1. Consider the problem of classifying a name as being Food or Beverage.
Assume the following training set:

- D1 Food: “turkey stuffing”
- D2 Food: “buffalo wings”
- D3 Beverage: “cream soda”
- D4 Beverage: “orange soda”

1. Apply kNN with k=3 to classify a new name:

- D5(Q) “turkey soda”

Use tf without idf, with cosine similarity. Would the result be the same if k=1? Why?

Solution:

	buffalo	cream	orange	soda	stuffing	turkey	wings	length
D1	0	0	0	0	1	1	0	sqrt(2)
D2	1	0	0	0	0	0	1	sqrt(2)
D3	0	1	0	1	0	0	0	sqrt(2)
D4	0	0	1	1	0	0	0	sqrt(2)
D5(Q)	0	0	0	1	0	1	0	sqrt(2)

sim(D1,Q) = 1/2
sim(D2,Q) = 0
sim(D3,Q) = 1/2
sim(D4, Q) = 1/2

if k=3 the neighbors are D1, D3, D4 of classes Food, Bevrage, Beverage;
therefore the class for the new document D5 is Beverage

if k=1 the class of D5 depends on how we solve ties.

2. For the previous training data, apply the Rocchio algorithm to classify a new name:

- “turkey soda”

Solution:

The prototype for class Food is P1 = D1 + D2 = <1,0,0,0,1,1,1>
and for the class Beverage P2 = D3 + D4 = <0,1,1,2,0,0,0>

sim(P1,Q) = 1 / (sqrt(4) sqrt(2)) = 1 / sqrt(8)
sim(P2,Q) = 2 / (sqrt(6) sqrt(2)) = 1 / sqrt(3)

=> Q in class Beverage because it is closer to P2

3. Cluster to following documents using K-means with K=2 and cosine similarity.

- D1: “go monster go”
- D2: “go karting”
- D3: “karting monster”
- D4: “monster monster”

Assume D1 and D3 are chosen as initial seeds. Use tf (no idf). Show the clusters and their centroids for each iteration. The algorithm should converge after 2 iterations.

Solution:

	go	karting	moster	length
D1	2	0	1	sqrt(5)
D2	1	1	0	sqrt(2)
D3	0	1	1	sqrt(2)
D4	0	0	2	sqrt(4) = 2

Iteration 1:

C1 = D1 = <2, 0, 1>
C2 = D3 = <0, 1, 1>

sim(C1,D1) = 1	>	sim(C2,D1) = 1 / sqrt(10)	=> D1 in cluster C1
sim(C1,D2) = 2 / sqrt(10) = 0.63	>	sim(C2,D2) = 1 / 2	=> D2 in cluster C1
sim(C1,D3) = 1 / sqrt(10)	<	sim(C2,D3) = 1	=> D3 in cluster C2
sim(C1,D4) = 2 / (2 sqrt(5))	<	sim(C2,D4) = 2 /(2 sqrt(2))	=> D4 in cluster C2

Iteration 2:

C1 = (D1 + D2) / 2 = <3/2, 1/2, 1/2> length(C1) = sqrt(11) / 2
C2 = (D3 + D4) / 2 = < 0, 1/2, 3/2> length(C2) = sqrt(10) / 2

sim(C1,D1) = (3+1/2)/(sqrt(5)sqrt(11)/2) = 7/sqrt(55)
> sim(C2,D1) = (3/2) /(sqrt(5)sqrt(10)/2) = 3/sqrt(50) => D1 in cluster C1

sim(C1,D2) = 4 / sqrt(22)	>	sim(C2,D2) = 1 / sqrt(20)	=> D2 in cluster C1
sim(C1,D3) = 2 / sqrt(22)	<	sim(C2,D3) = 4 / sqrt(20)	=> D3 in cluster C2
sim(C1,D4) = 1 / sqrt(11)	<	sim(C2,D4) = 3 / sqrt(10)	=> D4 in cluster C2

No changes in cluster assignment => Convergence