Introduction to Computational Text Analysis

Jae Yeon Kim

10 April, 2019

Motivation

- Misplaced hope and fear, confidence and skepticism
- Demystifying computational text analysis and machine learning
- Learn **basic** theories and techniques at the same time

What Is Language?

- Rationalist approach
- Empiricist approach
- Computational approach

What is NLP?

- It’s everywhere.
- It’s evolving.
- It has limitations.

The challenge of big data

- N of N samples $<$ P of P features
- High-dimensional data
- Pervasive problem across text, sound, and image data

Language processing

- Understanding
    - Analyzing
    - Representation
- Words -> Discourse
- Middle steps: Morphology, Syntax, Semantics

Preprocessing

- Tokenization: spiting lines of texts into the most basic units (n-grams)
- Removing stop words and other special characters among those units
- Normalization: standardizing those units (e.g., lemmatization)

Computational text analysis

- NLP + domain knowledge
- Dictionary-based methods
- Unsupervised machine learning (e.g., topic modeling)
- Supervised machine learning (e.g., text classification)