*Q1. Explain the role of machine algorithm in Spam filtering* **•E-mail provide perfect way to send millions of advertisement at no cost for the sender, and this unfortunate fact is nowaday extensively exploited by several organizations. • As a result, the e-mailboxes of millions of people get cluttered with all this so-called unsolicited bulk e-mail also known as "spam" or "junk mail". • Machine learning methods of recent are being used to successfully detect and filter spam emails. • Different categories of spam filtering1. Content Based Filtering Technique : Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning approaches, such as Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbore , Neural Networks. 2. Case Base Spam Filtering Method : Case base or sample base filtering is one of the popular spam filtering methods. Firstly, all emails both non-spam & spam emails are extract from each user email using collection model.** *Q2. Explain the role of machine learning algorithms in Natural Language processing* ***=* • The role of machine learning and Al in natural language processing (NLP) and text analytics is to improve, accelerate and automate the underlying text analytics functions and NLP features that turn unstructured text into useable data and insights • Machine learning for NLP and text analytics involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of text. • The techniques can be expressed as a model that is then applied to other text, also known as supervised machine learning. • It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as unsupervised machine learning.• The most popular NLP machine learning algorithms are: 1.Support Vector Machines 2 Bayesian Networks 3 Maximum Entropy 4 Conditional Random Field 5 Neural Networks/Deep learning.** *Q3. What problems are faced by SVM when used with real datasets?* ***1*. Unbalanced data – where the negative instances exceed the positive instances 2. Multilabel classification – SVM is design for binary classification, multilabel is computationally expensive 3. When large real datasets is used with support vector machine, it can extract a very large number of support vectors to increase accuracy and that can slow down the whole process. 4.To allow finding out a trade-off between precision and number of support vectors, Scikitlearn provides implementation called NuSVC, where the parameter nu (bounded between 0 and 1) can be used to control at same time number of support vectors & training errors. •NuSVC is defined as • class sklearn.svm.NuSVC(nu=0.5, kernel='rbf', degree=3, gamma=0.0, coef0=0.0, shrinking=True, probability=False, tol=0.001, cache\_size=200)** *Q6. With reference to Hierarchical Clustering, explain the issue of connectivity constraints****=* • Scikit-learn also allows specifying a connectivity matrix, which can be used as a constraint when finding the clusters to merge. • In this way, clusters which are far from each other (nonadjacent in the connectivity matrix) are skipped. • A very common method for creating such a matrix involves using the k-nearest neighbors graph function, that is based on the number of neighbors a sample :sklearn.datasets.make\_circles(n\_samples = 100, shuffle=True,noise=None, random\_state=None, factor=0.8) • It makes a large circle containing a smaller circle in 2d. A simple toy dataset to visualize clustering and classification algorithms. • Parameters : 1. n\_samples : int, optional (default=100) 2. shuffle : bool, optional (default=True) 3. noise : double or None (default=None) 4. random\_state : int, Random State instance or None (default) 5. factor : 0 < double < 1 (default=.8) Scale factor between inner and outer circle**

*Q5. Explain Evaluation methods for clustering algorithms* ***=* 1] Homogeneity • Homogeneity metric of a cluster labeling given ground truth. A clustering result satisfies homogeneity if all of its cluster contain only data point which are members of a single class. • This metric is independent of absolute value of the label: a permutation of the class or cluster label values won't change score value in any way. •To have homogeneity score, it's necessary to normalize this value considering initial entropy of class set H(C) : h=(1-(H(C|K)/H(C)) 2]Completeness•A complementaryrequirement is that each sample belonging to a class is assigned to the same cluster. • A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster. set H(k) : k=(1-(H(C|K)/H(C)) 3] Adjusted Rand Index • The adjusted rand index measures the similarity between original class partitioning (Y) and the clustering. • If total number of samples in the dataset is n, the rand index is defined as : R= (a+b)/(n/2)** *Q4. Explain role of machine learning the following common un-supervised learning problems****:* i) Object segmentation : Object segmentation is the process of splitting up an object into a collection of smaller fixed-size objects in order to optimize storage and resources usage for large objects. S3 multi-part upload also creates segmented objects, withobject representing each part. ii) Similarity detection : In contrast to symmetry detection, automatic similarity detection is much harder and more time-consuming. The symmetry factored embedding & symmetry factored distance can be used to analyze symmetries in pointssets. A hierarchical approach was used for building graph of all subparts of object.** *Q7. With reference Clustering, explain issue of “Optimization of clusters****”=* • The first method is based on the assumption that an appropriate number of clusters must produce a small inertia.•However, this value reaches its minimum (0.0) when the number of clusters is equal to the number of samples; therefore, we can't look for the minimum, but for a value which is a trade-off between the inertia and the number of clusters.• Given a partition of a proximity matrix of similarities into clusters, the program finds a partition with K classes that maximizes a fit criterion. Different options are available for measuring fit.•The default option (correlation) maximizes the the data matrix X and a structure matrix A in which a(i,j) = 1 if nodes i and j have been placed in the same class and a(i,j) = 0 otherwise.•Thus, a high correlation is obtained when the data values are high within-class and low between-class. This assume similarity data as input•For dissimilarity data,program maximizes negative of the correlation. Another measure of fit is the density function, which is simply the average data value within classes.•There is also a pseudo correlation measure that seeks to measure the difference between average value within classes & the average value between classes.** *Q8. Explain with example the variant of SVM, the Support vector regression***• Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). • The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. • First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. • In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem • But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration** • **However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.**