PAGE NO.:
DATE: / /

DS - 4.

Q.1. What are applications of text categorization?

→ 1] Indexing of texts using controlled vocabulary.
- In Information Retrieval (IR) systems each document is assigned one or more key terms describing its content.
- IR system is able to retrieve the documents according to the user queries.
- The key terms all belong to a finite set called controlled vocabulary.
- The task of assigning keywords from a controlled vocabulary of text documents is called text indexing.
- If keywords are viewed as categories, then text indexing is an instance of general TC problem.

2] Document sorting & text filtering.
- Sorting the given collection of documents into bins.
- Document sorting features - each document belong to exactly one category.
- Ex - e-mail classified into categories such as complaints, spams, deals.
- Text filtering activity can be seen as document sorting with only two bins - the 'relevant' & 'irrelevant' documents.
- Ex - An email client should filter away spam.

3] Hierarchical web page categorization.
- A common use of TC is the automatic classification of web pages under the

hierarchical catalogues.
- Useful for direct browsing & for restricting the query-based search to pages belonging to a particular topic.
- Hierarchical web page categorization can constraint the number of categories to which a document may belong.

3.2. What is the difference bet" single label & multi-label categorization, document-pivoted & category-pivoted categorization, hard & soft categorization?

→ A] single label & multilabel categorization.
- In single-label categorization, each document belongs to exactly one category.
- In multilabel categorization document may belong to any number of categories.
- Single-label categorization is simple generalization of the binary case.
- Multi-label categorization ~~is~~ can be solved by ~~it~~ |c| binary classifiers, one for each group.

B] Document-pivoted & category-pivoted categorization.
- Given a document, the classifier finds all categories to which the document belongs. This is called document-pivoted categorization.
- Alternatively, find all documents that should be filed under a given category, called a category-pivoted categorization.

PAGE NO.:
DATE: / /

c] hard & soft categorization:-
- A fully automated categorization system makes a binary decision on each document category pair. Such a system is said to be doing the hard categorization.
- In semiautomated approach decision to assign a document to a category is made by human & TC system provides a list of categories. Such a system is called soft categorization.

Q.3. Explain techniques for document representation.
→ - The common classifiers & learning algorithms cannot directly process the text documents in their original form.
- Documents are converted into manageable representation.
- Documents are represented by feature vectors.
- A feature is simply an entity without internal structure.
- A document is represented as a vector - a sequence of features & their weights.
- Bag - of words model -
  - Most common model
  - uses all words in document as features
  - Dimension of feature space is equal to the number of different words in all of the documents.
- The ~~F~~ methods of giving weights to features may vary.
- The simplest is the binary in which the

feature weight is zero or one.
- More complex weighing schemes are possible that take into account the frequencies of word in document, in category & in the whole collection.
- TF-IDF scheme-
  - gives the word w in the document d the weight
  - TF-IDF_weight(w,d) = TermFreq(w,d) * log(N / DocFreq(w))

where TermFreq(w,d) is frequency of word in the document,
N is no. of all documents
DocFreq(w) is no. of documents containing word w.

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.

- 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.

Q. 5. How dimensionality can be reduced by using feature extraction?

→ - One way is to create a new, much smaller set of synthetic features from original feature set.
- Rational - owing to polysemy, homonymy & synonymy, the words may not be the optimal features.

- Term clustering addresses the problem of synonymy by grouping together words with a high degree of semantic relatedness.

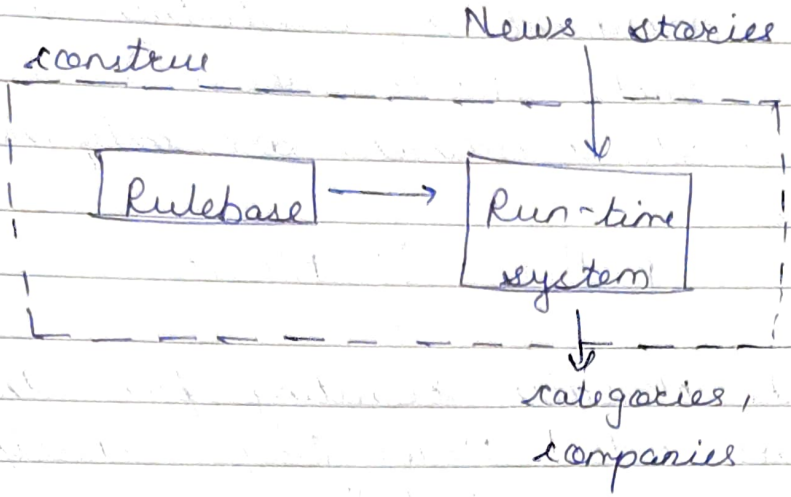
- A more systematic approach is Latent Semantic Indexing (LSI).
  Several LSI representations, one for each category, outperform a single global LSI representation.
- LSI usually performs better than chi-square filtering scheme.

3.6. Explain knowledge engineering approach to TC.

→ - It is focused around manual development of classification rules.
- A domain expert defines a set of sufficient conditions for a document to be labelled with a given category.
- Approach to TC → CONSTRUE system.
- A typical rule in CONSTRUE system is as follows:-
  if DNF (disjunction of conjunctive clauses) formula then category else ¬ category.

- 90% breakeven bet^ precision & recall on a small subset of the Reuters collection.
- Even accuracy is good. ML approach with less accuracy is better.

- CONSTRUE system - news story categorization system.



Specific goals -
- Accept Reuters news stories, including economic, financial & general news.
- Categorize each story into zero, one or several 150 categories.
- Recognize mentions of companies from a database of 10,000 company names.
- Process stories in an avg 5 & seconds.

Q. 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 Baye's 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 -

$$P(c|d) = \varphi(\beta.d) = \varphi(\Sigma_i \beta_i d_i)$$

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.

Q.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.

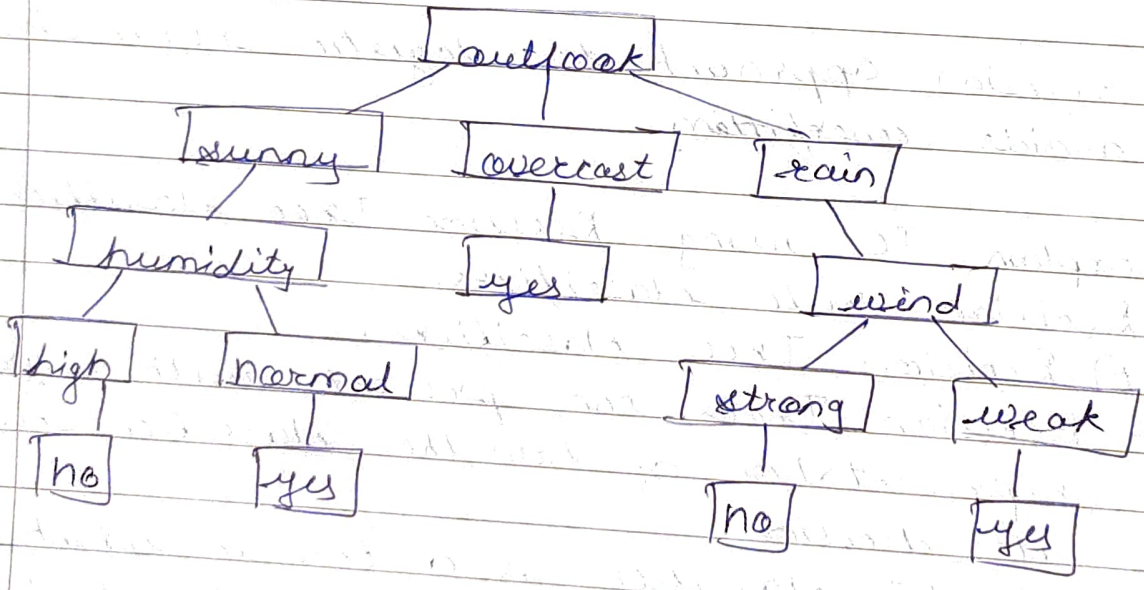
- A DT categorizes a document by starting at the root of tree & moving successively downward via the branches whose conditions are satisfied by document until a leaf node is reached.

- 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 -
  $d_1 \wedge d_2 \wedge \ldots d_n \rightarrow c$
  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).

- 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 ~~receive~~ 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.

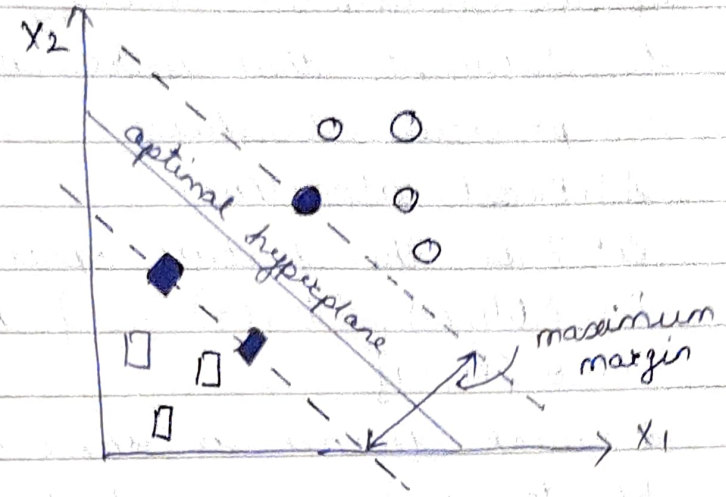
Q.12. Explain TC using example-based classifiers & support vector machines.

→ 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.13. Explain committees of classifiers. What are advantages of using classifier committees.

→ - Committees of classifiers stems from the intuition a team of experts, may produce better results.

A] Bagging method.
- Individual classifiers are trained in parallel on same training data.
- Assume there are k different classifiers.
- Must choose the method of combining their results.
- The simplest method is majority vote. → category is assigned to a document iff at least $(k+1)/2$ classifiers decide this way.
- Another possibility is weighted linear combination where categorization CSV is given by weighted sum of CSV's of k classifiers.

B] Boosting method -

- Classifiers are trained sequentially.
- Before training i^{th} classifier, the training set is reweighted with greater weight given to documents that were misclassified by the previous classifiers.
- AdaBoost algorithm -
  Let X be the feature space & let
  $D = \{(d_1, c_1), (d_2, c_2), \dots\}$ be training data,
  $c_i \in \{+1, -1\}$.
- A weak learner is some algo. that is able to produce a weak hypothesis
  $h: X \to \{\pm 1\}$ given the training data D together with a weight distribution W upon it.
- The goodness of hypothesis is measured by its error

$$\varepsilon(h, w) = \sum_{i : h(d_i) \neq c_i} w(i)$$

Q.14. How unlabeled data can be used to improve classification?

→
- Unlabeled documents usually exist in abundance.
- The common ways of incorporating knowledge from unlabeled documents
  ① Expectation Maximization (EM)
  ② cotraining.

- 1] Expectation Maximization (EM)
  - First the model is trained over the labelled documents.
  - Then the following steps are iterated

until convergence in a local maximum occurs :-

- E - step → the unlabeled documents are classified by the current model.
- M - step → the model is trained over the combined corpus.

B] Cotraining.

- cotraining works with documents, for which two views are available.
- Views providing two different document representations are sufficient for classification.
- In cotraining, unlabeled documents are classified by means of one of the view.
- Then used for training the classifier using the other view & vice versa.

Q.15. Explain ways of incorporating knowledge from unlabeled documents to improve classification.

→ Q.14.

Q.16. How to evaluate performance of text classifiers?

→ - TC experiment requires a document collection labelled with a set of categories.
- Divided into two parts: the training & test document sets.
- The training set is used for training classifier & test set is one on which the performance measures are calculated.
- A commonly used method is the n-fold cross-validation.
- The whole document collection is divided into

n equal parts & then the training-and-testing process is run n times, each time using a different part of collection as the test set.
- Then the result for n folds are averaged.

- The most common performance measures are recall & precision.
- A recall for a category is defined as the percentage of correctly classified documents among all documents belonging to that category.
- Precision is the percentage of correctly classified documents among all documents that were assigned to category by the classifier.
- Another measures the breakeven point, which is the value of recall & precision at the point on the recall-versus-precision curve where they are equal.
- F1 measure, equal to $\frac{2}{(1/\text{recall} + 1/\text{precision})}$, which combines the two measures in ad hoc way.

Q.18. What are applications of text clustering?

→ 1] Improving search recall
- Standard search engines & IR systems return list of documents that match a user query.
- clustering may help improve the recall
- Might significantly degrade precision.

2] Improving search precision.
- More documents means difficult task to browse.
- We must know exact search terms in order to find a document of interest.
- Clustering can group documents into a much smaller number of groups of related documents, ordering them by relevance.
- Experience has shown that the user needs to guide the clustering process so that the clustering will be more relevant to the user's specific interest.

3] Scatter / gather
- Scatter / gather browsing method uses clustering as a basic organizing operation.
- Purpose → To enhance the efficiency of human browsing of document collection when a specific search query cannot be formulated.
- During scatter / gathering browsing session, a document collection is scattered into a set of clusters and the short descriptions of the clusters are presented to user.
- The user selects one or more of the clusters that appear relevant.

4] Query specific clustering.
- The hierarchical clustering is appealing.
- The most related documents will appear in the small tight clusters.
- Recent experiments show better perform over document collections of realistic size.

Q.19. Define general clustering problem. Explain most popular similarity measure ~~metrics~~ metrics used in clustering algorithms.

→ A] General clustering problem -

A clustering task may include the following components :-

- Problem representation, including feature extraction, selection or both
- definition of proximity measure
- actual clustering of objects
- data abstraction
- evaluation.

B] Similarity measure metrics.

- The most popular metric is the usual Euclidean distance

$$D(x_i, x_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

which is particular case with $p = 2$ of Minkowski metric.

$$D_p(x_i, x_j) = \left(\sum_k (x_{ik} - x_{jk})^p \right)^{1/p}$$

- For text document clustering, the cosine similarity measure is the most common:

$$\text{sim}(x_i, x_j) = (x'_i \cdot x'_j) = \sum_k x'_{ik} \cdot x'_{jk}$$

where x' is the normalised vector $x = \frac{x}{|x|}$.

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 splitting 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 esternally.

- 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.

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

IMP

→
- Begins with each object in separate cluster
- Proceeds to repeatedly merge pairs of clusters that are most similar.
- Finishes when everything is merged into a single cluster.
- 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 between clusters C_i & C_j is:

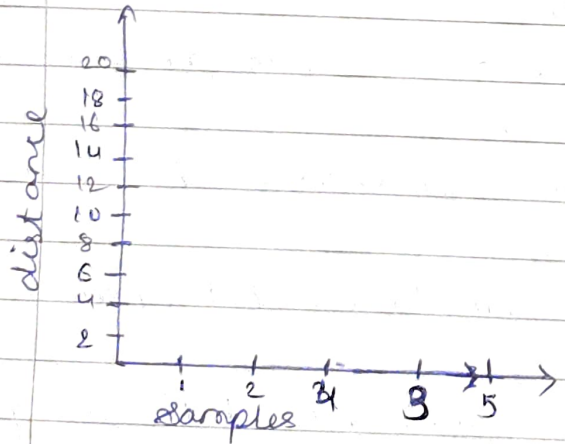
$$Sim(C_i, C_j) = \frac{1}{|C_i \cup C_j|(|C_i \cup C_j| - 1)} \sum_{x,y \in C_i \cup C_j, x \neq y} Sim(x,y)$$

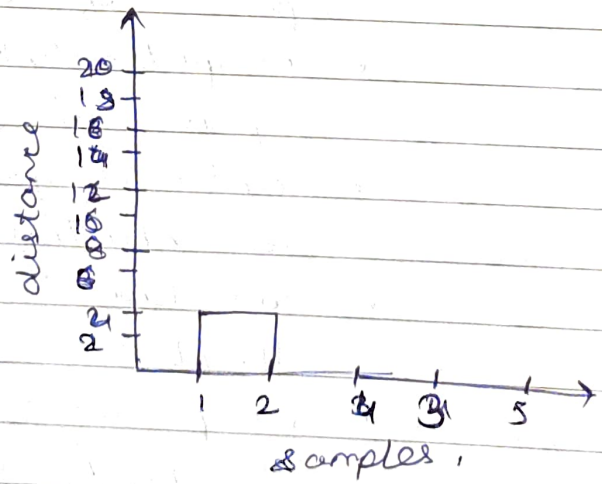
- Assuming that the similarity between individual vector is the cosine similarity, we have

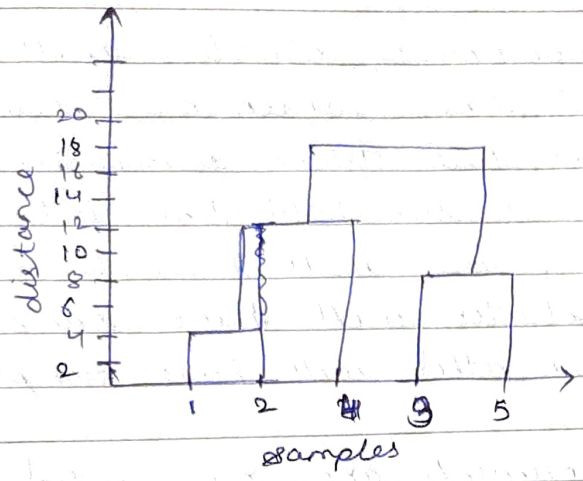
$$Sim(C_i, C_j) = \frac{(S_i + S_j) \cdot (S_i + S_j) - (|C_i| + |C_j|)}{|C_i \cup C_j|(|C_i \cup C_j| - 1)}$$

Use of dendrogram -

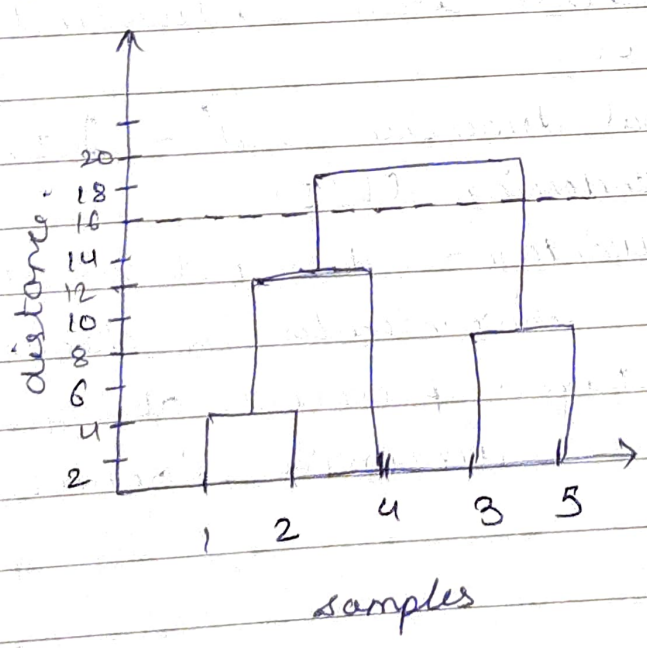
- To choose the number of clusters in hierarchical clustering, make use of concept called dendrogram.
- A dendrogram is a tree-like diagram that records the sequences of merges or splits.







- 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).

Q.23. How documents are represented for text clustering?

→ - To cluster, the documents must be converted into vectors in the feature space.
- common way :- the bag-of-words document representation
- One very imp problem arises for clustering - feature selection.
- Dimension of feature space range into tens & hundreds of thousands.
- Two possible ways of reducing the dimensionality
  - local methods - simply delete unimportant components from individual document vectors.
  - global dimension red^n - Latent Semantic Indexing (LSI)

  disadvantage - does not adapt to unique characteristic of document

  advantage - preserves ability to compare dissimilar documents.

PAGE NO.:
DATE: / /

Q.24. Explain dimension reduction using Latent Semantic Indexing (LSI)?

→ - The Singular Value Decomposition (SVD) of a matrix A is the factorization of A into the product of three matrices $A = UDV^T$.

- Singular value decomposition is a method of decomposing a matrix into three other matrices.

- Application of SVD is dimensionality red^n.

- Data with a large number of features, such as more features than observations may be reduced to a smaller subset of features that are most relevant to prediction problem.

- The result is a matrix with a lower rank that is said to approximate the original matrix.

- Leads to a low-dimensional representation of a high-dimensional matrix.

- An SVD of a real $m \times n$ matrix A is a representation of matrix as a product

$$A = UDV^T$$

where U is a column-orthonormal $m \times r$ matrix,
D is a diagonal $r \times r$ matrix,
V is column-orthonormal $n \times r$ matrix,
r denotes rank of A.

- The term column-orthonormal means that the column vectors are normalized & have a zero

dot product

$\therefore \quad UU^T = VV^T = I.$

- The diagonal elements of D are the singular values of A.

- There are many methods of computing the SVD of matrices.

- First a terms-by-documents rectangular matrix A is formed.
- Its columns are vector representations of documents.
- Thus, the matrix element A_{td} is non-zero when the term t appears in document d.
- Then the SVD of the matrix A is calculated:

$$A = UDV^T$$

- Next the dimension red^n takes place
- We keep k highest values in matrix D & set others to zero resulting in matrix D' }
- It can be shown that matrix $A' = UD'V^T$ is a matrix of rank k that is closest to A.

Q.25. Explain data abstraction in text clustering & evaluation of text clustering algorithms.

→ A] Data abstraction in text clustering -

- Data abstraction in clustering problems entails generating a meaningful & concise description of cluster.
- Useful for automatic processing
- Machine usable abstraction : cluster centroids or probabilistic models of clusters.

PAGE NO.:
DATE: / /

- text clustering: give the user meaningful cluster label
- for scatter/gather browsing: good labelling is required.
- generating cluster labels automatically -
  - title of medoid document
  - several words common to cluster documents
  - a distinctive noun phrase, is probably best label.

B] Evaluation of text clustering.
- Measuring the quality of an algorithm is a common problem.
- Quality of the results needs human judgment.
- Need a measure of how good clustering is for human consumption.
- Given a set of categorized documents
- common measure is purity.
- Assume $\{L_1, L_2, \ldots, L_n\}$ are the manually labeled classes.
- Let $\{C_1, C_2, \ldots, C_m\}$ are the clusters returned by the clustering process.

$$\text{Purity}(C_i) = \frac{\max_j |L_j \cap C_i|}{|C_i|}.$$

- Other measures are entropy, mutual info.

Q.7. Explain any two ML approaches to TC.