Sentiment Analysis using python:

Data Overview:

* Dataset of 50,000 movies review taken from IMDb
* Data is evenly split into 25k each for training and testing, moreover, each set has 12.5k positive and 12.5k negative
* IMDb let users to rate the movie on a scale of 1 to 10
* Data label anything with =< 4 is negative review, >= 7 is positive review whereas 5 and 6 are left out

**Step 1: Download and Combine Movie Reviews**

If you haven’t yet, go to [IMDb Reviews](http://ai.stanford.edu/~amaas/data/sentiment/) and click on “Large Movie Review Dataset v1.0”. Once that is complete you’ll have a file called aclImdb\_v1.tar.gz in your downloads folder.

**Shortcut:**If you want to get straight to the data analysis and/or aren’t super comfortable with the terminal, I’ve put a tar file of the final directory that this step creates here: [Merged Movie Data](https://github.com/aaronkub/machine-learning-examples/blob/master/imdb-sentiment-analysis/movie_data.tar.gz). Double clicking this file should be sufficient to unpack it (at least on a Mac), otherwise gunzip -c movie\_data.tar.gz | tar xopf — in a terminal will do it.

**Unpacking and Merging**

Follow these steps or run the shell script here: [Preprocessing Script](https://github.com/aaronkub/machine-learning-examples/blob/master/imdb-sentiment-analysis/preprocess_reviews.sh)

1. Move the tar file to the directory where you want this data to be stored.
2. Open a terminal window and cd to the directory that you put aclImdb\_v1.tar.gz in.
3. gunzip -c aclImdb\_v1.tar.gz | tar xopf -
4. cd aclImdb && mkdir movie\_data
5. for split in train test; do for sentiment in pos neg; do for file in $split/$sentiment/\*; do cat $file >> movie\_data/full\_${split}.txt; echo >> movie\_data/full\_${split}.txt; done; done; done;

### Step 2: Read into Python

For most of what we want to do in this walkthrough we’ll only need our reviews to be in a Python list. Make sure to point open to the directory where you put the movie data.

|  |
| --- |
| reviews\_train = [] |
|  | for line in open('../data/movie\_data/full\_train.txt', 'r'): |
|  | reviews\_train.append(line.strip()) |
|  |  |
|  | reviews\_test = [] |
|  | for line in open('../data/movie\_data/full\_test.txt', 'r'): |
|  | reviews\_test.append(line.strip()) |