ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG MÔN HỌC
DS310 – XỬ LÝ NGÔN NGỮ TỰ NHIÊN CHO KHOA HỌC DỮ LIỆU

1. THÔNG TIN CHUNG (General information)

Tên môn học (tiếng Việt):	Xử lý ngôn ngữ tự nhiên cho Khoa học dữ liệu
Tên môn học (tiếng Anh):	Natural Language Processing for Data Science
Mã môn học:	**DS310**
Thuộc khối kiến thức:	Đại cương ☐; Cơ sở nhóm ngành ☐;
	Cơ sở ngành ☐; Chuyên ngành ☑; Tốt nghiệp ☐
Khoa, Bộ môn phụ trách:	Khoa Khoa học và Kỹ thuật Thông tin
	Bộ môn Khoa học dữ liệu
Giảng viên biên soạn:	TS. Nguyễn Lưu Thuỳ Ngân (nganlt@uit.edu.vn)
	ThS. Nguyễn Văn Kiệt (kietnv@uit.edu.vn)
	ThS. Nguyễn Đức Vũ (vund@uit.edu.vn)
	CN. Lưu Thanh Sơn (sonlt@uit.edu.vn)
	CN. Huỳnh Văn Tín (tinhv@uit.edu.vn)
Số tín chỉ:	4
Lý thuyết:	3
Thực hành:	1
Tự học:	8
Môn học tiên quyết:	*Không*
Môn học trước:	*Học máy thống kê (DS102) hoặc Máy học (CS114)*

2. MÔ TẢ MÔN HỌC (Course description)

Môn học cung cấp cho sinh viên các kiến thức từ cơ bản và nâng cao về các kỹ thuật và công cụ áp dụng vào lĩnh vực Xử lý ngôn ngữ tự nhiên. Sinh viên được tiếp cận và làm quen với các bài toán phổ biến hiện tại và có khả năng nghiên cứu, giải quyết vấn đề liên quan đến lĩnh vực Xử lý ngôn ngữ tự nhiên và Khoa học dữ liệu.

3. MỤC TIÊU MÔN HỌC (Course goals)

Sau khi hoàn thành môn học này, sinh viên có thể:

- Hiểu được các kỹ thuật cơ bản về xử lý ngôn ngữ tự nhiên.
- Biết được các bài toán xử lý ngôn ngữ tự nhiên, khó khăn và thách thức của các bài toán.

- Áp dụng các công cụ máy học và học sâu để giải quyết các bài toán Xử lý ngôn ngữ tự nhiên.

Ký hiệu	Mục tiêu môn học	Chuẩn đầu ra trong CTĐT
G1	Hiểu và vận dụng các khái niệm và kỹ thuật cơ bản về xử lý ngôn ngữ tự nhiên: biểu diễn từ, vector từ, word embedding, …	*LO2, LO3*
G2	Hiểu và vận dụng các kiến thức về máy học vào xử lý ngôn ngữ tự nhiên: các mô hình máy học truyền thống, các mô hình Deep learning, …	*LO2, LO3*
G3	Hiểu, vận dụng và đánh giá một số bài toán xử lý ngôn ngữ tự nhiên phổ biến hiện tại và ứng dụng vào lĩnh vực Khoa học dữ liệu.	*LO10*

4. CHUẨN ĐẦU RA MÔN HỌC (Course learning outcomes)

CĐRMH	Mô tả CĐRMH (Mục tiêu cụ thể)	Mức độ giảng dạy
G1.1	Biết và hiểu được các khái niệm và kỹ thuật cơ bản về xử lý ngôn ngữ tự nhiên	*TU*
G1.2	Vận dụng các kỹ thuật cơ bản để xử lý cho bài toán cụ thể đặt ra.	*TU*
G2.1	Biết và hiểu được các mô hình máy học như: Logistic regression, Naive Bayes, và các mô hình học sâu: ANN, RNN, LSTM, GRU sử dụng trong xử lý ngôn ngữ tự nhiên	*TU*
G2.2	Vận dụng và đánh giá các kỹ thuật máy học cho bài toán cụ thể	*IT*
G3.1	Biết và hiểu được một số bài toán trong XLTNN hiện tại: phân tích văn bản mạng xã hội (social texts), hiểu ngôn ngữ (natural language understanding), dependencies parsing, dịch máy, …	*IT*
G3.2	Vận dụng và đánh giá các kỹ thuật đã học để giải quyết bài toán.	*IT*

5. NỘI DUNG MÔN HỌC, KẾ HOẠCH GIẢNG DẠY (Course content, lesson plan)

a. Lý thuyết

Buổi học (*45* tiết)	Nội dung	CĐRMH	Hoạt động dạy và học	Thành phần đánh giá
Buổi 1: (4 tiết)	**Các kỹ thuật cơ bản trong xử lý ngôn ngữ tự nhiên:** - Regular Expression. - Text Normailization. - Tách từ.	*G1.1*	**Dạy**: Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng, trả lời các câu hỏi của giảng viên và làm bài tập trên lớp. **Học ở nhà:** Đọc thêm: *Chương 2 sách Speech and Language processing (3rd).*	*A1*
Buổi 2: (4 tiết)	**Mô hình ngôn ngữ N-gram:** - Giới thiệu về N-gram. - Đánh giá một mô hình ngôn ngữ. - Các kỹ thuật smoothing cơ bản.	*G1.1, G1.2*	**Dạy**: Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng, trả lời các câu hỏi của giảng viên và làm bài tập trên lớp. **Học ở nhà:** Đọc thêm: *Chương 3 sách Speech and Language processing (3rd).*	*A1*
Buổi 3: (4 tiết)	**Naive Bayes và bài toán phân loại cảm xúc văn bản (sentiment analysis):** - Naive Bayes cho bài toán phân lớp. - Các độ đo đánh giá mô hình phân lớp. - Cross validation. - Statistical significance testing	*G2.1 G2.2*	**Dạy**: Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng, trả lời các câu hỏi của giảng viên và làm bài tập trên lớp. **Học ở nhà:** Đọc thêm: *Chương 4 sách Speech and Language processing (3rd).*	*A1*
Buổi 4: (4 tiết)	**Vector Semantic, Word embeeding và mạng neural cho language model:** - Lexical and Vector Semantic. - Từ và vector từ. - Độ đo tương tự cosine. - TF-IDF. - Mô hình Word2Vec. - Đánh giá mô hình vector.	*G1.1 G2.1*	**Dạy:** Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng và làm bài tập. **Học ở nhà:** Đọc thêm: *Chương 6+7 sách Speech and Language processing (3rd).*	*A1*

	- Kiến trúc mạng neural. - Vấn đề với hàm XOR. - Mạng truyền xuôi Feed forward. - Huấn luyện mạng neural.			
Buổi 5: (4 tiết)	**Gán nhãn chuỗi (sequence labeling) .** - Giới thiệu về mô hình sequence. - Part of speech tagging (POS Tagging). - Name Entity Recognition (NER). - Conditional Random Fields (CRF). - Các kỹ thuật đánh giá.	*G1.1* *G2.1* *G2.2*	**Dạy:** Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng và làm bài tập. **Học ở nhà**: Đọc thêm: *Chương 8 sách Speech and Language processing (3rd).*	*A1*
Buổi 6: (4 tiết)	**Các kiến trúc Deep learning:** - Nhắc lại về language model. - Mạng neural hồi quy. - Case study: LSTM, GRU. - Self-attention model. - Transformers.	*G2.1* *G2.2*	**Dạy**: Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng và trả lời các câu hỏi của giảng viên. **Học ở nhà:** Đọc thêm: *Chương 9 sách Speech and Language processing (3rd)*	*A1*
Buổi 7: (4 tiết)	**Văn phạm ngữ đoạn và phân tích cú pháp:** - Giới thiệu về Constituency. - Context-Free Grammar. - Treebanks. - Normal Form và Grammar Equivalence. - Lexicalized grammars. - Các kỹ thuật parsing.	*G1.1* *G1.2*	**Dạy**: Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng và trả lời các câu hỏi của giảng viên. **Học ở nhà:** Đọc thêm: *Chương 12+13 sách Speech and Language processing (3rd).*	*A1*
Buổi 8: (4 tiết)	**Phân tích cú pháp theo văn bản phụ thuộc:** - Dependency relation. - Dependency formalism. - Dependency treebanks. - Transition-based parsing và graph parsing. - Đánh giá mô hình.	*G3.1* *G3.2*	**Dạy ở lớp:** Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng và trả lời các câu hỏi của giảng viên. **Học ở nhà:** Đọc thêm: *Chương 14 sách Speech and Language processing (3rd).*	*A1*

Buổi 9: (4 tiết)	**Sematic parsing và rút trích thông tin** - Mô hình biểu diễn ngữ nghĩa. - First order logic. - Biểu diễn Event và State. - Description logic. - Relation extraction. - Các thuật toán extraction. - Extraction times	*G1.1* *G1.2*	**Dạy ở lớp:** Giảng viên giới thiệu qua các kiến thức. **Học ở lớp:** Sinh viên theo dõi bài giảng và trả lời các câu hỏi của giảng viên. **Học ở nhà:** Đọc thêm: *Chương 15 + 17 sách Speech and Language processing (3rd).*	*A1*
Buổi 10: (4 tiết)	**Hệ thống hỏi đáp** - Nhắc lại về Truy xuất thông tin. - Entity-linking. - Classic QA. - Open domain QA. - Đánh giá mô hình QA	*G3.1* *G3.2*	**Dạy**: Giảng viên giới thiệu qua các kiến thức. **Học ở lớp**: Sinh viên nghe giảng bài và đặt câu hỏi. **Học ở nhà:** Đọc thêm: *Chương 23 sách Speech and Language*	*A1*
Buổi 11: (5 tiết)	**Chatbot** - Đặc điểm của cuộc hội thoại. - Các đặc điểm của chatbot. - Dialogue State Architecture. - Thiết kế hệ thống chatbot theo tiếp cận dialogue. - Đánh giá hệ thống.	*G3.1* *G3.2*	**Dạy**: Giảng viên giới thiệu qua các kiến thức. **Học ở lớp**: Sinh viên nghe giảng bài và đặt câu hỏi. **Học ở nhà:** Đọc thêm: *Chương 24 sách Speech and Language*	*A1*
Buổi 12	**Sinh viên báo cáo đồ án**	*G3.1* *G3.2*	**Dạy:** Giảng viên theo dõi và góp ý cho bài thuyết trình của sinh viên. **Học ở lớp:** Sinh viên thuyết trình đồ án.	*A4*

b. Thực hành

- Hình thức thực hành: 2

Buổi học	**Nội dung**	**CĐR MH**	**Hoạt động dạy và học**	**Thành phần đánh giá**
Buổi 1 (5 tiết)	**Bài thực hành 1**: Bài toán phân loại văn bản (text classification).	*G2.2* *G3.2*	**Dạy**: Giảng viên hướng dẫn thực hành theo từng bước trên công cụ colab **Học ở lớp**: Sinh viên làm theo hướng dẫn	*A3*

			của giảng viên thực hành. **Học ở nhà**: Sinh viên hoàn thành bài lab và nộp lại theo yêu cầu của giảng viên.	
Buổi 2 (5 tiết)	**Bài thực hành 2**: Bài toán gán nhãn chuỗi (sequence labeling).	*G2.2* *G3.2*	**Dạy**: Giảng viên hướng dẫn thực hành theo từng bước trên công cụ colab **Học ở lớp**: Sinh viên làm theo hướng dẫn của giảng viên thực hành. **Học ở nhà**: Sinh viên hoàn thành bài lab và nộp lại theo yêu cầu của giảng viên.	*A3*
Buổi 5 (5 tiết)	**Bài thực hành 3**: Bài toán parsing.	*G2.2* *G3.2*	**Dạy**: Giảng viên hướng dẫn thực hành theo từng bước trên công cụ colab **Học ở lớp**: Sinh viên làm theo hướng dẫn của giảng viên thực hành. **Học ở nhà**: Sinh viên hoàn thành bài lab và nộp lại theo yêu cầu của giảng viên.	*A3*

6. ĐÁNH GIÁ MÔN HỌC (Course assessment)

Thành phần đánh giá *[1]*	**CĐRMH** *[2]*	**Tỷ lệ (%)** *[3]*
A1. Quá trình (Kiểm tra trên lớp, bài tập, đồ án, …)	*G1, G2*	*20%*
A2. Giữa kỳ		*0%*
A3. Thực hành	*G2, G3*	*30%*
A4. Đồ án	*G1, G2, G3*	*50%*

a. Rubric của thành phần đánh giá A1

Đánh giá quá trình học tập tại lớp	***Giỏi (9-10đ)***	***Khá (7-8đ)***	***TB (6-7đ)***	***Yếu (4-5đ)***	***Kém (<3đ)***
Mức độ chuyên cần	*Tham gia đặt câu hỏi, và trả lời câu hỏi trên lớp. Hoàn thành 100% bài tập về nhà.*	*Tham gia trả lời câu hỏi tại lớp và hoàn thành 80% bài tập về nhà.*	*Hoàn thành 70% các bài tập về nhà.*	*Hoàn thành 40% các bài tập về nhà.*	*Không tham gia vào bất cứ hoạt động nào tại lớp.*

b. Rubric của thành phần đánh giá A2

Không có

c. Rubric của thành phần đánh giá A3

	Giỏi (9-10đ)	***Khá (7-8đ)***	***TB (6-7đ)***	***Yếu (4-5đ)***	***Kém (<3đ)***
Làm bài tập thực hành hằng tuần	*Làm >80% bài tập*	*Làm 70% bài tập*	*Làm 50% bài tập*	*Làm dưới 50% bài tập*	*Không làm bài tập nào*
Báo cáo thực hành	*Nội dung đầy đủ, trình bày rõ ràng*	*Đáp ứng được 80% nội dung yêu cầu*	*Đáp ứng được 50% nội dung yêu cầu*	*Đáp ứng được 30% nội dung yêu cầu*	*Không đáp ứng hoặc sai nội dung yêu cầu*

d. Rubric của thành phần đánh giá A4

Đồ án cuối kì	***Giỏi (9-10đ)***	***Khá (7-8đ)***	***TB (6-7đ)***	***Yếu (4-5đ)***	***Kém (<3đ)***
Thuyết trình	*Nói trôi chảy, tự tin và thu hút; Đảm bảo đúng thời lượng trình bày.*	*Nói trôi chảy, tự tin nhưng không thu hút; Nói ngắn hoặc dài hơn thời gian qui định 3 phút.*	*Nói trôi chảy nhưng không tự tin và thu hút; Nói quá ngắn hoặc quá dài thời gian qui định 10 phút.*	*Không chuẩn bị nghiêm túc.*	*Không thuyết trình hoặc gian lận.*
Báo cáo giấy	*- Nội dung đầy đủ theo yêu cầu của giảng viên.* *- Phương pháp luận rõ ràng.* *- Bố cục bài báo cáo hợp lý.* *- Cách trình bày lôi cuốn.*	*- Nội dung đầy đủ theo yêu cầu của giảng viên.* *- Phương pháp luận rõ ràng.* *Bố cục bài báo cáo hợp lý.*	*- Nội dung đáp ứng 70% yêu cầu.* *- Phương pháp luận rõ ràng.*	*- Nội dung đáp ứng khoảng 30%.* *- Phương pháp luận không rõ ràng.*	*- Sai hoàn toàn nội dung hoặc phương pháp luận.*

7. QUY ĐỊNH CỦA MÔN HỌC (Course requirements and expectations)

- Sinh viên đến lớp học theo quy định chung của nhà trường.
- Sinh viên làm bài tập và thực hiện đồ án môn học đầy đủ.

8. TÀI LIỆU HỌC TẬP, THAM KHẢO

Giáo trình chính

1. Dan Jurafsky and James Martin (2019). *Speech and Language Processing (3rd)*.

Tài liệu tham khảo

1. CS224n: Natural Language Processing with Deep Learning (Winter, 2020) - Đại học Stanford (Hoa Kỳ).
2. COMS W475: Natural Language Processing (2020) - Đại học Columbia (Hoa Kỳ).
3. **Gomez-Perez**, Jose Manuel, **Denaux**, Ronald, **Garcia-Silva**, Andres. A Practical Guide to Hybrid Natural Language Processing, Springer (2020).

9. PHẦN MỀM HAY CÔNG CỤ HỖ TRỢ THỰC HÀNH

1. Google Colab.
2. Jupyter notebook.

Tp.HCM, ngày …… tháng …… năm ……

Trưởng khoa/bộ môn	**Giảng viên biên soạn**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)