1. Consider the table of term frequencies for 3 documents Doc1, Doc2, Doc3. Compute the **tf-idf** weights for the terms car, auto, insurance, best, for each document. The size of the collection is 806,791 documents.

term	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
Insurance	0	33	29
best	14	0	17

term	df_t
car	18,165
auto	6723
insurance	19,241
best	25,235

Sol:

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

terms	Doc1	Doc2	Doc3
Car	44.55	6.6	39.6
Auto	6.24	68.64	0
Insurance	0	53.46	46.98
Best	21	0	25.5

2.

Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1 R N R N N N N N R R

System 2 N R N N R R R N N N

a. What is the **MAP** of each system? Which has a higher MAP?

b. Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?

Sol:

a.

MAP (System 1) = (1/4)*(1+(2/3)+(3/9)+(4/10)) = 0.6

MAP(System 2) =(1/4)*(1/2 +2/5+3/6+4/7)=0.493

System1 has a higher average precision

b. MAP provides a single figure measure of quality across recall levels. For a good MAP score, it is essential to more relevant documents in the first few (3-5) retrieved ones.

3.

Suppose that a user's initial query is cheap CDs cheap DVDs extremely cheap CDs. The user examines two documents, d1 and d2. She judges d1, with the content CDs cheap software cheap CDs relevant and d2 with content cheap thrills DVDs non-relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using **Rocchio relevance feedback**, what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

Sol:

word	q	d1	d2
CDs	2	2	0
cheap	3	2	1
DVDs	1	0	1
extremely	1	0	0
software	0	1	0
thrills	0	0	1

For $1.0*q+0.75*d_1 + 1 -0.25 * d_2$, we get: $(3.5\ 4.25\ 0.75\ 1\ 0.75\ -0.25)^T$ or $(7/2\ 17/4\ 3/4\ 1\ 3/4\ -1/4)^T$. Negative weights are set to 0. The Rocchio vector thus is: $(3.5\ 4.25\ 0.75\ 1\ 0.75\ 0)^T$.

4.

Calculate Kappa Value between two judges.

		Judge 2 Relevance		
		Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Observed proportion of the times the judges agreed
$P(A) = (300 + 70)/400 = 370/400 = 0.925$
Pooled marginals
$P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$
$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$
Probability that the two judges agreed by chance
$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$
Kappa statistic
$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$