1. Consider the set of 6 documents (5 for training and 1 for testing).

docID	**words in document**	**in class = cricket ?**
1	*sachin dravid sachin*	Yes
2	*dravid federer schools*	No
3	*sachin sachin neville*	Yes
4	*sachin drive*	Yes
5	*saina sushil sachin*	No
6 (test set)	*sachin dravid neville federer*	?

Using the TF-IDF weight formula $w_{t,d} = (1+ \log_{10} tf_{t,d})\log_{10}(N/df_t)$, where N is the no. of documents in the collection, determine whether the Rocchio classification will assign document 6 to sports? Why?

2. We wish to build an SVM classifier that categorizes a point into Class A or B based on 3 data points: (1, 1) {Class A), {2, O) (Class A) and (2, 3) (Class B).
(a) Find the optimal separating hyperplane w.
(b) Find the margin p.
(c) Plot the points on a small graph and depict the separator geometrically. Mark points, lines and distances clearly.

3. Below is a training example where s_T and s_B represents whether the query word exits in title and body respectively. Assign weights to s_T and s_B to calculate the score of document-query pair such that the error is minimized. Error is defined as :

$$\epsilon(g,\Phi_j) = (r(d_j, q_j) - \text{score}(d_j, q_j))^2$$

Consider $r(d_j, q_j)$ as 1 for Relevant and 0 for Nonrelevant.

Example	DocID	Query	s_T	s_B	Judgment
Φ_1	37	linux	1	1	Relevant
Φ_2	37	penguin	0	1	Nonrelevant
Φ_3	238	system	0	1	Relevant
Φ_4	238	penguin	0	0	Nonrelevant
Φ_5	1741	kernel	1	1	Relevant
Φ_6	2094	driver	0	1	Relevant
Φ_7	3194	driver	1	0	Nonrelevant

4. Show that Rocchio classification can assign a label to a document that is different from its training set label.