Problem Set 3: Rule-based systems for language and categorization
CS328, Fall 2018, Anna Rafferty

For this problem set, you'll complete Problem 1 with your same partner from last time or alone if you completed the last problem set alone, unless I've confirmed a change in partnerships with you via email. For problem 1,if you're in a partnership, you should work on all parts together, and use paired programming, trading off who is controlling the keyboard and mouse. You are welcome to talk to others outside of your partner about the problems on this problem set, but the write up must be your (and if applicable, your partner's) own work. See the handout on homework to find out more details. You'll submit problem 1 under "Homework 3 - problem 1" on Moodle. If you work with a partner, only one of you should turn in your submission on Moodle.

Problems 2 and 3 do not involve coding and you will turn in an individual submission for them, but you are welcome to discuss these problems with others. You should **read and attempt each problem prior to discussing it with others.** The write up must be your own work. That means that **you should write your answers by yourself, without looking at notes from sessions with collaborators, and you must be able to explain any answer you write down.** You'll submit these problems under "Homework 3 - problems 2 and 3" on Moodle. When you fill out the collaboration form on Moodle, you'll note any collaborations for each individual problem.

Due date: All questions are due at 10PM on Friday, 28 September 2018. For problem 1, please submit a zip file with your `CS328-PS3-AnimalRules.ipynb` notebook and your python module. For problems 2 and 3, please submit a zip file with your `CS328-PS3-Learnability.ipynb` and the PDFs described in the final question. Within each zip file, please also include an HTML export of each notebook; you can produce this through File→Download as...→HTML or through File→Print preview. It's okay if some markup doesn't show correctly in the HTML version but shows up in your notebook - I just want a quick way to access your submissions outside of jupyter.
Starter files: The starter files for this assignment are available here.

1. (Categorization and induction) In this question, you'll be working with a dataset containing 50 animals and 85 features.[1] Features are things like whether an animal swims or whether it lives in the forest. Each animal-feature pair was rated by 8-10 people as to how well that feature reflected the animal. For example, "swims" might be highly rated for otter, but rated lower for giraffe. The people's ratings were averaged, and the matrix you've been given has been turned into a binary form: each feature-animal pair is a 0 or a 1. The continuous ratings were turned into the binary form using a threshold; e.g., a threshold of 0.5 would mean all animal-features pairs above 0.5 would be set to 1, and all those below 0.5 would be set to 0. Open the notebook `CS328-PS3-AnimalRules.ipynb` and follow the directions in that notebook.

2. (Learnability of language) In this question, you'll be considering formal models of learnability. Open the notebook `CS328-PS3-Learnability.ipynb` and follow the directions in that notebook.

3. We've started doing a little bit of probability in class already, and we'll be doing much more probability when we getting into the section on probabilistic models. To help you review your probability, we'll be using a section from CMU's Open Learner Initiative course on probability and statistics. You can access this section here: https://oli.cmu.edu/jcourse/webui/guest/join.do?section=probstat, and either create an account (if you want it to save your work) or click on Enter Course to enter without an account. We'll be working from UNIT 4: Probability.[2] You may find the review relatively quick if you've done lots of probability, or slower if it's been a while since you completed Math of CS. Please come to office hours and ask questions if you find it especially confusing. Work through the following pages, and include the information noted below in your submission for this assignment:

    - Module 8: Introduction (Probability): pages 89-96. Include in your submission a PDF named "Intro to Probability Checkpoint.pdf" that is a printout of the results of your completion of the Intro to Probability Checkpoint.
    - Module 9: Finding Probability of Events: pages 98-105. Include in your submission pdfs of each page that are printouts after you've done the interactive activities on pages 103-105 (so three PDFs total). Name each pdf as follows: ProbEvents103.pdf, where you replace 103 with the page number indicating which page it's from.

[1] See https://cvml.ist.ac.at/AwA2/ for more details and a fuller version of this dataset. There is also a readme in the data directory that tells you more if you're interested.

[2] Note that some of the examples in these units have an unnecessarily binary view of gender. I'm having you work through these activities because I think they're a good way to get up to speed with probability if you've forgotten some of it and I'm unaware of similar resources elsewhere, but I wish they had included more inclusive examples and additional resources are something I hope to find and incorporate into future versions of this class.