**PROJECT TITLE:**

**MARGINAL WORKERS IN TAMILNADU-A SOCIO ECONOMIC ANALYSIS(DAC)**

**PROBLEM DEFINITION:**

In this Indian scenario, voluntary migration (where the migrants move of their own choice) often takes place with the view to secure a livelihood. On the other hand, involuntary migration might take place due to natural disasters or war and the consequent persecution or in some of the cases because of environmental degradation. In between these two diverse classifications, there exist a great mass of people, whose migration is a direct result of their poverty, and whose migration is often, as much voluntary as involuntary.

**PRE-PROCESSING:**

**STEPS:**

Performing the Data processing transformation on the text data.

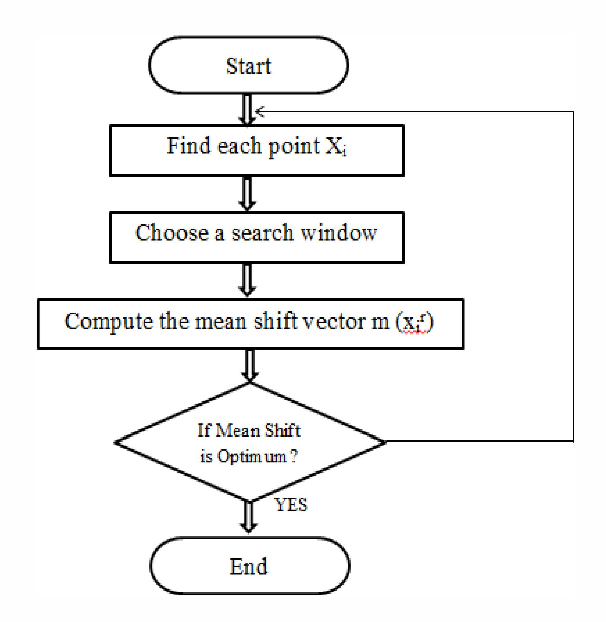
1. Transform characters to lower case.

2. Converting to Plain Text Document

3. Remove punctuation marks.

4. Remove digits from the documents.

5. Remove from the documents words which we find redundant for text mining (e.g. pronouns, conjunctions). We set this words as stopwords(“english”) which is a built in list for English language.

6. Remove extra whitespaces from the documents.

**CATEGORICAL TO NUMERICAL REPRESENTATIONS:**

1.cat.codes Attribute

2.replace

3.Label Encoder

**ALGORITHM:**

The objective of k-means is to minimize the total sum of the squared distance of every point to its corresponding cluster centroid. Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (? n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares where µi is the mean of points in Si.

**PROJECT WORKFLOW:**

