

IT4772E XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Phiên bản: 2020.17.05

1. THÔNG TIN CHUNG

Tên học phần:	Xử lý ngôn ngữ tự nhiên (<i>Natural Language Processing</i>)
Mã số học phần:	IT4772E
Khối lượng:	3(3-1-0-6) <ul style="list-style-type: none">- Lý thuyết: 45 tiết- Bài tập lớn: 15 tiết- Thí nghiệm: 0 tiết
Học phần tiên quyết:	-
Học phần học trước:	- IT3160: Nhập môn Trí tuệ nhân tạo
Học phần song hành:	Không

2. MÔ TẢ HỌC PHẦN

Học phần này cung cấp các kiến thức cơ bản về các bài toán xử lý ngôn ngữ tự nhiên (XLNNTN) thông qua máy tính: phân tích hình thái từ, tách từ, gán nhãn từ loại, phân tích cú pháp, phân tích ngữ nghĩa. Ngoài ra, sinh viên cũng được giới thiệu một số ứng dụng quan trọng của XLNNTN: hệ thống hỏi đáp tự động, dịch máy, trích rút thông tin. Thông qua nhiệm vụ của bài tập lớn, sinh viên sẽ có được kinh nghiệm xây dựng một ứng dụng thực tế có sử dụng các kỹ thuật XLNNTN đã được học. Ngoài ra môn học cũng cung cấp cho sinh viên các kỹ năng nghiên cứu, kỹ năng làm việc nhóm, thuyết trình và thái độ cần thiết để làm việc trong công ty sau này.

3. MỤC TIÊU VÀ CHUẨN ĐẦU RA CỦA HỌC PHẦN

Sinh viên hoàn thành học phần này có khả năng:

Mục tiêu/CDR	Mô tả mục tiêu/Chuẩn đầu ra của học phần	CDR được phân bổ cho HP/ Mức độ (I/T/U)
[1]	[2]	[3]
M1	Nắm vững các kiến thức cơ bản về các bài toán xử lý ngôn ngữ tự nhiên	[1.1,1.2,1.3]
M1.1	Có khả năng biểu diễn và mô hình hóa bài toán	[1.1,1.2](T)
M1.2	Biết cách áp dụng các kỹ thuật trong XLNNTN vào việc giải quyết bài toán	[1.3](IU)
M2	Nắm được một số phương pháp giải quyết các ứng dụng quan trọng của XLNNTN như hỏi đáp tự động, dịch máy, trích rút thông tin.	[1.2,1.3,2.2,2.3,2.4]
M2.1	Nắm được các khái niệm và các kỹ thuật cơ bản để phát triển một số ứng dụng XLNNTN	[1.2,1.3](T)

Mục tiêu/CDR	Mô tả mục tiêu/Chuẩn đầu ra của học phần	CDR được phân bổ cho HP/ Mức độ (I/T/U)
M2.2	Có khả năng tự tìm hiểu sâu hơn các kỹ thuật nâng cao để giải quyết các bài toán theo hướng XLNNTN	[2.2,2.3,2.4](IU)
M3	Có khả năng nghiên cứu, phân tích, đánh giá, vận dụng các kiến thức đã học vào việc xây dựng các hệ thống xử lý ngôn ngữ thông qua làm bài tập lớn theo nhóm	[2.1,2.2,2.3,3.1,3.2,4.2÷4.5]
M3.1	Có khả năng phân tích vấn đề, nghiên cứu, lựa chọn giải pháp và đưa ra hướng giải quyết cho bài toán	[2.1,2.2,2.3](T)
M3.2	Có khả năng triển khai xây dựng phần mềm xử lý ngôn ngữ	[4.2÷4.5](U)
M3.3	Biết cách làm việc nhóm, viết báo cáo và thuyết trình	[3.1,3.2](U)
M4	Nhận biết được các lĩnh vực nghiên cứu và triển vọng phát triển của xử lý ngôn ngữ tự nhiên	[1.3,2.2]
M4.1	Nắm được các hướng nghiên cứu và ứng dụng của XLNNTN như xây dựng các hệ thống dịch máy, chatbot, trợ lý ảo, phân tích quan điểm, tóm tắt văn bản, trích rút thông tin,...	[1.3,2.2](T)

4. TÀI LIỆU HỌC TẬP

Giáo trình

[1]

Sách tham khảo

- [1] James Allen (1994). *Natural Language Understanding*. The Benajmins/Cummings Publishing Company Inc.
- [2] Yoav Goldberg (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers
- [3] Dan Jurafsky and James Martin (2000, 2019). *Speech and Language Processing* (2nd edition, 3rd edition draft). Prentice Hall
- [4] Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge
- [5] Christopher Manning and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Bài báo khoa học

1. Phuong Le-Hong et al. (2008). *A hybrid approach to word segmentation of Vietnamese texts*. Proceedings of the 2nd International Conference on Language and Automat Theory and Applications, LATA 2008, Tarragona, Spain.
2. Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phương (2003). *Sử dụng bộ gán nhãn từ loại xác suất QTAG cho văn bản tiếng Việt*. Kỷ yếu Hội thảo ICT.rda'03
3. Phan Xuân Hiếu (2009). *Công cụ gán nhãn từ loại tiếng Việt dựa trên Conditional Random Fields và Maximum Entropy JvnTagger*.
4. Kenji Yamada, Kevin Knight (2001). *A syntax-based statistical translation model*. ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics July 2001 Pages 523–530

5. CÁCH ĐÁNH GIÁ HỌC PHẦN

Điểm thành phần	Phương pháp đánh giá cụ thể	Mô tả	CDR được đánh giá	Tỷ trọng
[1]	[2]	[3]	[4]	[5]
A1. Điểm quá trình (*)	Đánh giá quá trình			40%
	A1.1. Bài tập tại lớp	Tự luận	M1	20%
	A1.2. Bài tập lớn	Làm việc nhóm, viết báo cáo, thuyết trình	M1÷M4	20%
	A2.1. Thi cuối kỳ	Thi viết	M1,M2,M4	60%
A2. Điểm cuối kỳ				

* Điểm quá trình sẽ được điều chỉnh bằng cách cộng thêm điểm chuyên cần. Điểm chuyên cần có giá trị từ -2 đến +1, theo Quy chế Đào tạo đại học hệ chính quy của Trường ĐH Bách khoa Hà Nội.

6. KẾ HOẠCH GIẢNG DẠY

Tuần	Nội dung	CDR học phần	Hoạt động dạy và học	Bài đánh giá
[1]	[2]	[3]	[4]	[5]
1	1. Mở đầu 1.1 Giới thiệu chung 1.2 Các hướng nghiên cứu trong xử lý ngôn ngữ tự nhiên 1.3 Các mức phân tích trong xử lý ngôn ngữ tự nhiên 1.4 Các ứng dụng của xử lý ngôn ngữ tự nhiên	M4	Giảng bài	A2.1
2	2. Mô hình ngôn ngữ 2.1 n-gram	M1	Đọc trước tài liệu;	A1.1 A2.1

Tuần	Nội dung	CDR học phần	Hoạt động dạy và học	Bài đánh giá
[1]	[2]	[3]	[4]	[5]
	2.2 Đánh giá mô hình ngram 2.3 Làm tron 3. Tách từ tiếng Việt 3.1 Giới thiệu bài toán 3.2 Từ vựng tiếng Việt 3.3 Quy tắc cấu tạo từ tiếng Việt 3.4 Các cách tiếp cận trong tách từ 3.4.1 Tách từ dựa trên luật 3.4.2 Tách từ dựa trên học máy 3.5 Một số công cụ tách từ		Giảng bài; Bài tập tại lớp	
3	4. Gán nhãn từ loại 4.1 Giới thiệu bài toán 4.2 Các cách tiếp cận trong gán nhãn từ loại 4.2.1 Phương pháp Hidden Markov Model 4.2.2 Phương pháp dựa trên ràng buộc ngữ pháp 4.2.3 Phương pháp chuyển đổi 4.3 Một số nghiên cứu về gán nhãn từ loại tiếng Việt	M1	Đọc trước tài liệu; Giảng bài; Thảo luận	A1.1 A2.1
4	5. Phân tích cú pháp 5.1 Giới thiệu bài toán 5.2 Phân tích cú pháp dựa trên văn phạm phi ngữ cảnh 5.2.1 Giới thiệu 5.2.2 Phân tích kiểu trên xuống 5.2.3 Phân tích kiểu dưới lên 5.2.4 Một số vấn đề trong phân tích cú pháp	M1	Đọc trước tài liệu; Giảng bài; Bài tập tại lớp	A1.1 A2.1
5	5.3 Phân tích cú pháp theo cách tiếp cận xác suất 5.3.1 Một số khái niệm về xác suất 5.3.2 Phân tích cú pháp sử dụng xác suất	M1	Đọc trước tài liệu; Giảng bài; Bài tập tại lớp	A1.1 A2.1
6	5.4 Phân tích cú pháp phụ thuộc 5.4.1. Phân tích dựa trên trạng thái 5.4.2. Phân tích dựa trên đồ thị	M1	Đọc trước tài liệu; Giảng bài; Thảo luận	A1.1 A2.1
7	6. Phân tích vai nghĩa 6.1. Vai nghĩa là gì	M1	Đọc trước tài liệu;	A1.1 A2.1

Tuần	Nội dung	CDR học phần	Hoạt động dạy và học	Bài đánh giá
[1]	[2]	[3]	[4]	[5]
	6.2. FrameNet 6.3. Gán nhãn vai trò ngữ nghĩa 6.4. Ràng buộc lựa chọn		Giảng bài; Thảo luận	
8	7. Nghĩa từ vựng và phân giải nhập nhằng từ 7.1. Từ đồng âm, đồng nghĩa, đa nghĩa 7.2. Nghĩa từ vựng 7.3. Đo khoảng cách ngữ nghĩa từ dựa trên WordNet 7.4. Xác định nghĩa của từ dựa trên học máy 7.5. Xác định từ đồng nghĩa dựa trên học máy	M1	Đọc trước tài liệu; Giảng bài; Thảo luận	A1.1 A2.1
9	8. Hệ thống hỏi đáp 8.1 Giới thiệu bài toán 8.2 Một số cách tiếp cận trong xây dựng hệ thống hỏi đáp	M2	Đọc trước tài liệu; Giảng bài; Thảo luận	A2.1
10	9. Dịch máy 9.1. Giới thiệu bài toán 9.2. Dịch máy thống kê 9.2.1. Giống hàng câu 9.2.2. Giống hàng từ	M2	Đọc trước tài liệu; Giảng bài; Bài tập tại lớp	A2.1
11	9.3. Dịch máy dựa trên chuyển đổi cú pháp 9.4. Một số mã nguồn mở về dịch máy 9.5. Một số hệ thống dịch máy trên thế giới	M2	Đọc trước tài liệu; Giảng bài; Thảo luận	A2.1
12	10. Trích rút thông tin 10.1 Giới thiệu bài toán 10.2 Các bài toán trong trích rút thông tin 10.3. Nhận dạng thực thể 10.3.1. Nhận dạng thực thể sử dụng luật 10.3.2. Nhận dạng thực thể sử dụng học máy	M2	Đọc trước tài liệu; Giảng bài; Bài tập tại lớp	A2.1
13	10.4. Điền mẫu dựa trên thông tin trích rút	M2	Đọc trước tài liệu; Giảng bài; Thảo luận	A2.1
14	Các sinh viên báo cáo bài tập lớn	M1÷M4	Báo cáo bài tập lớn	A1.2

Tuần	Nội dung	CDR học phần	Hoạt động dạy và học	Bài đánh giá
[1]	[2]	[3]	[4]	[5]
15	Các sinh viên báo cáo bài tập lớn	M1÷M4	Báo cáo bài tập lớn	A1.2

8. NỘI DUNG BÀI TẬP LỚN

Sinh viên cài đặt một ứng dụng thực tế sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên. BTL được làm theo nhóm từ 1 đến 4 sinh viên

9. QUY ĐỊNH CỦA HỌC PHẦN

10. NGÀY PHÊ DUYỆT:

Chủ tịch Hội đồng

Nhóm xây dựng đề cương

Lê Thanh Hương
Nguyễn Kiên Hiếu
Nguyễn Kim Anh

11. QUÁ TRÌNH CẬP NHẬT

Lần cập nhật	Nội dung điều chỉnh	Ngày tháng được phê duyệt	Áp dụng từ kỳ/khóa	Ghi chú
1			
2			