In this assignment, we will study the URLs and Packets and explore various data mining concepts in relation to the URL. The practical exercises (2-3) will be assigned by Friday and will be due in approximately two weeks from the day they are assigned.

1. Get the Iris flower dataset from UCI repository and identify any correlations exceeding 0.5 between any pair of variables.

2. What are the parts of a URL? Explain with a rich enough URL that has at least three different parts.

3. Why would URL investigation help in security/privacy? Give at least three different applications from security/privacy in which it could be useful.

4. How could a URL be represented as a market-basket type of transaction? Give at least three different ways.

5. How could URLs be clustered? Give at least three different distance functions for URLs. How many of the distance functions you listed satisfy the definition of a metric? Prove or disprove for each whether it is a metric or not.

6. Does K-means assume that the distance function is a metric?

7. Repeat Questions 1 through 4 for a TCP/IP Packet with a small modification to 4, prove or disprove for only one distance function that you provide for packets.

8. The problem with association rules is that if the min_support threshold is too low, then too many itemsets are frequent and if it is too high, then too few itemsets are frequent. How may we set this threshold in practice? Does changing the min_confidence threshold affect the number of frequent itemsets? Why or why not? Explain carefully.

9. Reading exercise: study conditional independence of random variables. For the Iris flower data set from UCI repository, first discretize independently the range of values for each attribute into three categories (call them small, medium and large) representing intervals of equal size and then check whether petal length is conditionally independent of the sepal width when given the sepal length.

Academic Honesty Policy: No collaboration with anyone or anything in or outside the course is allowed on any homeworks, exams and programming assignments (yes, that excludes the internet as well) except if it is explicitly allowed on a problem. The appropriate help of the instructor and (if applicable) the TA is of course allowed and encouraged. Special Note for this assignment: Parts of this assignment specifically allow you to use the Internet, but it is still prohibited for the parts that do not explicitly ask you to use it.