1) What are the embedding techniques used in text mining?

- Word embedding : This method used to represent word as a numerical vectors. The goal is to capture the meaning of the words in the vector space, so that words that are semantically similar are close together in the vector space. Examples => word2vec and Glove.

- Sentence embedding: This method used to represent a sentence as a numerical vector. The goal is to capture the meaning of the sentence in the vector space, so that sentences that are semantically similar are close together in the vector space. There are several methods for generating sentence embeddings, including averaging the word vectors of the words in the sentence and using neural network-based approaches.

- Document-embedding: This method used to represent a document as a numerical vector. The goal is to capture the meaning of the document in the vector space, so that documents that are semantically similar are close together in the vector space. There are several methods for generating document embeddings, including averaging the word vectors of the words in the document and using neural-based approach.

- Paragraph embedding: This method used to represent a paragraph as a numerical vecto. The goal is to capture the meaning of the paragraph in the vector space, so that paragraphs that are semantically similar are close together in the vecto space. There are several methods for generating paragraph embeddings, including averaging the word vectors of the words in the paragraph and using neural network-based approach.

2) What kind of techniques can be used when there are more than one types of outliers? Describe one of them.

- Ensemble outlier detection
- Meta learning algorithms
- Clustering
- Anamoly detection
- Outlier ensembles

* Outlier ensembles

They involve training a variety of different outlier detection models and then combining their predictions using a voting or averaging scheme. The goal of outlier ensembles is to improve the robustness and reliability of outlier detection by combining the strengths of multiple models.

To implement an outlier ensemble, first step is to select a diverse set of outlier detection models to include in the ensemble. This can include models based on different algorithms, as well as models that have been trained on different subsets of the data.

Next, the models are trained on the data and their predictions are combined using a voting or averaging scheme. In a voting scheme, each model votes on whether an observation is an outlier, and the final decision is based on majority vote. In an averaging scheme, the outlier scores from each model are combined and the observations with the highest average scores are identified as outliers.

Outlier ensembles can be effective at identifying multiple types of outliers, as they are able to capture the strengths of multiple models and are less sensitive to the presence of individual outlying observations. However, it is important to carefully select the models to include in the ensemble, as the performance of the outlier ensemble will depend on the quality of the individual models.

3) What is the graph mining?

Graph mining is a type of data mining that focuses on the analysis of graphs or networks. A graph is a collection of nodes and edges that connect the nodes. Each node represents a unique entity, and the edges represent relations between the nodes. Graphs can be used to represent a wide variety of data, including social networks, communication networks...

Graph mining algorithms are used to extract meaningful patterns and insights from the data represented in a graph. This can involve tasks such as identifying the most central nodes in the graph, discovering comminities or clusters of nodes that are densely connected, and predicting missing or future edges. Graph mining techniques can be applied to both directed and undirected graphs, and can be used in a supervised or unsupervised setting.

- Community detection which involves identifying groups of nodes in a graph that are densely connected and have fewer connections to the rest of the graph. Community detection algorithms can be used to discover hidden structure in a graph, and can provide insights into the relationships and patterns within the data

One populer algorithm for community detection is the Louvain method, which is a fast and scalable method for finding communities in large graphs. It works by iteratively optimizing a modularity measure, which is a measure of the strength of the divisions within a graph. At each iteration, the algorithm reassigns nodes to different communities in order to maximize the modularity. The process is repeated until convergence, at which point the final community assignments are returned.

4)

* **Correlation analysis** A method that is used to measure the strength and direction of the linear relationship between two variables. Correlation coefficients, such as Pearson's correlation coefficient, can be used to quantify the strength of the relationship with values ranging from -1 to +1.

formula => $r = \Sigma (x-\bar{x})(y-\bar{y}) / \sqrt{(\Sigma (x-\bar{x})^2 \Sigma (y-\bar{y})^2)}$

* **Linear regression** A method that is used to model the linear relationship between a predictor variable and target variable. Estimates the parameters of the linear relationship using the least squares method, and can be used to make predictions about the target variable based on the predictor variable

formula => $y = b0 + b1x$

* **Logistic regression** A method that is used to model the relationship between a predictor variable and a binary target variable. Estimates the probability of the target variable taking on a certain value based on the value of the predictor variable, and can be used to make predictions about the target variable.

formula => $p = e^{(b0 + b1x)} / (1 + e^{(b0 + b1x)})$

* **ANOVA (Analysis of variance)** A method that is used to test whether there is a significant difference between the means of two or more groups. ANOVA can be used to evaluate the relationship between a categorical predictor variable and a continuous target variable.

formula => F= (between-group variance) / (within-group variance)

* **Chi-squared test**
A method that is used to determine whether there is a significant association between two categorical variables. The chi-squared test can be used to evaluate the relationship between two categorical variables, or between a categorical predictor variable and a categorical target variable.

formula => $x^2 = \Sigma$ (observed - expected)^2 / expected

5) explain one feature selection and feature extraction technique.

- feature selection technique

* Recursive Feouture Elimination (RFE):
This is a feature selection technique that involves recursively removing the least important features from the model until the desired number of features is reached. The importance of each feature is determined using a feature importance measure, such as coefficients in a linear model or feature importance scores in a decision tree model. RFE can be used with any model that has a feature importance measure, and is a useful technique for reducing the complexity of the model and improving the interpretability of the results.

- feature extraction technique

* Independant Component Analysis (ICA):
This is a feature extraction technique that involves decomposing the original features into a set of independent components that are maximally statistically independant from one another. ICA is often used to seperate mixed signals that have been combined linearly, and is particularly useful for extracting features from non-Gaussian data. One way to implement ICA is to use an iterative optimization algorithm that maximizes then non-Gaussianity of the independent components.