

**CHI TIẾT LỘ TRÌNH LỚP HỌC VIETNLP-01**

*(Nội dung giảng dạy sẽ được cập nhật phù hợp với tiến độ chung của lớp học)*

1. **Tên lớp học:** Lớp học nâng cao về xử lý ngôn ngữ tự nhiên (Advances in Natural Language Processing)
2. **Trình độ:** Nâng cao
3. **Phân bổ thời gian**   
   - Trên lớp:   
    + Thứ 5 (7 p.m ~ 9 p.m)  
    + Chủ nhật (9 a.m ~ 11:30 am)  
   - Tự học và nghiên cứu:  
    + 5 tiếng / tuần.
4. **Điều kiện tiên quyết**: để đạt được kết quả tốt trong lớp học thì người học cần:  
   - Nắm rõ kiến thức về Deep Learning cơ bản.  
   - Nắm rõ kiến thức về các cấu trúc mạng NLP cơ bản như: RNN (Recurrent Neural Network), LSTM (Long short-term memory), Bi-LSTM (Bidirectional Long Short-Term Memory), GRU (Gated recurrent units).   
   - Có thể tự xây dựng một mô hình Deep Learning bằng Python / TensorFlow hoặc Pytorch.  
   - Sử dụng thành thạo Google Colab hoặc Kaggle.
5. **Mô tả vắn tắt chương trình học**

Chương trình học bao gồm 5 phần chính:

* Ôn tập kiến thức cơ bản về NLP (**bổ sung**).
* Neural Machine Translation (bài toán dịch máy).
* Sequence Classification (bài toán phân loại).
* Question Answering (hệ thống hỏi đáp).
* Thảo luận cùng với các chuyên gia đầu ngành NLP (Guest Lecture - *dành cho lớp tích hợp*).

1. **Mục tiêu của lớp học**

Sau khi hoàn thành khóa học, học viên sẽ nắm vững kiến thức lý thuyết và thực hành về các mảng của xử lý ngôn ngữ tự nhiên được đề cập trong khóa học, bao gồm:

- Nắm được các khái niệm cơ bản của NLP, bao gồm tiền xử lý dữ liệu (các tokenizer, subword, language model, ...), các thành phần trong mạng học sâu (RNN, LSTM, Attention, Transformer), các kiến trúc mạng (LSTM, Attention, Transformer, BERT, RoBERTa) và cách các kiến trúc được sử dụng trong các bài toán cụ thể.

- Có thể tự xây dựng các kiến trúc mạng như LSTM, Transformer và huấn luyện mô hình cho các dạng bài toán trong NLP, cụ thể gồm các bài toán sau: dịch máy, gán nhãn chuỗi từ, phân loại, hỏi đáp.

1. **Nhiệm vụ của học viên**- Kênh thông tin trao đổi chính: **Slack VietAI\_NLP-01**. Tuân theo các nội quy kênh chat - thảo luận đúng theo các Channel được quy định. *(Ví dụ: Kênh “member-talk” để học viên có thể cùng nhau trò chuyện thêm các vấn đề xung quanh lớp học*).

- Trao đổi đóng góp ý kiến **trên tinh thần xây dựng**.

- Chủ động xem trước nội dung bài giảng, các yêu cầu chuẩn bị trước của học viên và tích cực đặt câu hỏi trên lớp và về nhà.

- Hoàn thành các bài tập trong lớp (chi tiết trong phần “*Tiêu chuẩn đánh giá*”).

1. **Đối với học viên Online**   
   - Các bạn Online có quyền tham gia khóa Offline khóa tiếp theo ở Sài Gòn miễn phí 100%. Các bạn có thể chọn tiếp tục tham gia lớp Online để học hỏi và học lại khóa 02 Offline miễn phí. Hoặc bảo lưu khóa hiện tại để tham gia từ đầu khóa Offline.  
     
   - Nội quy về học viên học Online:  
   Khi cần phát biểu sẽ mở Mic để phát biểu (Chú ý: tắt Mic sau khi phát biểu xong để không ảnh hưởng đến các học viên khác). Hoặc nhắn tin trên kênh thảo luận của Slack.

1. **Tài liệu học tập**   
   - Tiền điều kiện (*được đăng tải trên Google Classroom*):   
    + Tài liệu: Toán cho Machine Learning.  
    + Video: Chuỗi bài giảng về Deep Learning cơ bản.  
   - Slide bài giảng, lab thực hành, video record nội dung giảng dạy được đăng tải trên hệ thống **Google Classroom** của lớp học. Tài liệu sẽ được lưu trữ kể cả khi khóa học kết thúc để học viên có thể xem lại bất kỳ lúc nào.

- Trong trường hợp nếu học viên chia sẻ tài liệu ra bên ngoài sẽ bị coi là vi phạm quy định của lớp học.

1. **Tiêu chuẩn đánh giá**- Kaggle Compete: Sentiment Analysis - 10%  
   - Bài tập Machine translation - 25 %  
   - Bài tập Multi-classification problem - 25 %  
   - Project: Question Answering from a newspaper - 40 %

Trong trường hợp nộp bài muộn:

* Trong vòng 1 ngày: trừ 10% điểm.
* Trong vòng 3 ngày: trừ 30% điểm.
* Quá 3 ngày: 0 điểm.

1. **Thang điểm:** điểm của học viên được đánh giá theo thang điểm 10.  
   - Điểm tổng kết >= 9 điểm: Distinct.  
   - Điểm tổng kết >= 8 điểm: Merit.  
   - Điểm tổng kết >= 6 điểm: Pass.  
   - Điểm tổng kết < 6 điểm: Fail.
2. **Chính sách học bổng**   
   - Học bổng 50%, 30%, 20% lần lượt dành cho TOP 3 kết quả tổng kết lớp học.
3. **Nội dung chi tiết**

**Chú ý:** Nội dung giảng dạy sẽ được cập nhật phù hợp với tiến độ chung của lớp học.

|  |  |  |  |
| --- | --- | --- | --- |
| **Buổi** | **Nội dung giảng dạy** | **Tài liệu học tập** | **Chuẩn bị của học viên** |
| **Phần 1: Ôn tập kiến thức NLP** | | | |
| 1 | - Phần 1: Cơ hội nghề nghiệp với NLP.  - Phần 2: Giới thiệu về NLP + Word embedding: CBOW. | - [Slide] Giới thiệu NLP + Word2Vec  - [Practice] Xây dựng CBOW với dataset Truyện kiều | - Tìm hiểu trước về Word2Vec — Skip-gram and CBOW ([Link](https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314#:~:text=Continuous%20Bag%20of%20Words%20Model%20(CBOW)%20and%20Skip%2Dgram&text=In%20the%20CBOW%20model%2C%20the,used%20to%20predict%20the%20context%20.)).  - [[Video](https://www.youtube.com/watch?v=FT1ZZdcur5A&list=PLMm4sOMuA2QI5x_0KlNT3LuKDi6-ByboB&index=16)] Tokenization in NLP.  - [[Video](https://www.youtube.com/watch?v=akRbuXokLSo&list=PLMm4sOMuA2QI5x_0KlNT3LuKDi6-ByboB&index=17)] Word2vec. - [[Video](https://www.youtube.com/watch?v=JIafLwlGzBA&list=PLMm4sOMuA2QI5x_0KlNT3LuKDi6-ByboB&index=18)] Thực hành Train Sentiment Classification. |
| 2 | - Tổng quan kiến thức về Machine Learning và Deep Learning cơ bản.  - Kiến trúc mạng RNN & Bài toán Language Modeling. | - [Slide] Language model & Mạng RNN | - Xem trước qua các kiến trúc mạng RNN, LSTM, GRU.  - [[Video](https://www.youtube.com/watch?v=t0EoeTYU-fc&list=PLMm4sOMuA2QI5x_0KlNT3LuKDi6-ByboB&index=21)] Language Model and Recurrent Neural Network. |
| 3 | Phân loại câu sử dụng NN, Conv1D và RNN. | - [Codelab] Bài toán phân loại câu - “imdb-dataset” | - Ôn tập lại thư viện TensorFlow / Pytorch.  - Tìm hiểu trước bộ dữ liệu imdb-review.  - [[Paper](https://classroom.google.com/u/4/c/MjUwNDU4Njc4NzQ3/m/MjU0NjA5OTExMDkw/details)] Text Classification using Convolution.  - [[Video](https://www.youtube.com/watch?v=X6bYTZJkEDQ&list=PLMm4sOMuA2QI5x_0KlNT3LuKDi6-ByboB&index=22)] Xây dựng mạng SimpleRNN, LSTM, Bi-directional. |
| 4 | Kiến trúc mạng LSTM / Bidirectional RNN  Deep Bidirectional RNN. | - Slide bài giảng về LSTM / Bi-LSTM | - Đọc lại nội dung về cấu trúc mạng LSTM / Bi-LSTM. |
|  | **Assignment 01**  **Kaggle Compete**: Sentiment Analysis trên bình luận tiếng Việt  Deadline: **06/02/2021** | - Hướng dẫn sử dụng Kaggle.  - Sample Notebook.  - Phương thức đánh giá: % độ chính xác trên Private Leaderboard. | - Liên kết: https://www.kaggle.com/c/vietnlp-01-assignment-01 |
| 5 | Deep RNN + Seq2Seq + NMT | - [Slide]  - [[Practice](https://classroom.google.com/u/4/c/MjUwNDU4Njc4NzQ3/m/MjU4NDY3MTYxOTIz/details)] Seq2Seq | - Sequence to Sequence Learning with Neural Networks: https://arxiv.org/pdf/1409.3215.pdf  - Stanford NMT: http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf  - CS224 Note: http://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes06-NMT\_seq2seq\_attention.pdf |
| **Phần 2: Neural Machine Translation** | | | |
| 6 | Dịch máy - Machine Translation | - [Slide] | Stanford NMT: http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf  Original seq2seq+attention paper: https://arxiv.org/pdf/1409.0473.pdf  BLEU Score: https://www.aclweb.org/anthology/P02-1040.pdf |
| 7 | Thực hành xây dựng NMT | - [Slide] Attention  - [Slide] NMT Summary  - [Codelab] | - [Reading] Practical Methodology |
| 8 | Ôn tập kiến thức & Hướng dẫn Pytorch | - [Tutorial] Pytorch | - Xem lại các Video bài giảng và kiến thức đã học từ buổi 01 đến 07. |
| 9 | Transformer - Encoder | - [Slide] Transformer - Encoder | - Cần đọc kỹ lại về cơ chế Attention.  - Xem lại cấu trúc Encoder - Decoder. |
| 10 | Transformer - Decoder | - [Slide] Transformer - Decoder |  |
|  | Applications of transformers in NLP problems  **Assignment 2**: Machine translation | - Notebook trên Google Classroom | - Phần Encoder: 2 tuần  - Phần Decoder + ứng dụng trong NMT: 1 tuần |
| **Phần 3: Sequence Classification** | | | |
| 11 | Pretrained model and BERT | - [Slide] Giới thiệu về BERT, RoBERTa, XLM-RoBERTa và các bài toán ứng dụng  - [Notebook] Ứng dụng BERT trong bài toán:   * Question & Answering. * Multi Label Classification. | - BERT: https://arxiv.org/pdf/1810.04805.pdf |
|  | **Assignment 3**: Multi-label classification problem | Được tổ chức dưới dạng Kaggle Compete. | Deadline: **21/03/2021** |
| 12 | Bài toán classification. | - [Slide] Giới thiệu các loại input feature, model và metric thường dùng.  - [Codebase] Bài toán Sentiment Classification. |  |
| 13 | Bài toán Tagging | - [Slide] Giới thiệu Named Entity Recognition và các bài toán tương tự.  - [Practice] Bài toán NER cho Vietnamese. |  |
| **Phần 4: Question Answering** | | | |
| 14 | Question Answering System | - [Slide] |  |
| 15 | Machine Reading Comprehension | - [Slide] |  |
| 16 | Knowledge Graph Question Answering | - [Slide] |  |
| 17 | Codelab Xây dựng mô hình Machine Reading Comprehension | - [Codelab] | Tìm hiểu tập dữ liệu [Squad 2.0](https://rajpurkar.github.io/SQuAD-explorer/) |
| 18 | **Announce final project**: Question Answering from a newspaper | - [Slide]  - [Code base] |  |
| **Phần 5: Thảo luận cùng với các chuyên gia đầu ngành NLP** | | | |
| 19 | Guest Lecture | Nội dung phần này sẽ được điều chỉnh phù hợp với tiến độ của lớp học. |  |
| 20 | Project update |  |
| 21 | Guest Lecture |  |
| 22 | Project update/Guest Lecture |  |
| 23 | Project defend |  |