Homework 1

# Part 1

Install Python 3.6 or higher and the following libraries we will be using throughout this class:

* Numpy
* Pandas
* Scikit Learn (import as “sklearn” in python files)
* NLTK
* Pytorch (package is named “torch” for installation)

This list is preliminary and you may need to install additional libraries throughout the semester, I just want to make sure everyone knows how to install Python libraries.

# Part 2

Write a simple tokenize function using **regular expressions** in Python (re.findall) to pre-process the provided dataset containing Amazon appliance reviews in JSON format (see Canvas).

I have provided code for the reading in of the file into a list of documents, you will be in charge of coding everything else. You **can** use the code from Friday’s class as a starting point.

Special characters are shown in between square brackets for clarification, to indicate they are not part of the sentence. The square brackets themselves should **NOT** be matched. For example, [!] refers to the exclamation point character.

Make sure your tokenizer can handle the following situations.

* Treat all punctuation as their own tokens UNLESS they are in the middle of a number (1,000), or part of a word (see below).
* Instead of removing punctuation, remove all types of brackets.
* HINT: Look at how we removed punctuation in class
* Allows tokens to consist of both letters and integers. For example: "24cm", "32ft", …
* Allow contractions and possessives where ['] appears in the middle of the word and ≤ 3 letters appear after [']. For example: can't, you're, I'm, Marvin's, would've, …
* Allow [-] (dash) to appear in the middle of a token
* HINT: Look at how we were able to leave periods and commas in numbers

Upload your updated tokenizer to Canvas as Lastname\_firstname\_HW1.py