Weekly Homework 3

Dalton Rothenberger

1. A relevant feature in the context of this paper was a hexadecimal n-gram that provided information towards classifying if a given application was malicious or not. To determine whether a feature was relevant or not, the researchers in this paper used information gain to determine which hexadecimal n-grams were the most relevant. They chose relevant features based off the ones that provided the most information gain. I agree with this methodology because the n-grams that provide the most information gain are ones that are unique in comparison to an n-gram that might appear in every application provide less information. This makes a unique n-grams more of a tell tale sign of what classification the application might fall into. In the case of applications a given hex sequence might always appear in Trojans but not normal applications.
2. Yes, I believe true positive and false positives have meanings in multi-class settings. For example, if you had 3 potential diseases to diagnose a patient with such as diabetes, cancer, and Crohn's disease. A true positive in this case would occur when a patient with cancer is properly diagnosed to have cancer. A false positive would occur when a patient is diagnosed with cancer with cancer but actually have diabetes, Crohn's, or no disease.