**DEVI AHILYA VISHWAVIDHYALAYA**

**INDORE (M.P)**

# **SCHOOL OF DATA SCIENCE AND FORECASTING**

**M.SC DATA SCIENCE AND ANALYTICS**

**BATCH – (2021-23)**

**SUBJECT: Nature Language Processing**

(SUBJECT CODE: DS5B-607)

****

**First Assignment**

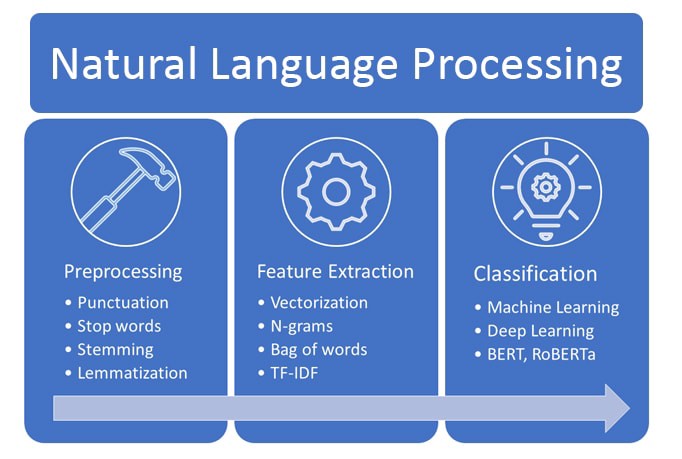
**(THIRD SEMESTER)**

**SUBMITTED BY: SUMITTED TO:**

**Himanshu Jain Mr. Avinash Navlani**

**Feature Engineering with NLP**

[**Natural Language Processing**](https://en.wikipedia.org/wiki/Natural_language_processing) is a field of artificial intelligence that studies the interactions between computers and human languages, in particular how to program computers to process and analyse large amounts of natural language data. NLP is often applied for classifying text data. **Text classification** is the problem of assigning categories to text data according to its content.



**Feature engineering** is one of the most important steps in machine learning. It is the process of using domain knowledge of the data to create features that make machine learning algorithms work. Think machine learning algorithm as a learning child the more accurate information you provide the more they will be able to interpret the information well. If we can use these contexts as features and feed them to our model then the model will be able to understand the sentence better. Some of the common features that we can extract from a sentence are the number of words, number of capital words, number of punctuation, number of unique words, number of stopwords, average sentence length, etc. Focusing first on our data will give us better results than focusing only on models. Feature engineering helps us to create better data which helps the model understand it well and provide reasonable results.

Feature Engineering for NLP Models

Bag of Words



Glove

Fast Text

Doc2Vec

TF-IDF

1) Bag of Words:

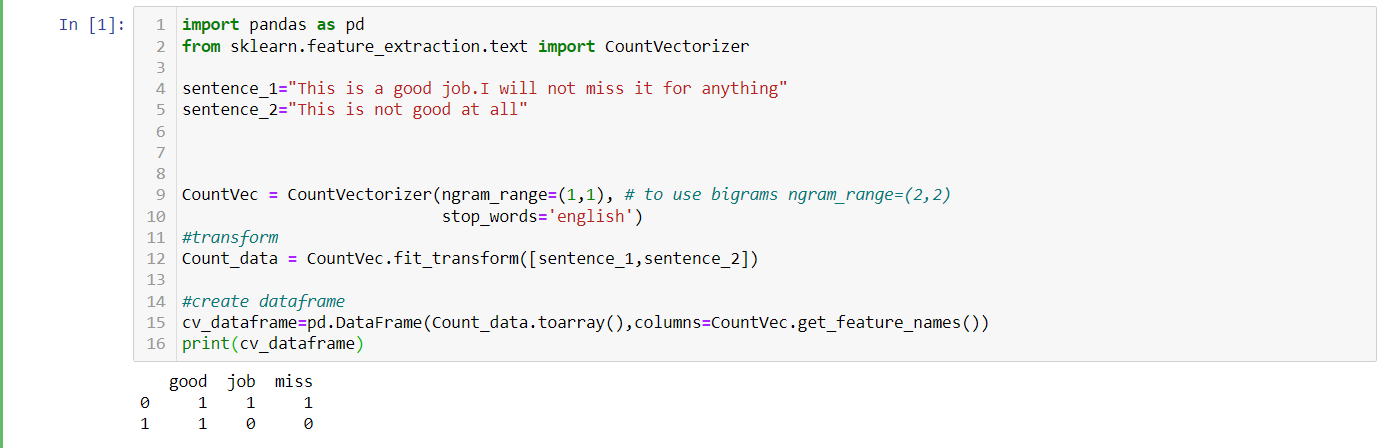
# **What is a Bag of Words in NLP?**

A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a **“bag”** of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

# **Why is the Bag-of-Words algorithm used?**

One of the biggest problems with text is that it is messy and unstructured, and [machine learning](https://www.mygreatlearning.com/blog/what-is-machine-learning/?highlight=what%20is%20machine%20learning) algorithms prefer structured, well defined fixed-length inputs and by using the Bag-of-Words technique we can convert variable-length texts into a fixed-length **vector**. Also, at a much granular level, the machine learning models work with numerical data rather than textual data. So to be more specific, by using the bag-of-words (B o W) technique, we convert a text into its equivalent vector of numbers.

## **Create a Bag of Words Model with Sklearn :**



2) TF-IDF:

The scoring method being used above takes the count of each word and represents the word in the vector by the number of counts of that particular word. What does a word having high word count signify?

Does this mean that the word is important in retrieving information about documents? The answer is NO. Let me explain, if a word occurs many times in a document but also along with many other documents in our dataset, maybe it is because this word is just a frequent word; not because it is relevant or meaningful.

TF-IDF for a word in a document is calculated by multiplying two different metrics:

The**term frequency (TF)** of a word in a document. There are several ways of calculating this frequency, with the simplest being a raw count of instances a word appears in a document. Then, there are other ways to adjust the frequency. For example, by dividing the raw count of instances of a word by either length of the document, or by the raw frequency of the most frequent word in the document. The formula to calculate Term-Frequency is