What is feature selection? Explain methods of feature selection used in text categorization.

→ - Number of different words is large even in relatively small documents.
- In big documents collection can be huge.
- Document representation vectors are sparse.
- Most of the words are irrelevant to categorization task. They can be dropped.
- The preprocessing step that removes the irrelevant words is called feature selection.

* Methods of feature selection.
- Most TC systems at least remove the stop words. - common words that do not contribute to the semantics of documents.
- Many systems perform aggressive filtering, removing 90 to 99 % features.

PAGE NO.:
DATE: / /

- To perform the filtering, a measure of relevance of each feature needs to be defined. The simplest such measure is document frequency $DocFreq(w)$.
More sophisticated measures of feature ~~relvo~~ relevance account the rel" bet" features & categories.

- Information gain.

$$IG(w) = \sum_{c \in C \cup \bar{C}} \sum_{f \in \{w, \bar{w}\}} P(f, c) \cdot \log \frac{P(c|f)}{P(c)}$$

Measures no. of bits of info obtained for prediction of categories.
The probabilities are computed as ratios of frequencies in training data.

- chi - square

$$\chi^2_{max}(f) = \max_{c \in C} \frac{|Tr| \cdot (P(f,c) \cdot P(\bar{f},\bar{c}) - P(f,\bar{c}) \cdot P(\bar{f},c))^2}{P(f) \cdot P(\bar{f}) \cdot P(c) \cdot P(\bar{c})}$$

measures maximal strength of dependence bet" feature & categories.

8. Explain TC using probabilistic classifiers & Bayesian logistic Regression.

→ A] Probabilistic classifiers.

- Probabilistic classifiers view the categorization status value CSV(d,c) as the probability P(c|d).
- The document 'd' belongs to the category c & compute this probability by an application of Bayes theorem:-

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)}$$

- The marginal prop probability P(d) need not ever be computed because it is constant for all categories.
- To calculate P(d|c) assumptions are

- document representation as a feature vector $d = (w_1, w_2, \dots)$
- all coordinates are independent.

Hence

$$P(d|c) = \prod_i P(w_i|c)$$

- The classifier resulting from this assumption are called Naive Bayes classifiers.
They are called "naive" because the assumption is never verified.

- Ex -

test	category
a great game	sports
election is over	not sports
very clean match	sports
a clean but forgotable game	sports
it was a close election	not sports

The probability that the sentence "a great game" is sports =

$$P(\text{sports}|\text{a great game}) = \frac{P(\text{a great game}|\text{sports}) \times P(\text{sports})}{P(\text{a great game})}$$

B] Bayesian logistic regression.

- Bayesian logistic regression is an old statistical approach that is applied to TC problem.
- Quickly gaining popularity owing to its apparently very high performance.
- Assuming categorization is binary, logistic Regression model has form -

$$P(c|d) = \varphi(\beta \cdot d) = \varphi\left(\sum_i \beta_i d_i\right)$$

where,

$c = \pm 1$ is used instead of $\{0, 1\}$.

$d = (d_1, d_2, \ldots)$ = document representation

$\beta = (\beta_1, \beta_2, \ldots)$ = model parameters vector.

φ = logistic link function.

$$\varphi(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

Bayesian approach to logistic regression avoids overfitting.

3.9. Explain TC using Decision Tree classifiers & Decision rule classifiers.

→ A) Decision Tree classifiers.

- Decision Trees can present with a graphical representation of how the classifier reaches its decision.
- A DT classifier is a tree in which the internal nodes are labelled by features, edges leaving a node are labelled by tests on the feature's weight, the leaves are labelled by categories.

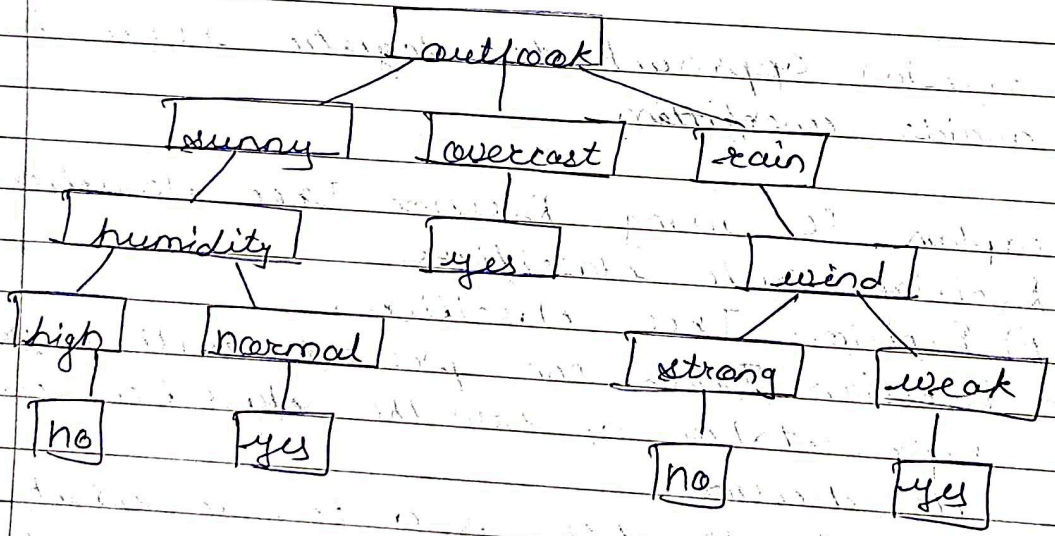
- A DT categorizes a document by starting at the root of tree & moving successively downward via the branches whose conditions are satisfied by document until a leaf node is reached.

- The document is then assigned to the category that labels the leaf node.

- Most of the DT based systems use some form of general procedure for DT induction such as - ID3, C4.5 & CART.

- The choice of features at each step is made by some measures such as info gain or entropy.

- Ex.



B] Decision rule classifiers.

- DR classifiers are like decision trees.
- The rules look very much like DNF rules of CONSTRUE
- Rule learning methods select best rule from set of all possible covering rules.
- DNF rules are often built in a bottom-up fashion.
- Ex -

$d_1 \wedge d_2 \wedge \ldots \wedge d_n \rightarrow c$

where d_i are features of document & c is category.

- Rule learner then applies a series of generalizations for maximizing the compactness of rules.
- Rule learners vary widely in their specific methods depending on heuristics & optimality criteria.
- One of the algorithm is RIPPER.
- Feature of Ripper is its ability to tune the performance by setting the loss ratio parameter.

Q.11. Explain TC using regression method, Rocchio method & neural networks.

→ A] Regression method

- Regression is a technique for approximating a real-valued fun^n using the knowledge of its values on a set of points.
- It can be applied to TC, which is problem of approximating the category assignment function.
- One method is linear least-square fit (LLSF).

Date

- category assignment fun^n is $|c| \times |F|$ matrix describes some linear transformation from the feature space.
- The LLSF model computes the matrix by minimizing the error on training collection according to formula

$$M = \arg\min_M \|MD - O\|_F.$$

where D is $|F| \times |$ training collection $|$ matrix, O is $|c| \times |$ training collection $|$ matrix $\|.\|_F$ is Frobenius norm.

B] Rocchio method.

- Rocchio classifier categorizes a document by computing its distance to the prototypical examples of the categories.
- A prototypical example for the category c is a vector $(w_1, w_2, \ldots)$ in the feature space computed by

$$w_i = \frac{\alpha}{|POS(c)|} \sum_{d \in POS(c)} w_{di} - \frac{\beta}{|NEG(c)|} \sum_{d \in NEG(c)} w_{di}.$$

where POS(c) & NEG(c) sets of all training documents that belong & do not belong to the category c resp.
w_{di} is the weight of i^{th} feature in the document d.

- Usually positive examples are more imp than negative ones. & so $\alpha >> \beta$.
- If $\beta = 0$, then prototypical example for a category is simply centroid of all documents belonging

to the category.
- It is easy to implement & computationally cheap.

C] Neural networks :-
- Neural networks can be built to perform TC.
- Input nodes of network receive the feature values, the output nodes produce the categorization.
- Link weights represent dependence relations.
- For classifying a document, its feature weights are loaded into input nodes
- The neural networks are trained by back propagation.
- If a misclassification occur, the error is propagated back through the network, modifying the link weights in order to minimize the error.
- Simplest kind of neural network is Perceptron.

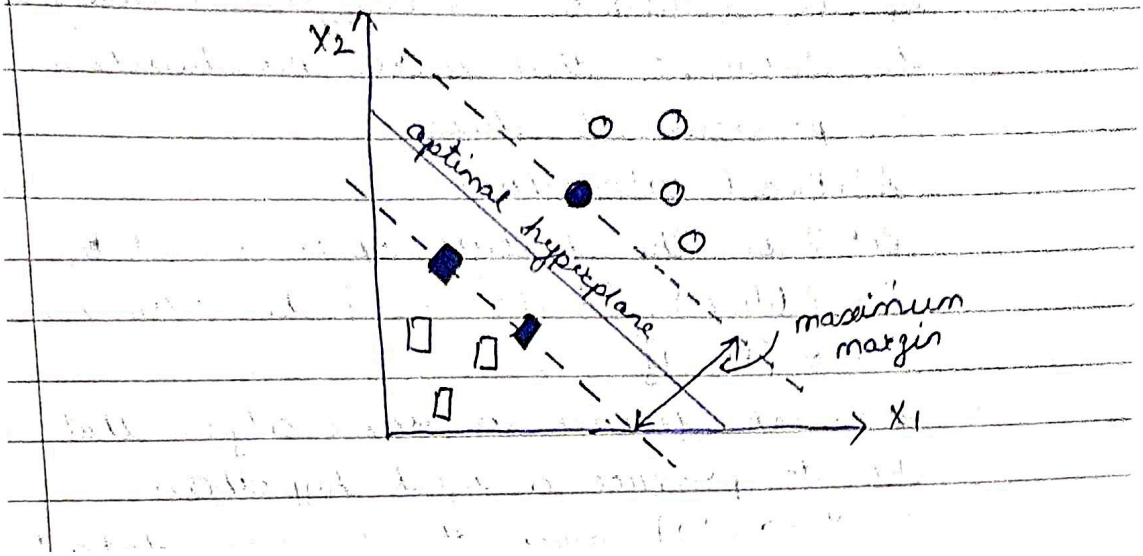
Q.12. Explain TC using example-based classifiers & support vector machines.

→ A] Example-based classifiers :-
- Example based classifiers do not build explicit declarative representations of categories.
- Rely on computing the similarity bet^n the document to be classified & training document.
- These methods have thus been called lazy learners.
- Training for such classifiers consists of simply

storing the representations of the training documents together with their category labels.
- Most prominent example of example-based classifier is KNN (k-nearest neighbour).
- To decide whether a document d belongs to category c, KNN checks whether the k training documents most similar to d belong to c.
- If the answer is positive, for a sufficiently large proportion of them, a positive decision is made, otherwise negative.
- Need to choose value of k.
- Researchers use $k = 20$ or $30 \leq k \leq 45$.
- Increasing the value of k does not significantly degrade the performance.

B] Support Vector Machines.
- SVM is very fast & effective for TC problem.
- A binary SVM classifier can be seen as a hyperplane in the feature space separating the points that represent the positive instances of category from the points that represent the negative instances.
- The classifying hyperplane is chosen during training as the unique hyperplane that separates the known positive instances from the known negative instances with the maximal margin.
- SVM hyperplanes are determined by small subset of training instances.
- SVM algorithm is different from categorisation algorithms.

- SVM classifier has imp advantage in overfitting problem.



Q.20. Explain document clustering algorithms.

→ - A flat clustering produces a single partition of a set of n objects into disjoint groups.
- Hierarchical clustering results in a nested series of partition.
- Hard clustering → every object may belong to exactly one cluster.
- Soft clustering → objects may belong to several clusters.
- Clustering optimization problems are computationally very hard.
- Agglomerative algorithms begin with each object in a separate cluster & successively merge clusters until a stopping criterion is satisfied.
- Divisive algorithms begin with a single cluster containing all objects & perform spitting until a stopping criterion is met.
- Shuffling algorithms iteratively redistribute objects in clusters.
- The most commonly used algorithms are
  - K-means (hard, flat, shuffling)
  - E-M based mixture resolving (soft, flat, probabilistic)
  - HAC (hierarchical, agglomerative).

1] K-Means algorithm

- The K-Means algorithm partitions a collection of vectors $\{x_1, x_2, \ldots, x_n\}$ into set of clusters $\{c_1, c_2, \ldots, c_k\}$

- Initialization - k seeds, either given or selected

randomly:

iteration - The centroids M_i of current clusters are computed:

$$M_i = |C_i|^{-1} \sum_{x \in C_i} x_i$$

stopping condition - The K means algorithm maximizes the clustering quality function Φ.

$$\Phi(C_1, C_2, \ldots, C_n) = \sum_{C_i} \sum_{x \in C_i} sim(x - M_i).$$

2] EM based probabilistic clustering algorithm

- Expectation Maximization (EM) is a general purpose framework for estimating the parameters of distribution in the presence of hidden variables in observable data.

- initialization - The initial parameters of k distributions are selected either randomly or externally.

- iteration -

E-step: Compute the $P(C_i|x)$ for all objects x by using current parameters of the distribution.

M-step: Reestimate the parameters of distribution to maximize the likelihood of objects.

- stopping condition - At convergence when the change in log-likelihood after each iteration becomes small.

3] HAC - as in Q. 22.

22. Explain hierarchical Agglomerative Clustering (HAC) algo for clustering text documents. How dendogram are used in the clustering process?

IMP

→ -

- initialization - Every object is put into a separate cluster.
- iteration - find the pair of most similar clusters & merge them.
- stopping condition - When everything is merged into a single cluster.

- Different versions based how the similarity bet^n clusters is calculated.
  - single-link : max of similarities bet^n pairs of objects.
  - complete-link : min of similarities bet^n pairs of objects.
  - center of gravity : similarity bet^n centroids of clusters
  - avg link : avg similarity bet^n pairs of objects.
  - group avg : avg similarity bet^n all pairs of objects in a merged cluster.

- complexity of HAC is $O(n^2 s)$.

- By definition, the group avg similarity bet clusters C_i & C_j is:

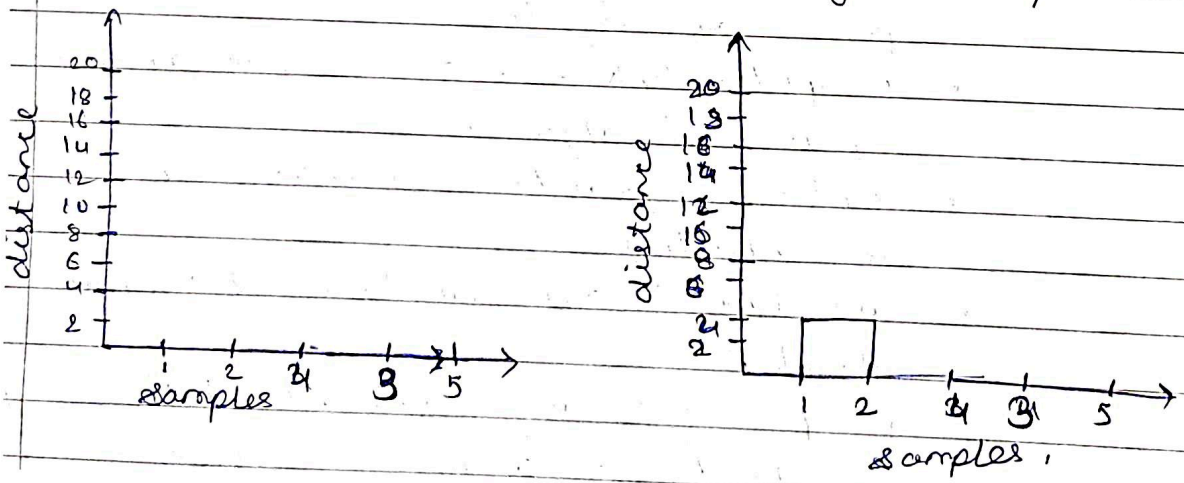
$$Sim(C_i, C_j) = \frac{1}{|C_i \cup C_j|(|C_i \cup C_j| - 1)} \sum_{\substack{x,y \in C_i \cup C_j \\ x \neq y}} Sim(x,y)$$

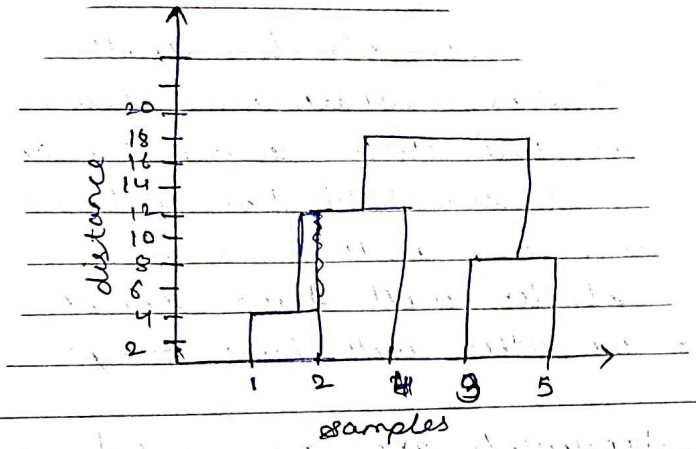
- Assuming that the similarity bet individual vector is the cosine similarity, we have

$$Sim(C_i, C_j) = \frac{(S_i + S_j) \cdot (S_i + S_j) - (|C_i| + |C_j|)}{|C_i \cup C_j|(|C_i \cup C_j| - 1)}$$

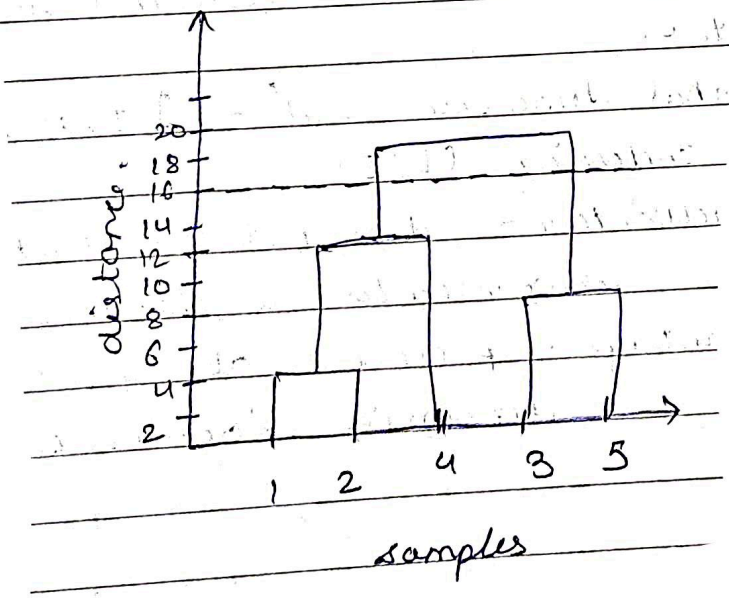
Use of dendrogram -

- To choose the number of clusters in hierarchical clustering, make use of concept called dendrogram.
- A dendrogram is a tree-like diagram that records the sequences of merges or splits.





- We can clearly visualize the steps of hierarchical clustering.
- More the distance of the vertical lines in the dendrogram, more the distance bet^ those clusters.
- Now we can set a threshold distance & draw a horizontal line.
- Generally, set the threshold in such a way that it cuts the tallest vertical line.
- Lets set this threshold as 16 & draw a horizontal line:



- The number of clusters will be the no. of vertical lines which are being intersected by the line drawn using the threshold.
- Since the threshold line intersects 2 vertical lines, we will have 2 clusters.
- One cluster will have sample ~~t1,2~~ (1, 2, 4) & the other will have sample (3, 5).