

# Homework 1

# Learning Goals

- Get basic infrastructure (Anaconda, environment) set up for this course
- Build the very first building block for our NLP models: a Vocabulary
- Reflect on dataset documentation, using data that we will use throughout the course

# 1. Installing Anaconda

- Anaconda lets you manage local environments for python and other tools
  - Avoid version conflicts across multiple projects
  - Get exactly the versions of packages you need
  - Helps reproducibility as well
- We've provided an environment in ``/dropbox/22-23/575j/env``
- Install:
  - `wget https://repo.anaconda.com/archive/Anaconda3-2021.11-Linux-x86_64.sh`
  - `sh Anaconda3-2021.11-Linux-x86_64.sh`
- `run_hw1.sh` shows you how to activate the environment

## 2. Implementing a Vocabulary

- At the base of every NLP system is a `Vocabulary` object, containing:
  - Token  $\rightarrow$  index
  - Index  $\rightarrow$  token
  - These provide the interface between strings (tokens), and integer indices that will be used in our models (e.g. for looking up embeddings)
- `/dropbox/22-23/575j/hw1/vocabulary.py`
  - `#TODO:` comments tell you where to write your own code
- Write small script to save various vocabularies from the SST dataset (see next slide)

# 3. Data Statement for SST

- For many assignments in this course, we will be using the [Stanford Sentiment Treebank](#)
  - Input: movie reviews
  - Output: discrete ratings (0-4) of the sentiment from very negative to very positive
  - Simple/cleaned version available in `/dropbox/22-23/575j/data/sst/`
- [Data Statements for NLP](#) (Emily M Bender and Batya Friedman)
  - Best practices for documenting dataset creation
    - Can help understand and mitigate biased models by clearly identifying the nature and source of the data (e.g. which populations)
  - For this assignment: answer (to the best of your ability, given the documentation of SST) the relevant questions that should go into a data statement
  - Note: also see updated schema here: <http://techpolicylab.uw.edu/data-statements/>