THƯ VIỆN TỰ HỌC DEEP LEARNING

NATURAL LANGUAGE PROCESSING (NLP)

Hiện nay, số lượng bạn có nhu cầu học NLP đang tăng cao, bao gồm các bạn Advanced và mới bắt đầu. Tuy nhiên, tài liệu đa số từ nước ngoài và chưa có thư viện tập hợp, hoặc có thì cũng rải rác và khó cho các bạn mới và tự học.

Đó là lý do VietAI kêu gọi & kết nối các bạn trong cộng đồng AI/ML/DL:

- **Những bạn Chia sẻ:** chia sẻ nguồn học/tài liệu NLP/paper từ mọi nơi trên thế giới
- **Những bạn được chia sẻ**: xem, tự học, tìm hiểu & chủ động chia sẻ.

Tất cả các bạn chia sẻ sẽ nhận **CREDIT** trực tiếp trên file và tùy thuộc vào mức độ đóng góp, VietAI sẽ hỗ trợ để cùng bạn tổ chức nhiều hoạt động chia sẻ bổ ích.

THƯ VIỆN

Dưới đây là mục lục cơ bản, các bạn paste đường link tìm được vào mục tương ứng. Nếu mục bạn cần chưa có, bạn có thể chủ động bổ sung.

1. General NLP tools/libraries

HuggingFace Transformers – State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0: https://github.com/huggingface/transformers

Stanza – A Python NLP Package for Many Human Languages: https://stanfordnlp.github.io/stanza

NLTK - Natural Language Toolkit https://www.nltk.org/

Spacy – Industrial-Strength Natural Language Processing: https://spacy.io

Flair NLP https://github.com/flairNLP/flair

HuggingFace datasets and evaluation metrics: https://github.com/huggingface/datasets

AdapterHub Repo for pre-trained adapter modules: https://adapterhub.ml/

Facebook platform for training and evaluating dialogue models: https://parl.ai/

Facebook AI Seq2Seq library: https://github.com/pytorch/fairseq

OpenNMT for Machine Translation: https://opennmt.net/

NVIDIA conversational AI library based on Pytorch Lightning: https://github.com/NVIDIA/NeMo

Stanford NLP group: https://nlp.stanford.edu/software/

Rasa: Open source conversational AI https://rasa.com/

DeepPavlov: Open source conversational AI Framework: https://deeppavlov.ai/

fastText - library for efficient text classification and representation learning: https://fasttext.cc

NVIDIA's Megatron for training large scale LM https://github.com/NVIDIA/Megatron-LM

Xatkit - The easiest way to build powerful bots and chatbots:
https://github.com/xatkit-bot-platform/xatkit

GenSim - Library for topic modeling, document indexing and similarity retrieval with large corpora: https://github.com/RaRe-Technologies/gensim

CoreNLP: library aims to make the application of linguistic analysis tools to a piece of text easy and efficient: https://github.com/stanfordnlp/CoreNLP

OpenNLP: A powerful tool with a lot of features and ready for production workloads if you're using Java https://opennlp.apache.org/

TextBrewer - A Pytorch-based toolkit for NLP containing different distillation methods:
https://github.com/airaria/TextBrewer

Ecco - A Python library to visualize and explore NLP language models
https://github.com/jalammar/ecco/

LINSPECTOR - 15 word-level diagnostic tasks for 24 languages (including vi) to examine multilingual representation
https://github.com/UKPLab/linspector

torchdistill - An open source framework built on top of Pytorch and dedicated for knowledge distillations
https://github.com/yoshitomo-matsubara/torchdistill

Spark NLP - NLP library built on top of Apache Spark ML
https://github.com/JohnSnowLabs/spark-nlp

Trankit - A Python toolkit utilizing multilingual pretrained Transformer to build joint token and sentence splitter, joint model for POS, morphological tagging and dependency parsing
https://github.com/nlp-uoregon/trankit

2. General NLP datasets/benchmarks

Stanford Question Answering Dataset (SQuAD): https://rajpurkar.github.io/SQuAD-explorer/

General Language Understanding Evaluation (GLUE) benchmark: https://gluebenchmark.com/

Machine Translation (WMT): http://www.statmt.org/wmt20/

NLP Progress: http://nlpprogress.com/

Quora Question Pairs: https://www.kaggle.com/c/quora-question-pairs

[datasets] The Multi-Genre NLI Corpus (MultiNLI): https://cims.nyu.edu/~sbowman/multinli/ (A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference)

[datasets] Textual Entailment Resource Pool (RTE):
https://aclweb.org/aclwiki/Textual_Entailment_Resource_Pool

[datasets] The WikiText Long Term Dependency Language Modeling Dataset:
https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/

[datasets] Open WebText – an open-source effort to reproduce OpenAI’s WebText dataset:
https://skylion007.github.io/OpenWebTextCorpus/

SemEval https://semeval.github.io/

Metatex curated NLP datasets
Dataset list - Natural Language Processing (metatext.io)

[Dataset] Asian Language Treebank (ALT) project (13 parallel corpora. Treebank of some languages): https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

[Website] The Big Bad NLP Database https://datasets.quantumstat.com/

[Website] Word Sense Disambiguation Dataset http://danlou.github.io/uwa/

[Website] NLP Datasets by Niederhoffer [Github, 4400 starts]
https://github.com/niderhoff/nlp-datasets

[Website] 25 Best Parallel Text Datasets for Machine Translation Training
https://lionbridge.ai/datasets/25-best-parallel-text-datasets-for-machine-translation-training/

[Website] 20 Best German Language Datasets for Machine Learning
https://lionbridge.ai/datasets/20-best-german-language-datasets-for-machine-learning/

XTREME - a multi-task benchmark evaluating cross-lingual generalization of multilingual representations

https://github.com/google-research/xtreme

3. General NLP learning resources

Stanford Natural Language Processing with Deep Learning (Free):
http://web.stanford.edu/class/cs224n/

Online version-XCS224N (Fee): Natural Language Processing with Deep Learning | Stanford Online

Deeplearning.ai Natural Language Processing Specialization (Fee/financial aid available)
https://www.deeplearning.ai/natural-language-processing-specialization/

CS124 - From Language to Information(Free): From Languages to Information - YouTube

BERT tutorials by Chris McCormick (Fee):
ChrisMcCormick.AI
The BERT Collection (chrismccormick.ai)

Neural Network for NLP - Graham Neubig CMU
http://phontron.com/class/nn4nlp2020/schedule.html

Multilingual NLP - Graham Neubig CMU
http://demo.clab.cs.cmu.edu/11737fa20/

[Book] Neural Network Method for Natural Language Processing (Y. Golberg)
https://www.morganclaypool.com/doi/abs/10.2200/S00762ED1V01Y201703HLT037

[Book] Cross-lingual Word Embedding (Sebastian Ruder)
https://www.morganclaypoolpublishers.com/catalog_Orig/product_info.php?products_id=1419

[Book] Speech and Language Processing (3rd ed. draft)
https://web.stanford.edu/~jurafsky/slp3/

[Book] Speech and Language Processing, 2nd Edition
https://www.amazon.com/Speech-Language-Processing-Daniel-Jurafsky/dp/0131873210

[Book] Foundations of Statistical Natural Language Processing
https://nlp.stanford.edu/fsnlp/

[Book] Natural Language Processing (draft 2018)
http://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf

[Book] Introduction to Information Retrieval (Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze)
https://nlp.stanford.edu/IR-book/information-retrieval-book.html

[Blog] 100 days of NLP
https://github.com/graviraja/100-Days-of-NLP/tree/master/architectures

[Website] The Natural Language Processing Dictionary
http://www.cse.unsw.edu.au/~billw/nlpdict.html

[Blog] Lil'Log
https://lilianweng.github.io/lil-log/

[Blog] Sebastian Ruder
https://ruder.io/

[Blog] Jay Alammar
jalammar.github.io

[Blog] Chris McCormick
Mccormickml.com

[Blog] Google AI blog
Google AI Blog

[Blog] Facebook AI blog
https://research.fb.com/blog/

[Blog] Salesforce Research blog
https://www.salesforce.com/research/

[Blog] Jlibovicky's blog (about MT)
https://jlibovicky.github.io

[Book] Real-world Natural Language Processing
https://drive.google.com/file/d/1nkCXR_AtOMV7NSmV3azj2petTYmWTVeT/view?usp=sharing

[Blog] FloydHub Blog
https://blog.floydhub.com/

[Blog] Marek Kei Blog - Thoughts on ML and NLP
http://www.marekrei.com/blog/

[Website] paperswithcode.com
https://paperswithcode.com/area/natural-language-processing

[Website] Trending on EMNLP-2020
https://emnlp-2020.herokuapp.com/

[Album] NLP-Papers
https://github.com/gyunggyung/NLP-Papers

[Album] 100 Must-Read NLP Papers
https://github.com/mhagiwara/100-nlp-papers

[Album] The Best NLP Papers From ICLR 2020
https://www.topbots.com/best-nlp-papers-from-iclr-2020/

[Slide] From perceptrons to word embeddings
https://drive.google.com/file/d/111AAxCQsr8uVsInyrEPxIbSkOBBipkGX/view?usp=sharing

[Slide] Intel Natural Language Processing Course
https://software.intel.com/content/www/us/en/develop/training/course-natural-language-processing.html

[Book] Natural Language Processing with Python
https://www.nltk.org/book/

[Book] Practical Natural Language Processing
http://www.practicalnlp.ai/

[Course] Extension to NLP course at Yandex School of Data Analysis
https://lena-voita.github.io/nlp_course.html

[Album] A collection of 400+ Survey Papers on NLP and ML
https://github.com/NiuTrans/ABigSurvey

[Website] The Super Duper NLP Repo (Demos in Colab) https://notebooks.quantumstat.com/

[Website] The Model Forge (Models with respective source URLs)
https://models.quantumstat.com/

[Book-Draft] Embeddings in Natural Language Processing
http://josecamachocollados.com/book_embNLP_draft.pdf

Top conferences such as: EMNLP, ACL, NAACL, …

4. Vietnamese NLP tools / libraries

VNCoreNLP
vncorenlp/VnCoreNLP: A Vietnamese natural language processing toolkit (NAACL 2018)

PhoBERT
https://github.com/VinAIResearch/PhoBERT

vELECTRA
fpt-corp/vELECTRA: Vietnamese ELECTRA pre-trained model of FPT.AI

Underthesea
https://github.com/undertheseanlp/underthesea

Pyvi
https://github.com/trungtv/pyvi

Coccoc tokenizer
https://github.com/coccoc/coccoc-tokenizer

NK-VECTOR (NLP with JS)
https://github.com/trinhdoduyhungss/nk-vector

VNTK (NLP with JS)
https://github.com/vunb/vntk

VietChunker (chunking)
https://vlsp.hpda.vn/demo/?page=resources

ETNLP (extract, evaluate, visualize multiple embeddings)
https://github.com/vietnlp/etnlp

PhoNLP - a BERT-based multitask learning toolkit for Vietnamese POS tagging, NER and DP
https://github.com/VinAIResearch/PhoNLP

5. Vietnamese NLP datasets / benchmarks

NLP Progress Vietnamese:
https://github.com/undertheseanlp/NLP-Vietnamese-progress

VLSP
https://vlsp.org.vn/

Vietnamese Text2SQL
https://github.com/VinAIResearch/ViText2SQL

NLP@UIT Research Group:
https://sites.google.com/uit.edu.vn/uit-nlp/datasets-projects

[Dataset] Vietnamese German Dataset
https://www.kaggle.com/flightstar/vietnamese-german-dataset

6. Vietnamese NLP learning resources

100 exercises of NLP: https://github.com/minhpqn/nlp_100_drill_exercises_ver_2020

Top 100 NLP Questions:
https://drive.google.com/file/d/1L_9FKt10dWnzTnM0DJdQrU3Esf1f5_c5/view?fbclid=IwAR1ixGmWLu7Yw6vd75rCLOSNFDrpeZIdrlKJxlPzUOIl2rzkLje_wckzOAE

7. NLP Research Group

[USA] University of Washington
Homepage: https://www.cs.washington.edu/research/nlp
https://github.com/uwnlp

[Germany] Technical University of Darmstadt
Homepage: https://www.informatik.tu-darmstadt.de/ukp/ukp_home/index.en.jsp
https://github.com/UKPLab

[UK] University of Edinburgh
Homepage: https://www.ed.ac.uk/
https://github.com/EdinburghNLP

[USA] Harvard University
Homepage: Harvard NLP
HNLP (github.com)

[Singapore] Nanyang Technological University

Homepage: https://ntunlpsg.github.io/

ABOUT US

VietAI is a non-profit organization. Our mission is to build a community of world-class AI talents in Vietnam to solve meaningful and impactful problems for not just Vietnam but also the world.

- Facebook: https://www.facebook.com/vietaipublic
- Tham gia **Test Deep Learning NLP miễn phí** vào **9AM Chủ nhật 27/12** tại: https://zalo.me/g/hvqlts099 hoặc https://www.facebook.com/groups/vietainlp

CREDIT TO

Các bạn sau khi đóng góp, chủ động điền tên, title và social contact (optional). VietAI sẽ monitor hằng ngày để đảm bảo danh sách được cập nhật đúng & đủ. Chân thành cảm ơn mọi người.

PhD. Thắng Lương - Senior Research Scientist at Google Brain @lmthang
MSc Bình Nguyễn - Machine Learning Engineer at Vin Bigdata. Acknowledged as a Google Developer Expert in Machine Learning in Vietnam. @nguyenvulebinh
Anh Tuan Nguyen - Deep Learning Research Engineer at NVIDIA Nguyen Anh
Nguyen Binh Minh - AI Engineer. @bkbinhminh9x
Vũ Anh - NLP Engineer. anhv.ict91@gmail.com
Ngô Quang Huy - Computer Science student, PTIT, ngoquanghuy1999lp@gmail.com
Trịnh Đỗ Duy Hưng - Computer science student, University of Greenwich (Da Nang Campus). @trinhhungsss492
Nguyễn Trọng Hiếu - Member of ICON Club, Ton Duc Thang University @hieunguyen1053
Lê Đình Duy - Ai Engineer @daemon.lee.33

Đoàn Nguyễn Thành Lương - Computer Science student, Ulsan National Institute of Science and Technology, @doannguyenthanhluong

Nguyễn Việt Hoa - Computer Science student, Technical University of Darmstadt
Việt Hoa Nguyễn

Trang Minh Chiến - Computer Science Student at HCM University of Science
chientrangminh@gmail.com

Trần Quốc Sơn - Computer Science student at Denison University
tran_s2@denison.edu

Phạm Văn Tín - Software Engineer student, Duy Tan University - @tinspham209

Hoàng Nghĩa Tuyến - Math Science student, Nanyang Technological University, Singapore - @hnt4499

PhD. Xuan-Son Vu - Postdoctoral Researcher at UMU, Sweden

Nguyễn Bá Tùng Lâm - Information Technology student, Hanoi University of Science and Technology - tunglam.ngxba@gmail.com

Lê Trọng Đức - Computer Science student at Hanoi University of Science and Technology - trongduclebk@gmail.com