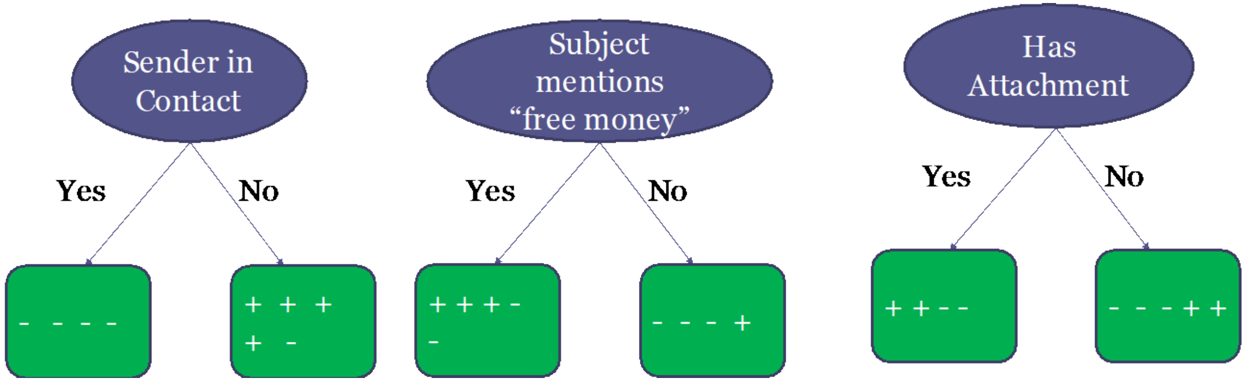
DS 200 Introduction to Data Sciences
Fall 2018

Homework 3 (40 points)

Due Oct 15, 10 pm

Name: ____

(1) (20 points) The figure below shows three possible feature test for the root node of a decision tree to predict spam e-mail messages (based on the example discussed in class). (a) Calculate the expected information gain for each feature test (please show your formula). (b) Which feature test will be selected by the Decision Tree Learning algorithm?



(2) (20 points) Construct a regular expression as the value for the parameter token_pattern so that CountVectorizer can extract hashtags, twitter user names (e.g., @realDonalTrump), and words from tweets as tokens. Explain how you construct the regular expression.