Introduction to Natural Language Processing

BUS 243 F: Spring 2023

Yeabin Moon

Lecture 1

Lecture Plan

- Each lecture consists of two parts (10 mins break between)
- Part 1
    - The course
    - NLP overview: what is it?
    - NLP application
    - Text representation
- Part 2
    - OOP
    - Regular expression
    - Matrix Algebra

Logistics in Brief

- Instructor: Yeabin Moon
- TA: Jiawei Fan
- Office hours: see the syllabus
- We’ve put a lot of resources on the Latte page
  - Syllabus / Announcement / course materials / homework
- Lecture materials uploaded before each lecture

Learning Goals

- The foundations of the effective / popular methods applied to NLP
  - More focus on basic ideas
    - Vector space models
    - Rule-based approach
    - Semantic analysis in word counts
- The challenges in understanding and producing human languages via computers
  - How to prepare text data for modeling using Python libraries
  - Limitation analysis
- What deep learning is and how it is different from other methods
  - Distributed representation

Course work and grading policy

- We have 6 meetings in total
- 3 x Assignments: 3 x 15 %: 45 %
    - HW1 is released today
    - Check every assignment due in the syllabus
        - 1 % off a day late
        - No accepted more than a day late unless given permission in advance
- 5 in-class quizzes: 5 x 4 %: 20 %
    - 15 mins at the beginning of every class (from next week)
- Final exam: 30 %
    - May 9th, 6:00 pm – 9:00 pm

Note for assignments

- Each assignment should be completed individually
- There are only 3 assignments
  - Strongly encourage you to work alone to gain the maximum
  - Understand allowed collaboration and how to document it
    - Do not take code off from the web
    - Must acknowledge working with other students
    - Must write your own assignment solutions

Why there are quizzes?

- Most of you have not taken any courses regarding NLP
- Readings are mandatory for this course
- Fundamental concepts in readings will be addressed in each quiz
- Main text: Hapke, Lane and Howard, *Natural Language Processing in Action*, 2019, Manning Publications
  - Brandeis students have online access to this book
- Additional readings will be posted in the Latte

How difficult this course would be?

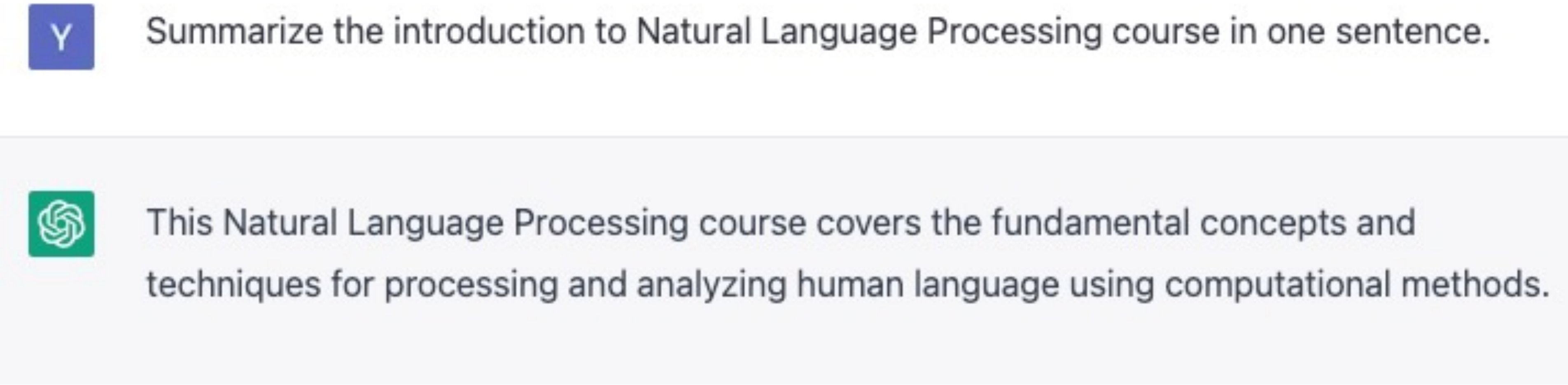
- This is not an introductory course to machine learning or deep learning
- You will not learn how to write neural nets in mathematical terms
- Thanks to modern ML libraries, you don’t really need to understand the math to build practical applications
    - However, I will explain some concepts as needed
- If you are interested in learning the theories and the math behind, that’s enough
- You do need to be at least comfortable enough to write in Python and know its ecosystems

Moneyball (2016)

It's not that hard, Scott. Tell him, Wash... It's incredibly hard. Hey, anything worth doing is!

Let’s Begin

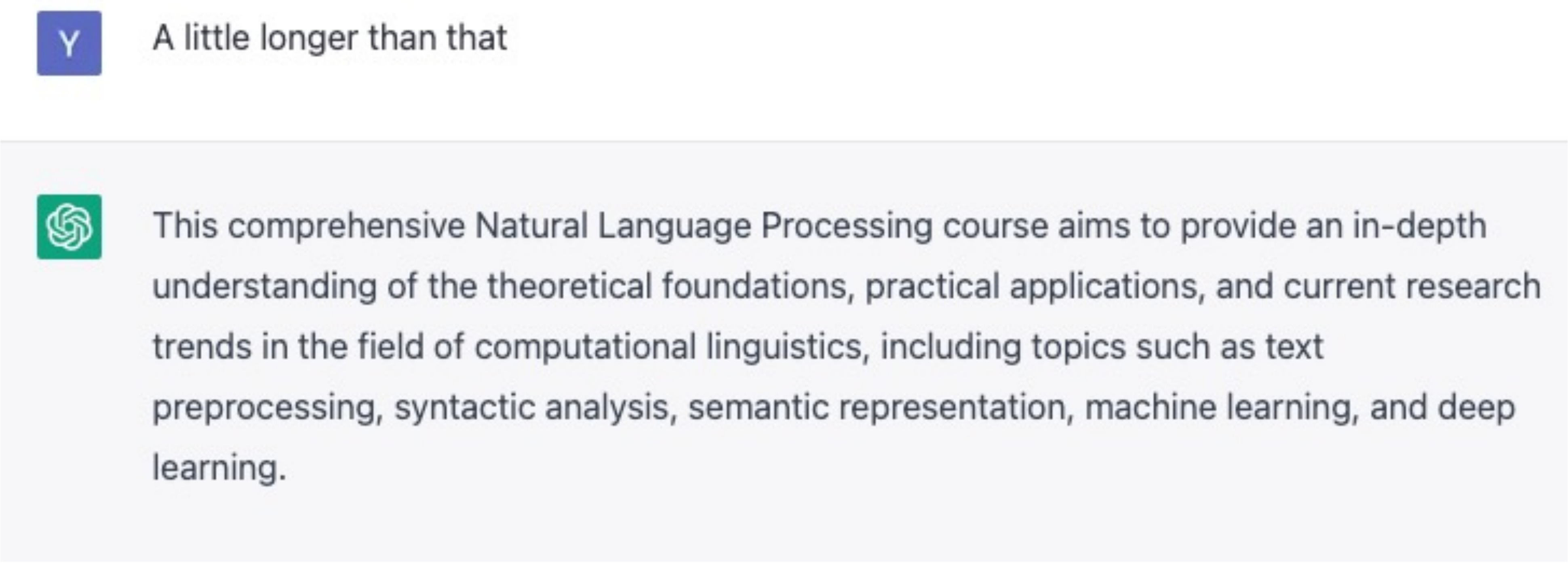
ChatGPT: Hmm…



Summarize the introduction to Natural Language Processing course in one sentence.

This Natural Language Processing course covers the fundamental concepts and techniques for processing and analyzing human language using computational methods.

ChatGPT: More amazing features would be



A little longer than that

This comprehensive Natural Language Processing course aims to provide an in-depth understanding of the theoretical foundations, practical applications, and current research trends in the field of computational linguistics, including topics such as text preprocessing, syntactic analysis, semantic representation, machine learning, and deep learning.

What is natural language processing (NLP)?

- NLP is a principled approach to processing human language
- Formally, it is a subfield of artificial intelligence (AI) that refers to computational approaches to process, understand, and generate human language
    - language processing is considered a huge part of human intelligence
- It might indicate a range of algorithms taking human-produced text as an input and producing some useful information

What does it mean for a language to be natural?

- You might wonder
  - Are there any unnatural languages?
  - Is English natural?
  - Is Spanish more natural than Korean?
- Another tricky term is a formal language
  - Is English formal?

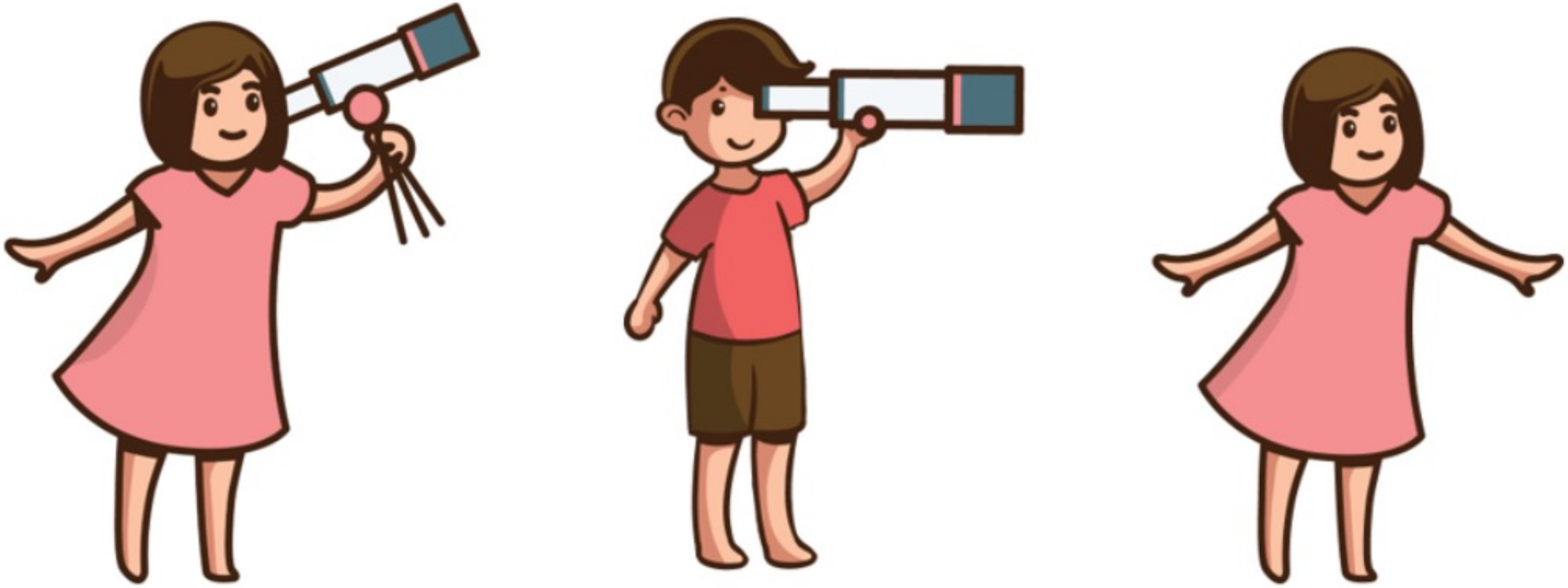
Natural vs. Formal

- The word *natural* is used to contrast natural languages with formal languages
  - all the languages humans speak are natural
- On the other hand, formal languages are types of languages that are invented by humans
  - Have strictly and explicitly defined syntax (grammatical rules) and semantics (meaning)
  - Programming languages are examples
  - When you run a compiler or an interpreter on the code you write in those languages, you either get a syntax error or not
  - the behavior of your program is always the same if it’s run on the same code

Natural language is hardly formal

- You can write a sentence that is *maybe* grammatical
    - Subjective, and worse, time varying
    - There are some grammar topics where even experts disagree with each other
- This is what makes human languages interesting but challenging, and why the entire field of NLP even exists
- Bottom line: human languages are ambiguous

He saw a girl with a telescope



Source of ambiguity

- Confused because you don’t know what “with a telescope” is about
  - Don’t know what this prepositional phrase (PP) modifies
  - PP-attachment problem is a classic example of *syntactic ambiguity*
- A syntactically ambiguous sentence has more than one interpretation of how the sentence is structured
  - interpret the sentence in multiple ways
- Another type is *semantic ambiguity*
  - *I saw a bat*
  - Meaning of a word is ambiguous, not its structure
    - Unless you are brutal enough

Welcome Remark

- Ambiguity is what makes natural languages rich but also challenging to process
- We can’t simply run a compiler or an interpreter on a piece of text and just *get it*
- We need to face the complexities and subtleties of human languages
- We need a scientific, principled approach to deal with them

Here comes a new challenger

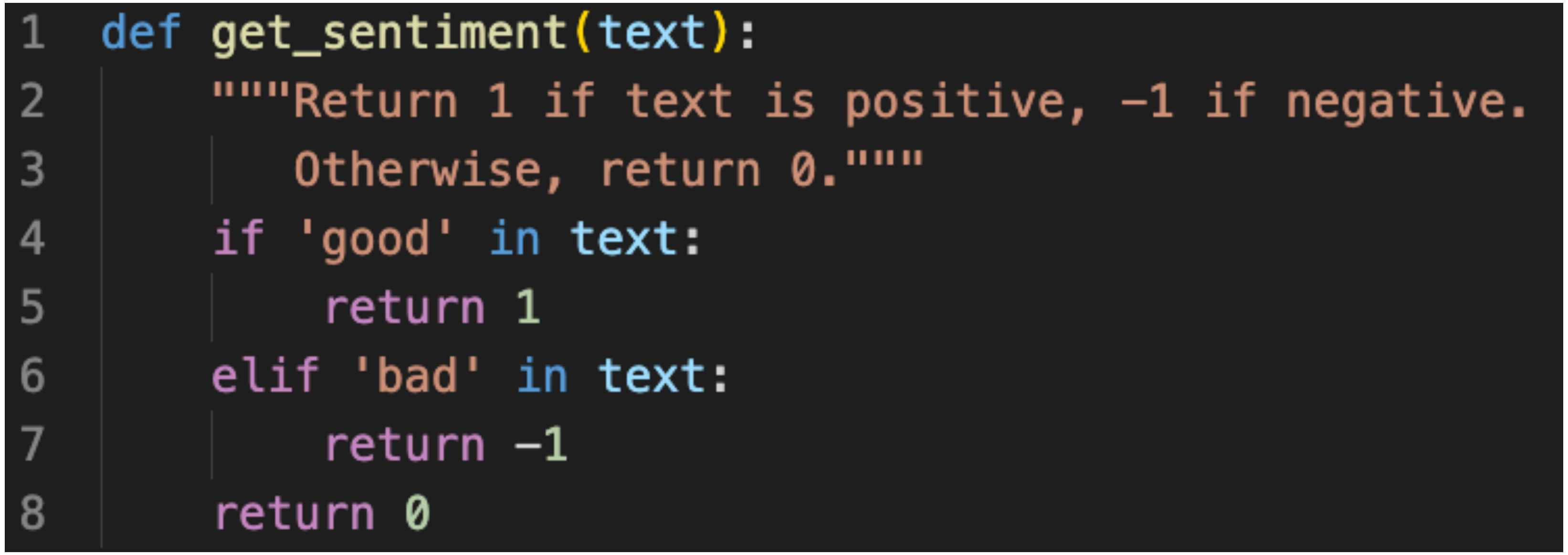
- Now let’s consider the following scenario and think how you’d approach this problem
- You are working as a junior data scientist at a midsized company that has a consumer-facing product line
- You got a giant TSV file containing all the responses to the survey questions about the product from the marketing team:
    1. How did you know about our product?
    2. How do you like our product?
    3. a free-response question, where our customers can write whatever they feel about our product

Here comes a new challenger

- The marketing team realized there was a bug in the online system and the answers to the second question were not recorded in the database at all
- Your task is whether you could recover the lost data
- Fortunately, data structure is fairly standard
    - It has several fields such as timestamps and submission IDs
    - At the end of each line is a lengthy field for the free-response question

First try

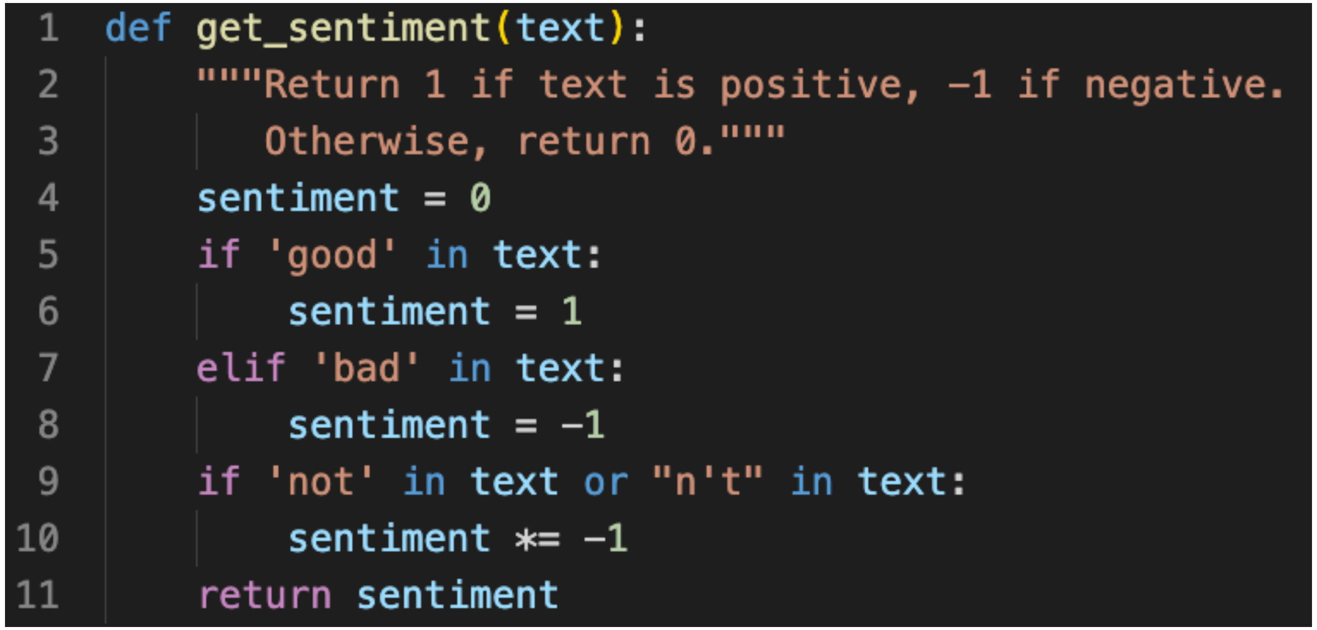
- Example responses:
  - A very good product!
  - Very bad. It crashes all the time!



```python
def get_sentiment(text):
    """Return 1 if text is positive, -1 if negative.
       Otherwise, return 0."""
    if 'good' in text:
        return 1
    elif 'bad' in text:
        return -1
    return 0
```

Again, natural language is ambiguous

- The code filtered a decent amount of data
- Alas, my code returns
  - I can’t think of a single good reason to use this product: positive
  - Not bad: negative
- Right. Negation!



```python
def get_sentiment(text):
    """Return 1 if text is positive, -1 if negative.
       Otherwise, return 0."""
    sentiment = 0
    if 'good' in text:
        sentiment = 1
    elif 'bad' in text:
        sentiment = -1
    if 'not' in text or "n't" in text:
        sentiment *= -1
    return sentiment
```

….

- The product is not only cheap but also very good!: negative
- Worse
  - I always wanted this feature badly!
  - It's very badly made
- How could a single word in a language have two completely opposite meanings?
- This course will save you

This course will save you?

- What does it mean? Another ambiguity

- This course will save you because it will teach you either

  1. How to deal with the problems described above

  2. They are impossible to solve

Rule-base vs. DL approach

- As you saw, a bunch of *if*s and *then*s would mitigate the issues
  - It is a rule-based approach
    - controversial definition of NLP
- You may also have heard of deep learning
  - a subfield of machine learning that usually uses *deep* neural networks
- As the amount of available data and computational resources increases, modern NLP makes a heavier and heavier use of machine learning and deep learning

Scope of the class

- Traditional methods such as counting words and measuring similarities between text are usually not considered to be machine learning techniques per se, although they can be important building blocks for ML-based models
- Let’s see the syllabus
- By the way, we only focus on English text documents and messages

We are not going to study

- Other languages
- Spoken statements
- Text generation
- Dialogue system

Readings

- Almost every NLP textbook is outdated
- For the beginners, class with a dedicated textbook is often helpful
- The textbook is somewhat chat-bot-application oriented
  - However, most contents are useful
  - Recommend reading the assigned chapters top-to-bottom

Class overview

- Besides an application, we will study Text Representation
- The overall theme is to examine how the machine understands text
- Roughly two ideas
  - Word counts
  - Locations (distributed representation)

Text representation

- Consider how to classify the documents
  - Find the most frequent / unique words
    - Surprisingly high accuracy
  - What’s the potential problems?
    - Words have a range of forms
    - Synonym?
  - What does the model mean in NLP?
    - Why do we ever need a neural net?

Road Map

- Count-based Text representation
  - Bag-of-words approach
  - Examine tokenizers
- Information retrieval
  - TF-IDF
  - PCA analysis
- Distributed representation
  - Find a meaning from the word usages
    - Word order matters

Google Colab

- Some codes in the textbook are outdated
  - Do not work well
- I will post the modified codes on Latte
- Use Google Colab to run them
- Fine to work on your machine if you know how to manage conda environment