Université de Sherbrooke

Département d'informatique

IFT 714 – Traitement automatique des langues naturelles

Plan d'activité pédagogique

Hiver 2024

Enseignant

Amine Trabelsi	
Courriel :	Amine.Trabelsi@USherbrooke.ca
Local :	
Téléphone :	
Disponibilités :	À spécifier au début de la session

Responsable(s) : Amine Trabelsi

Site web du cours : https://moodle.usherbrooke.ca

Horaire

Exposé magistral :	Lundi	13h30 à 14h20	salle D4-2024
	Mercredi	13h30 à 15h20	salle D3-2036

Description officielle de l'activité pédagogique[1]

Cibles de formation :	Connaître les fondements du traitement automatique des langues naturelles (TALN). Comprendre comment manipuler des données en TALN. Comprendre et appliquer des modèles de langage. Comprendre et appliquer des modèles de classification et d'étiquetage de documents textes. Comprendre et appliquer des modèles de traduction automatique et d'analyse grammaticale. Comprendre les fondements de la reconnaissance vocale.
Contenu :	Manipulation de données langagières. Expressions régulières. Distance d'édition. Modèle de langage N-gramme et techniques de lissage. Classification de documents avec modèle de Bayes naïf. Étiquetage de documents avec modèle de Markov caché. Traduction automatique : manipulation de corpus bilingues, évaluation de systèmes de traduction, modèles IBM et phrase-based. Analyse grammaticale : grammaire hors contexte, grammaire hors contexte probabiliste, grammaire lexicalisée. Concepts de base et technologies de la reconnaissance vocale.
Crédits	3
Organisation	3 heures d'exposé magistral par semaine 6 heures de travail personnel par semaine
Particularités	Aucune

[1] https://www.usherbrooke.ca/admission/fiches-cours/ift714

1 Présentation

Cette section présente les cibles de formation spécifiques et le contenu détaillé de l'activité pédagogique. Cette section, non modifiable sans l'approbation du comité de programme du Département d'informatique, constitue la version officielle.

1.1 Mise en contexte

Le traitement automatique du langage naturel (TALN), plus connu sous l'acronyme NLP pour « Natural Language Processing », est l'un des domaines les plus pertinents et les plus utiles de l'intelligence artificielle. Il se situe à l'intersection de la linguistique computationnelle et l'apprentissage automatique. En outre Il est en rapport avec des domaines tels que les sciences cognitives, la théorie de l'information, la science des données, les sciences politiques, la psychologie et l'éducation. Le traitement automatique du langage naturel est défini comme étant l'automatisation de l'analyse, de la génération et de l'acquisition du langage humain ("naturel") (mandarin, hindi, espagnol, arabe, anglais, … Inuktitut). Son utilisation massive est due à l'omniprésence du langage et du texte comme principal support de communication humaine, par exemple dans les courriers électroniques, les blogs, les médias sociaux, la recherche sur le Web, les agents conversationnels ou "chatbots", les rapports médicaux, la traduction Web, etc. Il est à noter que plusieurs facteurs ont contribué à l'essor du TALN. On peut citer entre autres la disponibilité d'une plus grande puissance de calcul, le déploiement du web, puis du web social, les avancées de l'apprentissage automatique, et les progrès réalisés dans la compréhension du langage dans un contexte social.

1.2 Cibles de formation spécifiques

À la fin de cette activité pédagogique, l'étudiante ou l'étudiant sera capable de :

1. maîtriser les concepts fondamentaux du traitement automatique du langage naturel (TALN) ;
2. maîtriser les différents composants du TALN (analyse, compréhension et génération de texte) ;
3. maîtriser les modèles et les algorithmes de classification de texte binaire et multinomiale ;
4. maîtriser les modèles de langue ;
5. maîtriser les concepts de la sémantique vectorielle et la construction des modèles de plongement lexical ou l'embedding de texte en représentations éparses et denses ;
6. maîtriser les méthodes et les algorithmes sur le traitement de l'information linguistique ;
7. maîtriser les méthodes et les algorithmes sur le traitement syntaxique et sémantique ;
8. comprendre les modèles basés sur les réseaux neuronaux pour le langage naturel.

1.3 Contenu détaillé

Thème	Contenu	Nbr. d'heures	Objectifs	Travaux	Lectures
1	Introduction aux concepts de base et pratiques du traitement automatique du langage naturel (TALN) et à la classification de texte : Vocation ; domaines connexes, concepts de base et différents composants ; bref historique ; difficultés du **TALN** (Ambiguïté : généralisation, "sparsité", variation, expressivité) ; complexité de la représentation linguistique ; analyse lexicale ; sémantique ; systèmes de dialogue ; classsifieurs de sentiments et de toxicité et leurs méfaits	3	1& 2		Chap.1 de [1]
2	Modèles de base pour la classification supervisée des données textuelles I : Classifieur de Bayes naïf (BN) simple et multinomial **;** représentation et principes d'apprentissage supervisé (règle de Bayes et maximum de vraisemblance) ; problèmes computationnels et méthodes de lissage des paramètres ; cas pratique et classification des sentiments avec ou sans lissage ; métriques individuelles et combinées.	5	1, 2 & 3	✔	Chap.4 [1]
3	Modèles de base pour la classification supervisée des données textuelles II : Régression logistique binaire et multinomiale ; représentation et principes d'apprentissage ; composants du classifieur (représentation des attributs ; fonction de classification probabiliste ; fonction objective "fonction de perte entropie-croisée") ; algorithme de la descente du gradient stochastique ; apprentissage en "batch" et "mini-batch" ; test et validation ; problèmes de généralisation ("overfitting") et solutions proposées (régularisation et).	3	1 & 3	✔	Chap.5 [1]
4	Modèles de base pour la classification supervisée des données textuelles III : Réseaux de neurones : Représentation (architecture, input, output, unités neuronales, fonctions d'activation) ; réseaux "feedforward" ou perceptron multicouches (MLP) ; représentation de la régression logistique multinomiale par un réseau de neurones monocouche ; réseau bicouche avec sortie scalaire ou vectorielle ; principes d'apprentissage d'un réseau ; graphes computationnels et mise à jour des paramètres par rétro-propagation des erreurs.	4	1, 3 & 8	✔	Chap.6 [1]
5	Modèles de Langue : : Modèles N-grams (rôle et domaines d'applications) ; évaluation extrinsèque et intrinsèque (stratégies et mode de calcul de la métrique de perplexité), visualisation de bi-grams ; aspects pratiques et calcul des probabilités des N-grams ; "overfitting" et inférence directe et par lissage (simple et composée), "backoff" et interpolation ; sélection des paramètres de lissage par validation croisée N-fois.	5	1, 2 & 4	✔	Chap.3 [1]

6	"Embeddings" I : Rappel linguistique sur la signification des mots (synonymie, contraste sémantique, similarité, filiation des mots, champs sémantiques, antonymie, connotation affective et évaluation du sentiment) ; sémantique vectorielle ou "embbeddings" (origine, premiers modèles) ; différents types d'embbeddings ; représentation d'un corpus de documents par une matrice "terme-document" ; visualisation vectorielle du corpus ; matrice "terme-contexte" représentant la cooccurrence des mots dans un contexte, les métriques de similarité : le tf-idf ("term frequency – inverse document frequency") ; calcul de l'information ponctuelle mutuelle (IPM) et l'information ponctuelle positive mutuelle (et IPPM) sur une matrice terme-contexte ; pondérations de IPM et IMPP pour atténuer les effets des mots rares.	3	1 & 5	✔	Chap.6 [1]
7	"Embeddings" II : Avantage des vecteurs denses par rapport aux vecteurs épars ; Word2vec ("skip-gram", "CBOW") : algorithmes de plongement statique pour générer des vecteurs courts et denses ; apprentissage autosupervisé du classifieur "skip-gram" avec sondage négatif (SGSN) ; propriétés sémantiques (similarité analogique et relationnelle) des "embeddings" ; utilité des "embeddings" pour étudier le mode de changement de sens dans le temps et le biais culturel.	3	1 & 5	✔	Chap.6 [1]
8	Modèles de Langue Neuronaux (MLN) : Comparaison des modèles MLN / modèles N-grams ; le modèle de langue simple "Feedforward" ; préentraînement des "embeddings" ; entrainement simultané des "embeddings" et du réseau "Feedforward" ; stratégies d'évaluation et métriques utilisées pour juger du niveau de la performance des modèles crées.	3	1,3,4,5 & 8		Chap.7, 9 [1]
9	Traitement neuronal des séquences : Structures neuronales d'apprentissage profond qui traitent directement de la nature séquentielle et spatiale (contextuelle) du langage : réseaux neuronaux récurrents (RNN). Étude des RNNs en guise de modèles de langage, RNNs pour la génération de texte, RNNs pour l'étiquetage de séquences, RNNs pour la classification, RNNs empilés/RNN bidirectionnels pour la classification des séquences, RNNs à mémoire longue et courte (LSTM), et unités récurrentes à accès limité "gated recurrent units" (GRUs) pour gérer les effets de contexte, palier aux problèmes de "vanishing gradient" dans les RNNs classiques et mieux capturer les dépendances pour les séquences relativement distantes.	3	1,2,4,6,7 & 8	✔	Chap. 9, 10 [1]

10	Modèles encodeurs-décodeurs (seq-to-seq) et traitement des séquences : Modèles encodeurs-décodeurs avec RNNs (architectures et formulation mathématique). Entraînement du modèle encodeur-décodeur ; mécanisme d'attention (description, propriétés et formulation mathématique) ; problèmes des architectures basées sur les RNNs ; transformateurs pour le traitement des séquences ; mécanisme d'autoattention simple ; mécanismes d'autoattention à plusieurs niveaux ("Multihead") ; bloc de transformateurs ; encodage positionnel, les transformateurs comme modèles de langues autorégressifs ; génération et synthétisation contextuelles ; encodeur-décodeur avec transformateurs.	4	1,2,3,4,5,6,7 & 8		chap. 9, 10, 11 [1]

1. Le cours doit comprendre au moins trois travaux pratiques couvrant tous les sujets marqués «✔» dans le tableau.
2. Les lectures indiquées ne sont là qu'à titre indicatif. L'enseignant est libre de choisir un autre document de référence.

2 Organisation

Cette section propre à l'approche pédagogique de chaque enseignante ou enseignant présente la méthode pédagogique, le calendrier, le barème et la procédure d'évaluation ainsi que l'échéancier des travaux. Cette section doit être cohérente avec le contenu de la section précédente.

2.1 Méthode pédagogique

Le cours comprend trois heures de cours magistraux. Si le temps le permet, des sessions dédiées à la présentation des projets seront organisées en fin de session.

2.2 Calendrier

Semaine	Date	Thème
1	2024-01-08	1
2	2024-01-15	2
3	2024-01-22	3
4	2024-01-29	4
5	2024-02-05	4
6	2024-02-12	5
7	2024-02-19	6
8	2024-02-26	Examen périodique
9	2024-03-04	Relâche
10	2024-03-11	7
11	2024-03-18	8
12	2024-03-25	9
13	2024-04-01	10
14	2024-04-08	10
15	2024-04-15	Projet
16	2024-04-22	Examen final

Le rythme réel du cours peut varier en fonction des connaissances préalables et des intérêts des étudiantes et des étudiants. La personne enseignante peut exclure certains sujets ou inclure plus de sujets au cours de la session.

2.3 Évaluation

Devoirs (un minimum de 4 devoirs)	55 %
Projet (4 évaluations/livrables)	42 %
Participation	3 %

2.3.1 Qualité de la langue et de la présentation

Conformément à l'article 17 du règlement facultaire d'évaluation des apprentissages[2] l'enseignante ou l'enseignant peut retourner à l'étudiante ou à l'étudiant tout travail non conforme aux exigences quant à la qualité de la langue et aux normes de présentation.

2.3.2 Plagiat

Le plagiat consiste à utiliser des résultats obtenus par d'autres personnes afin de les faire passer pour sien et dans le dessein de tromper l'enseignante ou l'enseignant. Vous trouverez en annexe un document d'information relatif à l'intégrité intellectuelle qui

[2] https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/documents/Etudiants_actuels/Etudiants_actuels/Informations_academiques_et_reglements/2017-10-27_Reglement_facultaire_-_evaluation_des_apprentissages.pdf

fait état de l'article 9.4.1 du Règlement des études[3]. Lors de la correction de tout travail individuel ou de groupe une attention spéciale sera portée au plagiat. Si une preuve de plagiat est attestée, elle sera traitée en conformité, entre autres, avec l'article 9.4.1 du Règlement des études de l'Université de Sherbrooke. L'étudiante ou l'étudiant peut s'exposer à de graves sanctions qui peuvent être soit l'attribution de la note E ou de la note zéro (0) pour un travail, un examen ou une activité évaluée, soit de reprendre un travail, un examen ou une activité pédagogique. Tout travail suspecté de plagiat sera transmis au Secrétaire de la Faculté des sciences. Ceci n'indique pas que vous n'ayez pas le droit de coopérer entre deux équipes, tant que la rédaction finale des documents et la création du programme restent le fait de votre équipe. En cas de doute de plagiat, l'enseignante ou l'enseignant peut demander à l'équipe d'expliquer les notions ou le fonctionnement du code qu'elle ou qu'il considère comme étant plagié. En cas d'incertitude, ne pas hésiter à demander conseil et assistance à l'enseignante ou l'enseignant afin d'éviter toute situation délicate par la suite.

2.4 Échéancier des travaux

- Toutes les dates limites pour les livrables sont fixées à 23 h 59 heure locale. Un livrable remit 24 h après la date limite sera pénalisé de 3 %. Une soumission tardive de 2 jours (24 à 48 heures après la date limite) sera pénalisée de 10 %, et une soumission tardive de 3 jours sera pénalisée de 20 %. Les week-ends et les jours fériés sont également comptés comme des jours de retard. La soumission d'un livrable 72 heures (3 jours) après la date limite ne sera pas acceptée.

2.5 Utilisation d'appareils électroniques et du courriel

Selon le règlement complémentaire des études, section 4.2.3 [4], l'utilisation d'ordinateurs, de cellulaires ou de tablettes pendant une prestation est interdite à condition que leur usage soit explicitement permise dans le plan de cours.

Dans ce cours, l'usage de téléphones cellulaires, de tablettes ou d'ordinateurs est autorisées. Cette permission peut être retirée en tout temps si leur usage entraîne des abus.

Tel qu'indiqué dans le règlement universitaire des études, section 4.2.3[5], toute utilisation d'appareils de captation de la voix ou de l'image exige la permission de la personne enseignante.

Note : Je réponds aux questions posées par courriel à l'extérieur des périodes de cours.

Je vérifie et réponds aux e-mails pendant mes heures de travail du lundi au vendredi, de 8 h 30 à 16 h 30. Je ne verrai pas ou ne répondrai pas régulièrement aux e-mails en dehors de ces heures.

3 Matériel nécessaire pour l'activité pédagogique

4 Références

[1] JURAFSKY AND MARTIN : Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3.

[3] https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

[4] https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/documents/Etudiants_actuels/Etudiants_actuels/Informations_academiques_et_reglements/Sciences_Reglement_complementaire.pdf

[5] https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

UNIVERSITÉ DE SHERBROOKE

L'intégrité intellectuelle passe, notamment, par la reconnaissance des sources utilisées.
À l'Université de Sherbrooke, on y veille!

Extrait du *Règlement des études* (Règlement 2575-009)

9.4.1 DÉLITS RELATIFS AUX ÉTUDES

Un délit relatif aux études désigne tout acte trompeur ou toute tentative de commettre un tel acte, quant au rendement scolaire ou une exigence relative à une activité pédagogique, à un programme ou à un parcours libre.

Sont notamment considérés comme un délit relatif aux études les faits suivants :

a) commettre un plagiat, soit faire passer ou tenter de faire passer pour sien, dans une production évaluée, le travail d'une autre personne ou des passages ou des idées tirés de l'œuvre d'autrui (ce qui inclut notamment le fait de ne pas indiquer la source d'une production, d'un passage ou d'une idée tirée de l'œuvre d'autrui);

b) commettre un autoplagiat, soit soumettre, sans autorisation préalable, une même production, en tout ou en partie, à plus d'une activité pédagogique ou dans une même activité pédagogique (notamment en cas de reprise);

c) usurper l'identité d'une autre personne ou procéder à une substitution de personne lors d'une production évaluée ou de toute autre prestation obligatoire;

d) fournir ou obtenir toute aide non autorisée, qu'elle soit collective ou individuelle, pour une production faisant l'objet d'une évaluation;

e) obtenir par vol ou toute autre manœuvre frauduleuse, posséder ou utiliser du matériel de toute forme (incluant le numérique) non autorisé avant ou pendant une production faisant l'objet d'une évaluation;

f) copier, contrefaire ou falsifier un document pour l'évaluation d'une activité pédagogique;

[…]

Par plagiat, on entend notamment :

- Copier intégralement une phrase ou un passage d'un livre, d'un article de journal ou de revue, d'une page Web ou de tout autre document en omettant d'en mentionner la source ou de le mettre entre guillemets;
- reproduire des présentations, des dessins, des photographies, des graphiques, des données… sans en préciser la provenance et, dans certains cas, sans en avoir obtenu la permission de reproduire;
- utiliser, en tout ou en partie, du matériel sonore, graphique ou visuel, des pages Internet, du code de programme informatique ou des éléments de logiciel, des données ou résultats d'expérimentation ou toute autre information en provenance d'autrui en le faisant passer pour sien ou sans en citer les sources;
- résumer ou paraphraser l'idée d'un auteur sans en indiquer la source;
- traduire en partie ou en totalité un texte en omettant d'en mentionner la source ou de le mettre entre guillemets ;
- utiliser le travail d'un autre et le présenter comme sien (et ce, même si cette personne a donné son accord);
- acheter un travail sur le Web ou ailleurs et le faire passer pour sien;
- utiliser sans autorisation le même travail pour deux activités différentes (autoplagiat).

Autrement dit : mentionnez vos sources