UNIVERSITÉ DE SHERBROOKE

IFT 607
Traitement automatique des langues naturelles

Plan de cours
Automne 2014

Enseignant

Hugo Larochelle
Courriel : Hugo.Larochelle@USherbrooke.ca
Local : D4-1024-1
Site Web: http://www.dmi.usherb.ca/~larocheh/cours/ift607.html

Horaire

Exposé magistral :	Mercredi	10h30 à 11h20	salle D4-2021 (sauf le 28 août)
	Jeudi	8h30 à 10h20	salle D4-2021

Description officielle de l'activité pédagogique[1]

Objectifs — Connaitre 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.

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.

Crédits — 3

Organisation — 3 heures d'exposé magistral par semaine
6 heures de travail personnel par semaine

Préalable — IFT 615

[1] http://www.usherbrooke.ca/fiches-cours/ift607

1 Présentation

1.1 Mise en contexte

Les recherches en traitement automatique des langues naturelles (TALN) est un domaine de l'intelligence artificielle visant le développement de techniques automatisées pour la manipulation de données langagière, sous une forme textuelle ou sonore. L'objectif ultime est de donner à une machine la capacité de comprendre le langage humain, écrit ou parlé. Bien que cet objectif soit encore lointain, les applications immédiates de ces techniques incluent le développement d'interfaces textuelles plus naturelles, la traduction automatique de documents, la détection de pourriels, la recherche d'information dans une collection de documents à partir de requêtes, les systèmes de questions-réponses, et plusieurs autres. Des progrès substantiels ont tout de même été accomplis dans les dernières années (système questions-réponses Watson d'IBM, système de reconnaissance vocale Siri d'Apple, etc.) laissant croire que les technologies de TALN seront de plus en plus présentes dans le développement de systèmes informatisés.

Le cours IFT 607 couvre les outils fondamentaux sur lesquelles reposent la plupart des techniques en TALN. Le cours couvre également des applications répandues du TALN, telles la classification et l'étiquetage de documents, la traduction automatique et l'analyse grammaticale.

1.2 Objectifs spécifiques

À la fin de cette activité pédagogique, l'étudiante ou l'étudiant devrait connaitre, comprendre et être capable d'appliquer les approches de base en TALN :

1. prétraitement de données textuelles;
2. modèles de langage N-gramme;
3. modèles de Bayes naïf pour la classification de documents;
4. modèles de Markov cachés pour l'étiquetage de documents;
5. modèles IBM et *phrase-based* pour la traduction automatique;
6. modèles à base de grammaire probabiliste hors-contexte.

À la fin du cours, l'étudiant aura acquis les outils fondamentaux nécessaires à l'exploration d'autres sujets en TALN non-traités dans le cours.

1.3 Contenu détaillé

Thème	Contenu	Objectif	Nb. d'heures
Traitement de données textuelles	Expressions régulières	1	6
	Segmentation, lemmatisation et normalisation de mots		
	Distance d'édition		
	Extraction de dictionnaires et de caractéristiques		
Modélisation de textes	Modèle de langage N-gramme Évaluation de performance (perplexité)	2	6
	Techniques de lissage d'un modèle N-gramme Structures de données pour modèle N-gramme		
	Interprétation maximum de vraisemblance et a posteriori		
	Classification de documents par modèle de Bayes naïf Évaluation de performance (précision, rappel, F1)	3	3
	Étiquetage de mots par modèle de Markov caché	4	3
Traduction automatique	Données textuelles bilingues	5	12
	Modèles de traduction IBM		
	Modèle de traduction *phrase-based*		
	Décodage		
	Évaluation de systèmes de traduction (BLEU)		
Analyse grammaticale	Grammaire hors-contexte Forme normale de Chomsky	6	9
	Grammaire hors-contexte probabiliste		
	Algorithme CKY		
	Lexicalisation de grammaire probabiliste hors-contexte		
	Évaluation de systèmes d'analyse grammaticale		

2 Organisation

2.1 Méthode pédagogique

Le cours comprend trois heures d'exposé magistral et six heures de travail personnel par semaine. Le contenu du cours sera présenté à l'aide de diapositives qui seront mises en ligne progressivement au cours de la session. Du temps en travail dirigé (exercices en classe) est également prévu, où l'étudiant pourra tester et vérifier ses connaissances théoriques du contenu du cours.

En plus d'un examen intra et d'un examen final, quatre devoirs permettront d'évaluer les connaissances des étudiants en les mettant en application dans le cadre d'exercices de programmation.

2.2 Calendrier du cours

Le calendrier détaillé des séances de cours est disponible dans la section **CONTENU** du site web du cours.

2.3 Évaluation

Devoirs (4) : 40 %
Examen intra : 20 %
Examen final : 40 %

L'attribution des notes finales se fait selon les règles suivantes :

Note chiffrée	**Note finale**
note ≥ 90	A+
90 > note ≥ 85	A
85 > note ≥ 80	A-
80 > note ≥ 77	B+
77 > note ≥ 73	B
73 > note ≥ 70	B-
70 > note ≥ 65	C+
65 > note ≥ 60	C
60 > note ≥ 57	C-
57 > note ≥ 54	D+
54 > note ≥ 50	D
50 > note	E

2.3.1 Qualité du français et de la présentation

Conformément aux articles 36, 37 et 38 du Règlement facultaire d'évaluation des apprentissages[2], 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

Un document dont le texte et la structure se rapportent à des textes intégraux tirés d'un livre, d'une publication scientifique ou même d'un site Internet doit être référencé adéquatement. Lors de la correction de tout travail individuel ou de groupe une attention spéciale sera portée au plagiat, défini dans le Règlement des études comme « le fait, dans une activité pédagogique évaluée, de faire passer indûment pour siens des passages ou des idées tirés de l'œuvre d'autrui. ». Le cas échéant, le plagiat est un délit qui contrevient à l'article 8.1.2 du Règlement des études[3] : « tout acte ou manœuvre visant à tromper quant au rendement scolaire ou quant à la réussite d'une exigence relative à une activité pédagogique. » À titre de sanction disciplinaire, les mesures suivantes peuvent être imposées : a) l'obligation de reprendre un travail, un examen ou une activité pédagogique et b) l'attribution de la note E ou de la note 0 pour un travail, un examen ou une activité évaluée. Tout travail suspecté de plagiat sera référé au Secrétaire de la Faculté des sciences.

2.4 Devoirs

Devoir	Thème	Pondération
1	Modèles de langage et lissage	10 %
2	Étiquetage de documents	10 %
3	Traduction automatique	10 %
4	Analyse grammaticale	10 %

[2] http://www.usherbrooke.ca/accueil/fileadmin/sites/accueil/documents/direction/politiques/2500-008-sciences.pdf

[3] http://www.usherbrooke.ca/programmes/etude

Directives particulières

- Les devoirs doivent être effectués de façon individuelle;
- L'implémentation d'algorithmes dans le cadre des devoirs doit se faire dans le langage de programmation Python. Le code soumis doit être compatible avec (c'est-à-dire exécutable sous) la version 2.6.5 de Python, soit celle installée dans les laboratoires sous Ubuntu;
- Toute soumission en retard vaut zéro, sauf celles motivées par des raisons valables et conformes au règlement des études (par exemple, maladie avec attestation d'un médecin).

3 Matériel utile pour le cours

3.1 Références (recommandées)

- Daniel Jurafsky & James H. Martin. *Speech and Language Processing*. Prentice Hall, 2nd Edition, 2008.
  Sujets : manipulation de données textuelles, modèles N-gramme, étiquetage de documents, analyse grammaticale, reconnaissance vocale
- Christopher D. Manning & Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
  Sujets : manipulation de données textuelles, modèles N-gramme, classification et étiquetage de documents, traduction automatique, analyse grammaticale
- Philipp Koehn. *Statistical Machine Translation*. Cambridge University Presse, 2010.
  Sujet : traduction automatique
- Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. Association for Computational Linguistics, 2003.
  Sujet : analyse grammaticale

3.2 Ressources en ligne

- Plan de cours
- Présentations (PDF) des cours magistraux
- Forum de discussion

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

8.1.2 Relativement aux activités pédagogiques

L'expression délit désigne d'abord tout acte ou toute manœuvre visant à tromper quant au rendement scolaire ou quant à la réussite d'une exigence relative à une activité pédagogique.

Sans restreindre la portée générale de ce qui précède, est considéré comme un délit :

a) la substitution de personnes ou l'usurpation d'identité lors d'une activité évaluée ou obligatoire;

b) le plagiat, soit le fait, dans une activité évaluée, de faire passer indûment pour siens des passages ou des idées tirés de l'œuvre d'autrui;

c) l'obtention par vol ou par toute autre manœuvre frauduleuse de document ou de matériel, la possession ou l'utilisation de tout matériel non autorisé avant ou pendant un examen ou un travail faisant l'objet d'une évaluation;

d) le fait de fournir ou d'obtenir toute aide non autorisée, qu'elle soit collective ou individuelle, pour un examen ou un travail faisant l'objet d'une évaluation;

e) le fait de soumettre, sans autorisation préalable, une même production comme travail à une deuxième activité pédagogique;

f) la falsification d'un document aux fins d'obtenir une évaluation supérieure dans une activité ou pour l'admission à un programme.

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.