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/23-24/574/env`
  - **NB:** will soon, homework is **due April 6**
- Install:
  - `wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh`
  - `sh Miniconda3-latest-Linux-x86_64.sh`
- `/dropbox/23-24/574/hw1/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 —> index
  - Index —> 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/23-24/574/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/23-24/574/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
  - NB: also see updated schema here: http://techpolicylab.uw.edu/data-statements/