SCI6134 – Outils linguistiques et gestion documentaire

Hiver 2018 – Syllabus

Professeure

Lyne Da Sylva
Bureau C-2030, Pavillon Lionel-Groulx
514-343-6444
Lyne.Da.Sylva@UMontreal.CA
http://dasylva.ebsi.umontreal.ca

Horaire du cours : mercredi 8h30-11h30
Local du cours : C-2043 (Pavillon Lionel-Groulx)
Site web du cours : sur Studium (www.studium.umontreal.ca)
Heures de disponibilité : mercredi 15h-17h et sur rendez-vous

Description du cours (annuaire)

Application d'outils linguistiques à la chaîne documentaire pour la gestion (semi-)automatique de textes. Principes et fonctionnement des outils. Constitution de corpus. Applications à divers types d'analyses documentaires.

Objectifs du cours

Objectifs généraux

- faire connaître les contextes d'applications des traitements linguistiques automatiques à l'intérieur de la chaîne documentaire et les types de logiciels utiles au traitement
- familiariser les étudiants avec la constitution et la structuration de corpus numériques en vue d'exploitations automatiques ou assistées par ordinateur
- amener les étudiants à prendre conscience de l'état d'avancement variable des technologies linguistiques disponibles

Objectifs spécifiques

À la fin de ce cours, l'étudiant(e) sera en mesure :

- d'utiliser quelques logiciels pour effectuer des traitements linguistiques ou statistiques sur des documents textuels
- d'analyser diverses statistiques des documents traités et évaluer leur pertinence pour l'analyse documentaire
- de comparer la performance d'outils similaires
- de constituer un corpus de documents textuels conformes à une thématique ou à un objectif précis
- de situer les technologies linguistiques disponibles et leur lieu d'application dans la chaîne documentaire

Contenu du cours

Ce cours vise à initier les étudiants aux technologies linguistiques qui peuvent être utiles à l'analyse documentaire automatique ou assistée par ordinateur.

Le cours se présente en deux temps : il débute par une introduction aux concepts de base en linguistique et en linguistique informatique ainsi qu'une présentation des outils de base qui peuvent mis être à contribution. Ensuite, l'application de ces éléments sera illustrée dans divers contextes de gestion docu-

mentaire : en priorité l'indexation automatique, le résumé automatique et la classification automatique, mais aussi l'acquisition de documents, la diffusion d'information, la constitution de thésaurus, la veille informationnelle, etc.

Les étudiants effectueront diverses expérimentations pour comprendre l'utilité, les caractéristiques, la portée et les limites de ces outils logiciels.

Les étudiants seront ainsi amenés à évaluer la contribution potentielle des diverses technologies aux tâches d'indexation, de condensation et de classification documentaires.

Méthodes pédagogiques

- cours magistraux et conférenciers invités
- petits travaux de groupes
- travaux pratiques
- démonstrations de logiciels
- lectures obligatoires

Évaluation

Travaux pratiques

Les travaux pratiques se font seuls ou en équipe de deux personnes. Les séances de travail supervisées ne sont pas nécessairement suffisantes pour compléter les travaux; il sera parfois nécessaire de revenir au laboratoire sur une base individuelle (en réservant un poste de travail à l'avance). Il faut également prévoir du temps pour préparer le rapport.
Les rapports doivent être remis à la professeure au plus tard au début du cours, le jour de l'échéance.
Pour les travaux réalisés en équipe, la professeure se réserve le droit d'évaluer séparément chaque membre d'une équipe.

Cinq travaux pratiques (60%)	Remise	Pondération
- TP1 : extraction de mots et fréquences d'un ensemble de textes	24 janvier	10%
- TP2 : extraction de termes d'un ensemble de textes	31 janvier	10%
- TP3 : segmentation manuelle et automatique de textes	7 février	10%
- TP4 : constitution d'un corpus de textes	28 février	15%
- TP5 : indexation d'un corpus de textes	21 mars	15%

Travail de recherche

Un travail final (individuel) (40%) : une recherche portant sur un scénario de gestion documentaire à l'aide d'outils automatiques (avec analyse critique).

-	Présentation orale	18 avril	10%
-	Rapport écrit	25 avril	30%

Calendrier provisoire en date du 19 décembre 2017

Date	Thème	*Dimension linguistique/TAL*	*Travaux*
10/01	Présentation du cours. Les applications du TAL dans la chaîne documentaire.	Introduction au *TAL* et à la *linguistique*.	
17/01	Défis du traitement intellectuel des documents par des moyens automatiques	a) identifier les mots, ou *Le lexique* : reconnaître, compter b) identifier les variantes des mots, ou *La morphologie*	*Présentation TP1*
24/01		c) identifier le sens des mots, ou *La sémantique lexicale* d) identifier les multitermes, ou *La terminologie*	*Remise TP1* *Présentation TP2*
31/01		e) identifier les phrases, ou *La segmentation en phrases* f) identifier les unités thématiques, ou *La segmentation en unités thématiques*	*Remise TP2* *Présentation TP3*
7/02	La création, la représentation et le stockage des documents	Regrouper des textes numériques, ou *Les corpus* et *La linguistique de corpus*	*Remise TP3* *Présentation TP4*
14/02	L'analyse documentaire I – introduction	a) comprendre le sens d'un texte, ou *La linguistique textuelle*	
21/02	L'analyse documentaire II – l'indexation	b) extraire un certain nombre de termes d'indexation, ou *Les propriétés statistiques du lexique*	
28/02	L'analyse documentaire III – la condensation	c) extraire des phrases pour produire un résumé, ou *La distribution des phrases dans un texte* d) comprendre un texte pour générer un résumé, ou *Les modèles de compréhension*	*Remise TP4* *Présentation TP5*
7/03	*Semaine de lecture*	*Pas de cours*	
14/03	L'acquisition des documents Conférencier invité : Philippe Langlais, DIRO (UdeM)	a) numériser un document papier, ou *La reconnaissance optique des caractères (ROC)* b) donner accès à des documents dans d'autres langues, ou *La traduction automatique*	
21/03	Le repérage des documents	a) utilisation de thésaurus, ou *L'expansion de requêtes* b) recherche translinguistique (CLIR), ou *La traduction*	*Remise TP5*
28/03	La classification des documents. La sélection des documents. La diffusion des documents.	a) identifier la(les) thématique(s) de base d'un document, ou *La catégorisation* b) comparer cette thématique à celle d'un plan de classification, ou *La classification* c) comparer des documents entre eux, ou *Le clustering* Liens avec la classification automatique	
4/04	Les ressources documentaires – la constitution de thésaurus, ontologies, etc. Conférencière invitée : Magali Lachapelle, SRC	a) identifier et structurer un vocabulaire contrôlé, ou *L'extraction de terminologie* b) rendre des vocabulaires contrôlés interopérables, ou *L'alignement d'ontologies*	
11/04	Atelier pour le travail final		*Atelier de travail*
18/04	Présentations orales		*Présentation orale*
25/04	Remise des travaux		*Remise des travaux*

La matière et sa répartition entre les cours sont sujettes à changement en fonction de la vitesse de progression et de la disponibilité des conférenciers.

Politiques

Délais et dates de remise des travaux

Les retards seront traités conformément à la politique de l'EBSI (voir le Guide de l'étudiant).

Règlement disciplinaire sur le plagiat ou sur la fraude concernant les étudiants

Il est attendu que tous les étudiants inscrits au cours respectent le code d'honneur de l'EBSI (http://www.ebsi.umontreal.ca/sout/code-honneur.html). Le plagiat à l'Université de Montréal est sanctionné par le Règlement disciplinaire sur la fraude et le plagiat concernant les étudiants. Pour plus de renseignements, consultez le site www.integrite.umontreal.ca.

Qualité de la langue

La professeure tiendra compte de la qualité du français dans l'évaluation des travaux et peut enlever jusqu'à 10 % de la note (voir Guide de l'étudiant).

Afin de minimiser les problèmes dus à la qualité de la langue, on conseille l'utilisation d'un logiciel de correction grammaticale et orthographique comme Antidote, installé sur les postes des laboratoires informatiques.

Mode de communication

Le mode de communication privilégié entre la professeure et les étudiant(e)s est le courriel. Veuillez vous assurer que vous êtes officiellement inscrit(e) au cours et maintenez à jour l'adresse de courriel enregistrée dans votre profil informatique à l'Université de Montréal. Vous devez lire votre courriel très régulièrement (au moins une fois par jour), des informations importantes concernant le cours ou les TP pouvant être diffusées par ce moyen.

Très important : pour toute correspondance concernant le cours, veuillez inscrire obligatoirement au début du champ sujet du message la chaîne suivante : [SCI6134] (incluant les crochets).

Mode d'évaluation

L'évaluation des travaux se fait selon le barème présenté dans le Guide de l'étudiant, avec les précisions suivantes :

Lettre	Signification	Points	Critères d'évaluation
A+	Exceptionnel	4,3	La note A+ est réservée aux travaux exceptionnels qui dépassent les exigences demandées.
A A-	Excellent Très bon	4,0 3,7	Travaux excellents ou très bons qui répondent exactement aux exigences demandées.
B+ B B-	Bon	3,3 3,0 2,7	Travaux qui répondent aux exigences demandées, avec certaines erreurs mineures ou relativement mineures.
C+ C	Passable	2,3 2,0	Travaux qui ne rencontrent qu'à moitié les exigences demandées ou qui comportent quelques erreurs importantes.
C- D+ D	Échec	1,7 1,3 1,0	Travaux qui ne rencontrent que partiellement les exigences demandées ou qui comportent des erreurs graves.
E F		0,5 0,0	La note E est attribuée aux travaux qui ne répondent pas aux exigences demandées. La note F est attribuée lorsqu'un travail ou un examen n'est pas remis ou lorsqu'un travail est remis après la date d'échéance fixée par le professeur, ou dans un cas de plagiat, copiage ou fraude.

Ressources

Notes de cours

Les notes de cours seront mises en ligne sur le site web du cours au fur et à mesure de l'avancement de la session. Il est à noter que les notes de cours sont un support à ce qui est présenté en classe et ne suffisent pas, à elles seules, pour comprendre la matière couverte.

Logiciels

Les logiciels utilisés dans le cadre du cours incluent un logiciel, Indexo, qui est installé sur les postes des laboratoires d'informatique de l'EBSI (local C-2027, C-2035 et C-2043 du pavillon Lionel Groulx) et dont l'accès est restreint aux étudiants du cours.

Si des versions gratuites ou de démonstration des logiciels sont disponibles pour réaliser les travaux en dehors du laboratoire, les étudiants en seront avisés.

Autres

- textes et corpus numériques
- manuels d'utilisation et/ou de référence des logiciels étudiés

Bibliographie générale

Bouillon, Pierrette et al. 1998. *Traitement automatique des langues naturelles,* Paris ; Louvain-la-Neuve : Duculot.

Clark, a; Fox, C.; Lappin, S. 2010. *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley-Blackwell, Oxford, UK.

Farzindar, Atefeh; Inkpen, Diana. 2015. *Natural language processing for social media*. San Rafael, California : Morgan & Claypool.

Jurafsky, Daniel; Martin, James H. 2009.*Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, New Jersey : Pearson Prentice Hall.

Lallich-Boidin, Geneviève; Maret, Dominique. 2005. *Recherche d'information et traitement de la langue. Fondements linguistiques et applications*. Villeurbanne : Presses de l'enssib.

Mitkov, Ruslan. 2005. *The Oxford Handbook of Computational Linguistics*, Oxford : Oxford University Press.