**Problem Statement: Determination of Short Tail Keywords for Marketing:**

The marketing manager of an AI conference requires recommendations for short-tail keywords for marketing purposes in order to attract general public to enhance public understanding of the domain as well as attract potential investors.

You are given a dataset and its feature description are given below:

1. Title -  Title of the paper
2. Authors - Author(s) of the paper
3. Groups - Author-selected, high-level keyword(s)
4. Keywords - Author-generated keywords
5. Topics - Author-selected, low-level keywords
6. Abstracts - Paper abstracts

Link to the dataset- <https://archive.ics.uci.edu/ml/datasets/AAAI+2014+Accepted+Papers>

Can you design a pipeline of how to achieve the task? As evident it is a clustering problem, can you come up with the optimum number of clusters (short tail keywords) while solving this problem? If yes, how?

Solution:

Summary:

The aim of this problem statement is to find different groups of short tail keywords using clustering. The solution can be further broken into two parts-

1. To find optimal number of clusters
2. To find the best suitable Machine learning algorithm for clustering

Machine learning Pipeline: Steps to solve above mentioned two problems are as given below –

1. Understanding the dataset

The first and foremost step to solve any machine learning problem should be to thoroughly understand the data. Looking into the dataset given in the link, its feature description are given below:

1. Title -  Title of the paper
2. Authors - Author(s) of the paper
3. Groups - Author-selected, high-level keyword(s)
4. Keywords - Author-generated keywords
5. Topics - Author-selected, low-level keywords
6. Abstracts - Paper abstracts

There are total 396 records in the dataset with all the features of Object data type i.e. Strings

1. Feature Exploration

On exploring the dataset available, it can be figured out

1. ’Authors’ feature which contains authors of the research paper is irrelevant to find keywords.
2. ‘Abstracts’ feature which contains paper abstracts is too detailed in nature.
3. ‘Keywords’ feature contains keywords from the ‘Title’ of the research paper and hence ‘Title’ feature can also be dropped.
4. Since we are looking for short tailed keywords(Less specific keywords with one or two words) we don’t need title specific Keywords and thus ’Keywords’ feature is of no use.
5. ‘Topics’ feature contains author selected low level keywords which represent the specific topics of the ‘Group’ feature, same can be dropped as we only require short tailed keywords and no specific keywords.

Hence we can consider ‘Groups’ feature for clustering

1. Feature Engineering

Feature Engineering is done on the ‘Groups’ feature and it is converted into numeric form. Relevant steps to convert ‘Groups’ feature into numeric form are

1. Splitting the string into words using space as the delimiter.
2. Removal of special characters like-., /()
3. Removal of Stop words like and, is, are etc.
4. Finding root words by Applying Stemming and Lemmatization.
5. Applying Bag of words algorithm to the words which gives the frequency of each word present in the text.
6. Exploratory Data Analysis

Frequencies of words can be plotted to understand the distribution of words

1. Modelling

Once the words are converted into numbers using the above steps. Clustering can be applied to find similar group of words.

Elbow method can be used to find optimal number of clusters that can be formed using the data. Once number of clusters to be made are known, it is then provided to the clustering algorithm as number of clusters an algorithm should create is a hyper parameter.

Visualizing the clusters can help in determining the Short tail keywords.

1. Evaluation of the model

There is no one way of evaluating the output of clustering model. A clustering model is good enough if maximum variance is explained within each of the clusters, objects within the cluster share similar properties and clusters are further away from each other.