CS 412: Introduction to Data Mining	**Summer 2017**
Homework 1	
Handed Out: June 5, 2017	*Due: June* 15, 2017 *11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Compass (http://compass2g.illinois.edu) for collecting this assignment. Contact TAs if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The homework MUST be submitted in pdf format. Handwritten answers are not acceptable. Name your pdf file as YourNetid-HW1.pdf
- You need to explain the logic of your answer/result for every question. A result/answer without any explanation will not receive any points.
- It is OK to discuss the problems with the TAs and your classmates, however, it is NOT OK to work together or share code. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the CS Honor code (http://cs.illinois.edu/academics/honor-code) on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.
- Please use Piazza if you have questions about the homework. Also feel free to send TAs emails and come to office hours.

2 Question 1 (15 points)

Given a dataset "scores.txt" (contained in data.zip), which includes the records of students exam scores (sample from the population) for the past few years of an online course. The first column is the student id, the second column is the mid-term scores, and the third column is the final scores. The columns are separated by tab. Based on the dataset, give out the following statistical description of data. If the result is not integer, then round it to 2 decimal places.

1. (9') Calculate the following for the midterm scores:
   (a) (3') min, max
   (b) (3') Q1, median, Q3
   (c) (3') mean, std
2. (6') Draw the Q-Q plot of the midterm scores and final scores. Do midterm scores tend to be lower or higher than final scores?

		J Sainsbury	
		0	1
King Kullen	0	43	31
	1	19	107

Table 1: Item supplement summary

3 Question 2 (15 points)

Given the inventories of two supermarkets King Kullen (KK) and J Sainsbury (JS) in "inventories.txt" (contained in data.zip), compare the similarity between this two supermarkets by using the different proximity measures. if the result is not integer, then round it to 2 decimal places.

1. (5) Given 200 items, the following table summarizes how many items are supplied by corresponding supermarket in Table 1. In Table 1, for $KK = 0$, $JS = 0$, it corresponds the number of items among the 200 items that are served neither by KK nor JS. For $KK = 1$, $JS = 0$, it corresponds the number of items among the 200 items that are served by KK but not JS. So on and so forth. Based on Table 1, calculate the Jaccard coefficent of J Sainsbury and King Kullen.
2. (5') Treat the counts of the 100 items as a feature vector of the two supermarkets. Calculate the Minkowski distance between the two vectors with regard to different h values:
   (a) h = 1,
   (b) h = 2,
   (c) h = ∞.
3. (5') Calculate the cosine similarity between J Sainsbury and King Kullen with regard to the feature vectors.

4 Question 3 (10 points)

Consider the student score file "score.txt" (contained in data.zip). Normalize the midterm scores using z-score normalization.

1. (5') Compare the sample mean and std before and after normalization.
2. (5') Given original score of 90, what is the corresponding score after normalization?