Projet IDEX Formations Innovantes INFOSHS
Statistiques Lexicales

Présentation	Cette formation donnera des bases pour travailler sur des tâches d'extraction d'information à partir de corpus textuels. Les tâches visés sont la categorisation automatique des textes, le clustering, la découverte de tendances dans des corpus annotés avec des marques temporelles. La formation alternera cours et travaux pratiques.
Mode de validation	Une attestation de présence sera délivrée à chaque étudiant ayant suivi l'ensemble de la formation. Un certificat de réussite sera délivré aux étudiants ayant passé avec succès le QCM et l'épreuve de validation.
Programme détaillé	• Fréquence des mots dans les corpus, loi de Zipf • Extraction de bigrammes, trigrammes • Comparaison de distributions de fréquence • Information Mutuelle • Perplexité des corpus (utilisation du logiciel SRILM) • Détection de fréquences "anomales" • Comment faire un cloud tag (nuage de tags)
Objectifs pédagogiques	Être capable d'extraire mots, n-grammes à partir d'un texte et les utiliser pour découvrir des informations additionnelles qui ne sont pas forcement explicites dans le texte, aussi par comparaison avec des autres textes.
Formateur(s)	**Davide Buscaldi :** maître de conférences à l'Université Paris 13, auteur de plus de 70 articles scientifiques publiés dans des conférences et journaux internationaux, avec plus de 10 ans d'experience dans la recherche en TAL. Ses domains d'interêt sont la recherche d'information sémantique, les ontologies et la fouille de textes.
Pré-requis	Connaissance du langage Python et de la librairie NLTK. Notions fondamentales de probabilité et statistique.
Public visé	Étudiants en Master 1, 2. Doctorants de 1ère année - Disciplines : informatique, langues, linguistique, littérature. S'adresse plus spécifiquement à des étudiants concernés par l'utilisation de corpus textuels.
Effectif	maximum 24 personnes

Informations pratiques

Email (contact pédagogique, pour toutes questions sur le contenu de la formation)	Email de contact : davide.buscaldi@lipn.univ-paris13.fr
Date(s)	Lundi 20 juin : 9h-12h et 14h-17h Mardi 21 juin : 9h-12h et 14h-17h
Lieu	IUT de Villetaneuse, salles informatiques SHS ; **se munir du présent document pour contrôle à l'entrée du bâtiment.**
Durée de la formation	2 jours (12h)
Nombre de jours validés par le Cfdip	
Lien vers l'inscription (obligatoire)	http://goo.gl/forms/hBxhXfWVul