NLP4Free

A Free Natural Language Processing Microcourse

https://github.com/nlpfromscratch/nlp4free

Part 1 - Introduction to NLP

Myles Harrison,
Founder and Trainer

Agenda

01 Front Matter

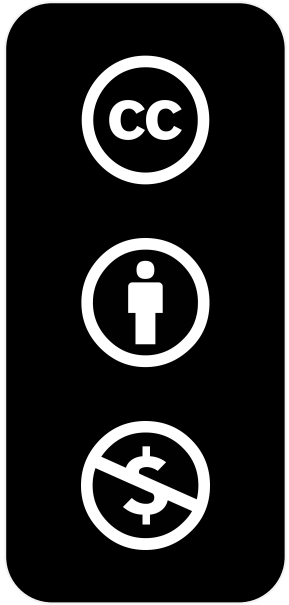
02 Welcome to the Course

03 Course Materials & Topics

04 A Brief Introduction to NLP

05 What's Ahead

Licensing and Usage



Credit Where Due

Though this work is based upon my experiences consulting and teaching data and machine learning, including that for natural language, all materials within have been compiled or created by myself.

When I have included or relied upon others' materials such as images, code, or text, I have done my best to cite as appropriate and provide links to the source.

Welcome!

I'm excited that you've taken an interest in the course. I hope that you will find it valuable as a resource.

The course will cover the fundamentals of natural language processing (NLP), introducing the reader to concepts, tools, and techniques for working with language and machine learning.

While there are technical bits, this is not intended to be a deep comprehensive look at applying NLP techniques, but rather a place to begin for those unfamiliar with the field.

Let's get started.

Image source: https://koa.com/blog/best-locations-to-watch-the-sunrise-in-north-america

Materials & Delivery

Course materials are slides available in PDF format, as well as Jupyter notebooks available both for download and to run in Google Colab, and pre-recorded Youtube videos.

There is no live component nor assessment and all materials are to be reviewed at the pace desired by the reader.

Course Contents

The course will cover NLP at a high level from basics all the way up to advanced techniques using machine learning, deep learning, and touch on generative AI and large language models (LLMs) which represent the current state of the art:

1. Introduction to NLP
2. Acquiring and Preprocessing Text
3. Machine Learning and Sentiment
4. Unsupervised Methods for NLP
5. Deep Learning for Natural Language

What's the deal with this ChatGPT thing?

ChatGPT is an example of a *large language model (LLM)*, a type of deep learning model trained with hundreds of millions or billions of parameters on very large bodies of text. Large language models currently represent the state of the art in NLP.

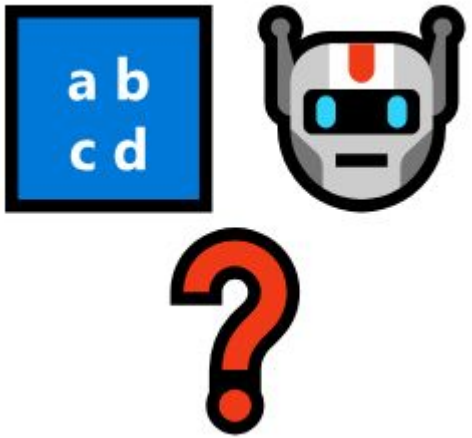
While we're here, ChatGPT is not sentient, nor is it an example of an Artificial General Intelligence (AGI). Let's take a step back…

What is Natural Language Processing (NLP)?

Natural language processing lies at the intersection of the domains of linguistics, computer science, and artificial intelligence.

In this course, we are primarily concerned with NLP as it pertains to the field of data science and AI, in this meaning referring to teaching computers to process - and perhaps even "understand" - text written in ordinary language, and perform associated tasks.

Though the term *processing* usually refers specifically to altering and preparing data, in the domain of AI, NLP is often used to refer more generally to any language problem - including those of applying machine learning (ML) to language - since these still require processing text data beforehand.



Areas of NLP

The field of NLP can be broken down into high level areas and associated tasks, as non-exhaustively shown here.

Some areas are highly specialized and far beyond the scope of this course.

Document Classification

Natural Language Generation

Named Entity Recognition

Machine Translation

Speech Recognition

Sentiment Analysis

Conversational Systems

Text to Speech

Document Summarization

Applications of NLP

Some examples of use cases for natural language processing and machine learning for specific industry verticals are provided here.

Finance

Summarizing earnings reports, financial statements, filings, etc.

Retail

Generative models for copywriting automation

Medicine

Categorizing and classifying free-form clinical notes

Media

Automated captioning of television and films

A Brief History of NLP (according to Wikipedia)

Symbolic
(1950's- 1970's)

Rules-based methods for language tasks such as translation and conversation.

Statistical / ML
(1980's- 2000's)

Advent of statistical techniques and application of machine learning.

Neural
(2000's - Present)

Breakthroughs in deep learning leading to rapid advances in the field up to today.

Sources: 1, 2, 3

Acquiring and Preprocessing Text

This refers to the where and how of getting text data, and also methods and techniques for preparing it for whatever task need be accomplished.

Since all NLP tasks require text data and it to be processed beforehand, this is a foundational area.

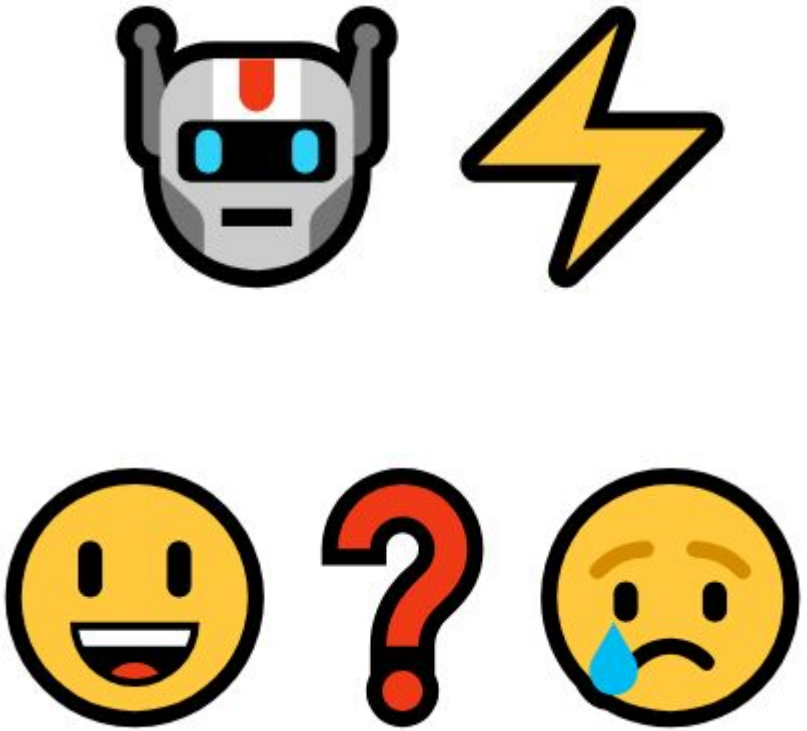
These will be the topics of Part 2.

Machine Learning and Sentiment

Machine learning is the application of statistical methods and algorithms applied to data in order to find patterns, solve problems, or perform tasks.

Sentiment analysis is a subdomain of natural language processing concerned with the emotional tone or content of text.

These topics will be covered together in Part 3 with an applied example.



Unsupervised Methods for NLP

This type of machine learning is not given specific labelling or prediction tasks, and instead works by finding patterns in the data.

Unsupervised learning is very important for making sense of large bodies of text, and can also be used for transformation of data before applying other machine learning methods.

We will cover unsupervised methods including topic modeling and word embeddings in the Part 4.

Deep Learning for Natural Language

Also known as neural networks, this type of machine learning seeks to emulate how the human brain functions, and represents the state of the art for nearly all NLP tasks.

We will introduce the fundamentals of deep learning and move into its applications to language in the final section.

End of Part 1

Introduction to NLP