires au traitement

Mettre en place les différentes étapes de préprocessing et de vectorisation

Utiliser les techniques appropriées en fonction des objectifs : classification / topic modelling / analyse de sentiment

Appliquer et évaluer des modèles sur des données réelles

LE PROGRAMME

dernière mise à jour : 03/2023

1) Environnement Python pour le NLP

- L'environnement de développement Python / Anaconda / Jupyter Notebook.
- Les principaux types de données : chaînes, booléennes, nombres, listes, tuples et dictionnaires.
- Les structures de contrôles : les boucles for et while, le test if/elif/else.
- Les fonctions : création, passage de paramètres, valeurs par défaut, arguments variables.
- Numpy : vecteurs, matrices, slicing, concaténation.
- Pandas : l'analyse de données tabulaires (CSV, Excel...), statistiques, pivots, jointures, filtres.

Travaux pratiques : Manipulation de Python dans un notebook Jupyter. Exercice de mise en pratique avec pandas et numpy.

2) Prétraitement des données textuelles

- Identifier ce que sont des données textuelles et présentation des librairies spaCy et nltk.
- Tokenisation des mots.
- Suppression des stop-words, de la ponctuation et des éléments non essentiels à l'analyse.
- Lemmatisation vs racinisation (stemming).

Travaux pratiques : Preprocessing sur des corpus de textes avec les 2 librairies, comparaison des résultats et des façons d'implémenter. Création de listes de stop-words, comparaison lemmatisation et de racinisation.

PRÉREQUIS

Connaissances de programmation en Python.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

3) Extraction d'informations

- Identification de la nature grammatical des mots à l'aide du Part Of Speech Tagging.
- Identifier des personnes, lieux etc avec le Named Entity Recognition.

Travaux pratiques : Mettre en place le Part Of Speech Tagging et le Named Entity Recognition. Analyse des résultats, filtres sur certaines catégories grammaticales, sur les noms propres.

4) Représentation vectorielle des données textuelles

- Bag of words.
- Pondération tf-idf.
- Approche avec des n-grams.
- Les embeddings : word2vec, gloVe, fastTesxt..

Travaux pratiques : Transformation d'un corpus de texte en utilisant différentes approches : bag of words, tf-idf, word2vec, gloVe.
Comparaison des vecteurs.

5) Machine learning sur des données textuelles

- Rappels sur les étapes de construction d'un modèle prédictif.
- Classification.
- Analyse de sentiment.
- Topic modelling.

Travaux pratiques : Modélisation en utilisant différents types de vecteurs (bag of words vs embeddings).
Analyse de sentiment sur des tweets.

6) Procédures d'évaluation de modèles

- Les techniques de rééchantillonnage en jeu d'apprentissage, de validation et de test.
- Test de représentativité des données d'apprentissage.
- Mesures de performance des modèles prédictifs.
- Matrice de confusion.

Travaux pratiques : Construire et évaluer un modèle NLP de façon appliquée..

LES DATES

CLASSE À DISTANCE	PARIS
2024 : 15 mai, 08 juil., 16 oct.	2024 : 01 juil., 09 oct.