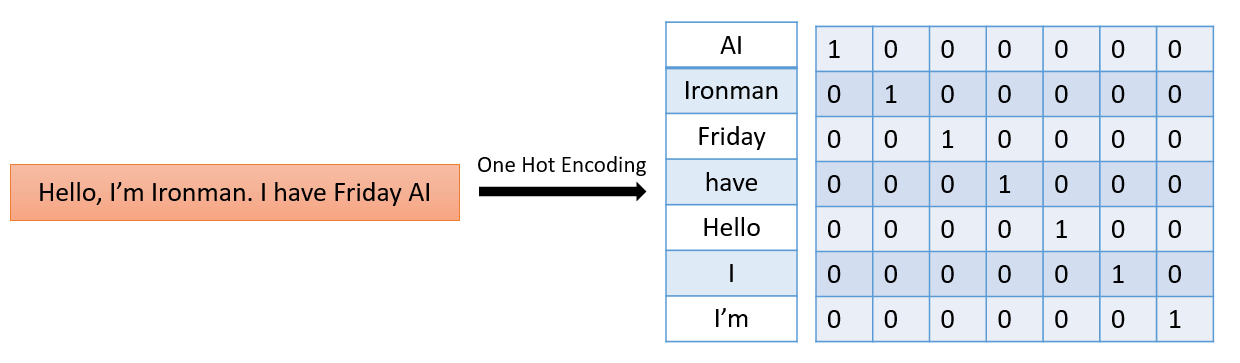
**Feature Engineering in Natural Language Processing and its Types:**

Machine learning algorithms and deep learning algorithms, being mathematical models at their very core, can only work with and train on numbers. In the case of typical classification and regression models, the dataset primarily consists of numbers or can be converted into numbers using techniques such as label encoding. However, in the case of natural language processing, more advanced techniques have to be employed to convert the text data (string datatype) into numerical features or vectors that predictive algorithms can train on. This process of quantifying string data so that it can be used in ML for applications such as sentiment analysis, text summarization, translation, etc. is known as feature extraction.

Feature engineering, more formally, can be defined as the process of extracting and transforming variables from raw data. It plays an important role in the NLP pipeline and is usually done after text pre-processing, which involves tokenizing, lowercasing, cleaning, and lemmatizing the text data, and before the creation of a predictive model. The idea of feature engineering in NLP was introduced in the 1990s, when a more statistical method began to be applied to NLP, which required converting the text into a number. The early feature extractors were simple in their design, and word vectors were created on the basis of a word’s position in a sentence or by the frequency of the word in a document. Later feature extractors captured the semantic relationship between words to create word vectors, while the most advanced systems use deep learning.

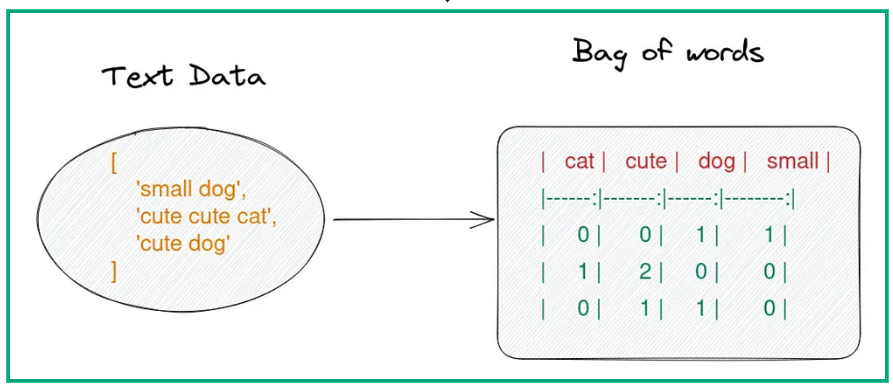
This document is going to primarily deal with two simple models of feature extraction: one hot encoder and a bag of words. However, before proceeding, it is important to define certain terms that will be used in this document. A corpus is the entire text document that has been organised into a tabular form to create a dataset. A corpus consists of many documents, which are unique texts in the form of reviews, comments, etc. The simplest unit in NLP is a word or token, which the feature extractor encodes in the form of numbers. A corpora is a collection of corpuses that are used for training large language models, such as GPT Transformers.

One Hot Encoding is a method used to convert text data into binary vectors on the basis of a word’s position in the sentence. The two binary operators are 1 and 0, and the process of feature extraction using this method involves creating a matrix with an equal number of rows and columns, with both the rows and columns representing the word of the sentence. In all places where the column name becomes equal to the row name, a value of 2 is assigned to that position in the matrix, while other spaces are filled with zeros. An example of one hot encoding of a sentence can be clearly seen in the image given below.



While One Hot Encoding does the job of creating a word vector, it does have some disadvantages. The main disadvantage of this system is that it creates a sparse matrix, and if the text input is large, the matrix created will be very large and composed of  data points, where  represents the number of words in the text. This sparsity often increases the time for computation. Also, One Hot Encoding does not capture the relationship between the words, leading to a loss of information.

The second feature extractor is the Bag of Words, or BoW, method. This method creates a word vector on the basis of the frequency distribution of words in the document. The first step in this method is creating a list of unique words, or a vocabulary of words. Initially, a value of zero is assigned to all the unique words. Following this, the frequency of words is analysed, and a vector is created based on the number of times a word or token appears in a sentence. After this, the vector is used to train the model. An example of this method is given below:



The Bag of Words Method is widely used in natural language processing, but it does have some disadvantages, such as sparsity, loss of information, and out-of-vocabulary errors. If the vocabulary of words is large, the matrix created for the text will be sparse, which will increase computational runtime. In the Bag of Words method, the position of the word in the sentence is lost, and the semantic relationship between the words is also gone. The out-of-vocabulary error arises in a predictive model when the text input from the user consists of vocabulary different from the vocabulary of the trained text.

The Bag of Words and One Hot Encoding techniques for feature extraction use simple methods for converting strings into numerical data. The main difference between the two methods is that One Hot Encoding captures the position of the token in the sentence, whereas Bag of Words simply records the frequency like a histogram. The metrics produced by both methods are sparse, but the matrix produced by One Hot Encoding is sparser. Apart from these differences, both of these methods do not represent the semantic relations between the words, so they are not preferred.

In conclusion, feature extraction is a crucial step in natural language processing and has been instrumental in the development of applications that can analyse and communicate with humans in natural language. One Hot Encoding and the Bag of Words method are techniques that were popular in the past, but they are quickly being replaced by more advanced vectorization tools such as TF-IDF, Word2Vec, and n-gramme. As NLP grows in popularity, the number of feature extractors that use deep learning and transformers is bound to grow.