Linguistique outillée et traitements statistiques avec R

Infos pratiques

> **ECTS** : 3.0

> **Nombre d'heures** : 24.0

> **Période de l'année** : Enseignement neuvième semestre

> **Méthodes d'enseignement** : En présence

> **Forme d'enseignement** : Cours magistral

> **Ouvert aux étudiants en échange** : Oui

> **Composante** : Philo, Info-Comm, Langages, Littératures & Arts du spectacle

Présentation

Loin de se réduire à des requêtes lancées au hasard sur de grandes bases de données textuelles, la linguistique de corpus se nourrit des théories linguistiques pour les étayer ou les remettre en cause empiriquement. Ce cours a pour objectif de rendre les étudiant·e·s autonomes en leur donnant les moyens de constituer leurs propres outils pour l'exploration des corpus et la quantification de leurs données dans un seul et même environnement de programmation : R.

Le cours se décompose en deux parties. Après une présentation des objectifs de la linguistique de corpus et une brève typologie des corpus, la première partie aborde successivement :

- les bases de la programmation en R,
- la manipulation des chaînes de caractères,
- l'élaboration d'outils d'exploration de corpus,
- la constitution de jeux de données tabulées,
- la quantification sommaire des données ainsi que leur visualisation.

La seconde partie est consacrée au traitement statistique des données linguistiques. Sont abordés les points suivants :

- les statistiques descriptives,
- les tests statistiques,
- les mesures d'association et les réseaux lexicaux,
- les méthodes dites de clustering et leurs visualisations,
- les modèles de sémantique distributionnelle (SVD, PPMI, word2vec, BERT).

Enseignant : DESAGULIER Guillaume (MCF-HDR Paris 8)

Objectifs

Ce cours a pour objectif de rendre les étudiant·e·s autonomes en leur donnant les moyens de constituer leurs propres outils pour l'exploration des corpus et la quantification de leurs données dans un seul et même environnement de programmation : R.

Évaluation

Evaluation :

2 cas de figure sont prévus dans les modalités de contrôle des connaissances et compétences : 1/ en 2 sessions ou 2/ en session unique. OPTER pour l'un ou l'autre. 2.En session unique (devoir récapitulatif après les cours)

M3C en 2 sessions

* Régime standard session 1 – avec évaluation continue (au moins 2 notes, partiel compris) : ……..

ou

Régime standard session 1 – avec évaluation terminale (1 seule note) : ……..

Un examen

* Régime dérogatoire session 1 : ……..

Un examen

* Session 2 dite de rattrapage : ……..

M3C en session unique

* Régime standard intégral – avec évaluation continue (au moins 2 notes) –

! ATTENTION : cette formule ne prévoit pas d'épreuve en session 2 mais une 2ème chance organisée sur la

période du semestre - elle ne peut être appliquée à des EC isolément mais doit concerner tte la formation – elle ne peut appliquée aux EC ETAB et aux formations qui ont des EC ETAB – son application n'est pas adaptée en Licence.

: 1 épreuve, 100% ; devoir écrit ; 10 jours

Pré-requis nécessaires

Bases en programmation, maîtrise des outils de bureautique.

Bases en mathématiques.

Compétences visées

Rendre les étudiant·e·s autonomes en leur donnant les moyens de constituer leurs propres outils pour l'exploration des corpus et la quantification de leurs données. Comprendre l'environnement R.

Bibliographie

Brezina, Vaclav (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Desagulier, Guillaume (2017). *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Springer.

Winter, Bodo (2019). *Statistics for Linguists: An Introduction Using R*. Routledge

Ressources pédagogiques

Notebook en ligne

Contact(s)

> Guillaume Desagulier
Responsable pédagogique
gdesagul@parisnanterre.fr