1. Assume P is a distribution over a finite set S. Then

$$H(P) = \sum_{s \in S} p(s) \log(1/p(s)) \tag{1}$$

Here is a list of items: [1,2,1,1,2,3,1,1,2] Write a Python function that determines the entropy.

   - In Info Gain doc on Github

2. What does it mean when $H(P) = 0$?

   - H(P) = 0 means the distribution P conveys no information. This indicates that the outcome of P is certain. This is the case if p(s) = 0 or p(s) = 1.

3. When computing entropy, we allow $\log 1/0 \to 0$–why?

   - As p(s) approaches 0, $p(s) * \log(1/p(s))$ approaches 0, even though $\log(1/p(s))$ is undefined.

4. When does H reach a maximum value–use calculus to determine this.

   - $H'(P) = \frac{d}{dp(s)} \sum_{s \in S} p(s) \log(1/p(s))$
   - $H'(P) = \sum_{s \in S} \frac{d}{dp(s)} (p(s) \log(1/p(s)))$
   - $H'(P) = \sum_{s \in S} -\log(p(s)) - 1$
   - $0 = \sum_{s \in S} -\log(p(s)) - 1$
   - For $\log_2$ and a set S of cardinality 2, the P distribution that satisfies the equation is [0.5,0.5]

5. Let X = [[1,1],[1,2],[1,1],[2,1],[2,3],[2,3],[2,3]]. Find the information gain using the first element of the list as the "splitting attribute".

   - In Info Gain doc on Github

6. Generally entropy (machine learning) is defined using random variables. Let AB be joint discrete random variables over some sample space Ω. If the distributions are independent, what is the entropy?

$$\sum_{a \in \Omega} \sum_{b \in \Omega} p(a, b) \log(1/p(a, b)) \tag{2}$$

7. What is the conditional probability of X, given the first element of the list, what is the probability of the second?

	(x,1)	(x,2)	(x,3)
(1,y)	2/3	1/3	0/3
(2,y)	1/4	0/4	3/4

8. How is Baye's Theorem related to entropy?

    - You can use Bayes' Theorem to classify data. Given a set of attributes $\{a_1, a_2, ...a_n\}$, you can classify each object into class Y_i. The method is to select the class that maximizes $P(Y = i|\{a_1, a_2, ...a_n\})$.
    - Similarly, we use entropy calculations to repeatedly classify our data in set S by finding attribute A with a distribution $\{a_1, a_2, ...a_n\}$ that maximizes $(InformationGain(S, A)|\{a_1, a_2, ...a_n\})$.

9. X = [[1,2,1],[1,3,1],[2,3,0],[2,1,0],[2,2,1],[1,1,1],[1,1,0]]. The "attributes" are the first list members and class the last. Build the decision tree for this using entropy.

    - In progress...

10. Why is a tree pruned?

    - A tree is pruned to prevent overfitting. To prune a node, you remove the part of the tree rooted at that node and make it a leaf node. You select nodes to prune only if the hypothesis performs no worse if the node is pruned (Mitchell 69).

11. What does overfitting mean?

    - A hypothesis is overfit if there is a different hypothesis that performs less well on the training instances, but performs better on a set of test instances. This can happen when the training data contains noise and the ID3 algorithm constructs a more complex tree to classify this noisy data. Accuracy of a hypothesis on test data tends to increase with the size of the tree up to a certain point and then may decline. This decline is evidence that the hypothesis overfits the training examples (Mitchell 67).

12. Why do you suppose entropy uses log? It's not random by the way.

    - We want the entropy function to yield 0 when the p(s) is 1. log is the only function that fits that criteria.