# 1.Corpus Preparation

## 1.1. Justification

The scenario of classifying emails as spam or non-spam based on their content in terms of words is a common problem in the field of natural language processing and email filtering. With the increasing volume of emails that people receive on a daily basis, it is becoming increasingly important to be able to automatically filter out unwanted or unsolicited emails, such as spam.

The Spambase dataset from UCI Machine Learning repository is a widely used dataset for this problem, containing a collection of emails with their corresponding labels as spam or non-spam. This dataset can be used to train machine learning models that can automatically classify emails as spam or non-spam based on their content, which can save a lot of time and effort for email users.

Two commonly used machine learning models for classification problems are K Nearest Neighbors (KNN) and Decision Trees. KNN is a non-parametric and lazy learning algorithm that makes predictions based on the similarities between new instances and the instances in the training dataset. Decision Trees, on the other hand, are a type of supervised learning algorithm that make predictions by recursively splitting the input space into smaller regions based on a set of rules learned from the training data.

Both KNN and Decision Trees have been shown to be effective in solving the problem of email classification, and hence are suitable for the given scenario. By training and evaluating these models on the Spambase dataset, we can determine which one performs better for this particular problem and use it for filtering emails in real-world applications.

## 1.2. Relevance

The relevance of the scenario lies in its practical application. As email has become an integral part of our daily communication, it has also become a target for spammers who send unsolicited and unwanted emails. Filtering these emails can be a tedious and time-consuming task, which is why the development of automated spam filters is crucial.

By using machine learning models such as KNN and Decision Trees to classify emails as spam or non-spam based on their content, we can save valuable time and effort for email users while also improving the accuracy of spam filtering. The Spambase dataset from UCI Machine Learning repository provides a suitable dataset for training and testing these models on a real-world problem.

Moreover, this scenario is relevant to the field of natural language processing, which has become increasingly important in recent years due to the growth of big data and the need to extract meaningful insights from unstructured data such as text. By using machine learning models to classify emails based on their content, we can improve our understanding of how language is used in different contexts, which can have applications in areas such as sentiment analysis, language translation, and information retrieval.

Overall, the relevance of the scenario lies in its practical application to a real-world problem, as well as its contribution to the development of machine learning algorithms for natural language processing tasks.

## 1.3. Compatibility

The compatibility of the above scenario lies in the fact that it is well-suited for solving using machine learning algorithms. The problem of classifying emails as spam or non-spam based on their content is a classic classification problem, and machine learning models such as KNN and Decision Trees are particularly well-suited for solving such problems.

The Spambase dataset, which contains a collection of emails with their corresponding labels as spam or non-spam, is a widely used dataset for training and testing machine learning models on this problem. Moreover, the dataset is preprocessed and ready to use, which makes it convenient for researchers and practitioners who want to experiment with different machine learning algorithms.

KNN and Decision Trees are two of the most commonly used machine learning algorithms for classification problems, and they are well-suited for this particular problem. KNN is a non-parametric algorithm that makes predictions based on the similarity between the new instance and the instances in the training dataset. Decision Trees, on the other hand, make predictions by recursively splitting the input space into smaller regions based on a set of rules learned from the training data.

Overall, the compatibility of the above scenario lies in the fact that it is a classic classification problem that can be solved using well-established machine learning algorithms such as KNN and Decision Trees, and there exists a suitable dataset, Spambase, which is ready to use for training and testing these models.

# 2. KNN