Introduction to Natural Language Processing

Spring 2023

Thursday 9:35 am – 12:25 pm

Instructor: Yeabin Moon ([yeabinmoon@brandeis.edu](mailto:yeabinmoon@brandeis.edu))

Subject to Change (October 27, 2022)

**Course WebPage:** *Updated Soon*

**Office Hours:** After class, or by appointment

**Course Description**

Natural language processing (NLP) is becoming increasingly widespread. Applications of NLP have become embedded in our everyday lives, and these applications are based somewhere between formal linguistics and statistical physics. Especially over the past decade, neural network approaches have become the de facto standard for many NLP tasks. This course aims to provide a survey of these foundations, but we will take NLP in a narrow sense to cover the text analysis only. The course assumes a background in multivariate calculus, linear algebra, and proficiency in Python. The goal of this course is to enable you to build your language applications using the *PyTorch* framework.

**Assignments:** *Updated Soon*

**Exams:** *Updated Soon*

**Main Reference**

We will use ***Natural Language Processing with PyTorch*** by Delip Rao and Brian McMahan (O’Reilly, 2019) as a main reference. There is a partial list of useful books that will be touched during the course.

* Steven Bird, Ewan Klein, Edward Loper, *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* (**REF1**)
* Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana, *Practical Natural Language Processing* (**REF2**)

You have an **online access** for all the references listed above through Brandeis Library. Other useful reference is *Introduction to Natural Language Processing* by Jacob Eisenstein for avid students of mathematical exposition.

**Prerequisites**

1. Competency in Python (Bus215f)

* All class exercises will be using Python. You should be familiar with NumPy and data structures in Python. Note that you should be fine if you have ample experience in coding with a different language.

1. Calculus, Linear Algebra, Probability, and Statistics (Econ213a)

* You should know college-level calculus and the basics of probabilities

1. Machine Learning

* If you have basic machine learning or deep learning experience, the course would be much easier. You can take it without knowing them. If you need a top-bottom textbook treatment, I highly recommend: “*Hands-on machine learning with scikit-learn and TensorFlow*” by Geron Aurelien

**Course Outline**

The class covers the major applications of neural network in NLP. We will mainly examine the practical use cases and delve into theories where necessary. Each week will be dedicated to one concept. However, some additional concepts would be introduced due to the compact module-class structure. The following outline provides a high-level overview of the course. The mandatory readings are indicated by \*, and you must be prepared before the class meeting.

1. **Introduction (Week 1)**
2. The foundations of the effective modern methods for ML applied to NLP

* Chapter 1 \*
* Chapter 1, REF2 \*
* Turing, Alan M. “*Computing machinery and intelligence.*” Parsing the turing test. Springer, Dordrecht, 2009. 23-65.
* Olah, Christopher. “[*Understanding LSTM Networks.”.*](https://colah.github.io/posts/2015-08-Understanding-LSTMs/index.html) August 27, 2015

1. Python Reviews (***Offered online***)

* Object Oriented Programming: Class
* PyTorch Basics

1. **Language structure (Week 2)**
2. Word Vectors
   * Mikolov et al. “[*Efficient Estimation of Word Representations in Vector Space*](https://arxiv.org/pdf/1301.3781.pdf)” \*
   * Vector space models, Chapter 3 (REF2) \*
3. Use of linguistic data: text corpora and lexical resources

* Chapter 2 \*
* Chomsky hierarchy: <https://people.cs.umass.edu/~mccallum/courses/inlp2007/lect2-regex.ppt.pdf>

1. Tagging

* Chapter 5 (REF1) \*

1. **Introduction to Neural Networks (Week 3)**

* Chapter 3 \*
* Basic Search and Learning model
* Introduction to convergence algorithm
* Supervised gradient-based learning in PyTorch

1. **Multilayer problem and examples (Week 4)**
2. XOR problem

* Introduction to multilayer perceptron
  + Example: Surname classifier model

1. Deep Learning for Text Classification
   * Chapter 4 \*
   * Chapter 4 (REF2)
   * Classifying Surnames by Using convolutional neural network in PyTorch
   * Long short-term memory networks revisit
2. **Bag of Words and Embedding (Week 5)**
3. Introduction to Transfer learning

* Chapter 5 \*
* Encoding and decoding problems

1. Transformer models and Transfer learning in NLP these days
   * Attention Mechanisms
   * Hugging Face Application
2. **Introduction to Recurrent Neural Networks (Week 6)**
3. Sequence Modeling in NLP

* Chapter 6 \*

1. Revisit the Surname classifier model

* Understanding the hidden states

1. Epilogue
   * Review: NLP model pipeline