**Clustering**

**Overview:**

* Spot applications of clustering
* Recognise the difference between Classification and Clustering
* Understanding how the K-mean Algorithm works.

Clustering is a way to group items together based on some measure of similarity.

Users in a group must be Maximize intra cluster similarity

“similar” to each other

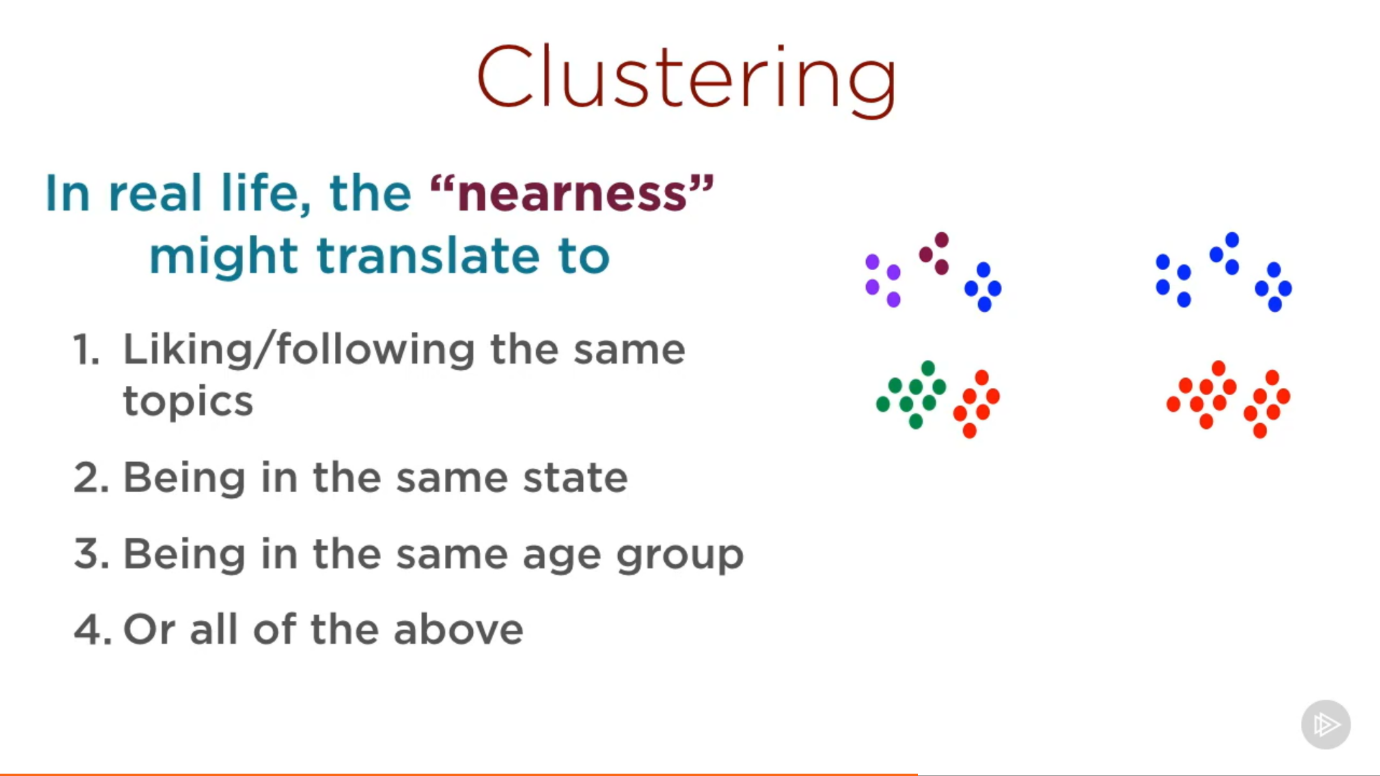
Users in different groups must be dissimilar Minimize inter cluster similarity

to one another

All users can be represented Age

using some features Location

Frequency of usage for each topic



**Clustering Large Data Sets into Meaningful Groups: Contrasting Clustering and Classification:**

**What’s Difference**

**Classification:** Classifying data into pre-defined categories

**Clustering:** Grouping data into set of categories.(Categories are not defined beforehand)

**Classification**

**Steps:**

* Take one instance
* Classify it into a pre-defined category (labels)
* Do this based on training data which has already been classified

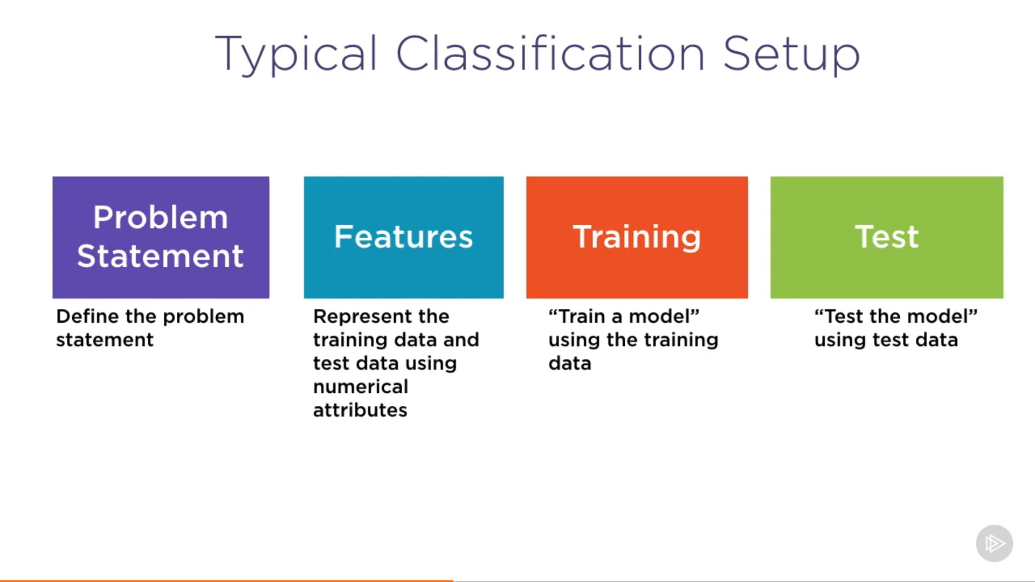
**Clustering**

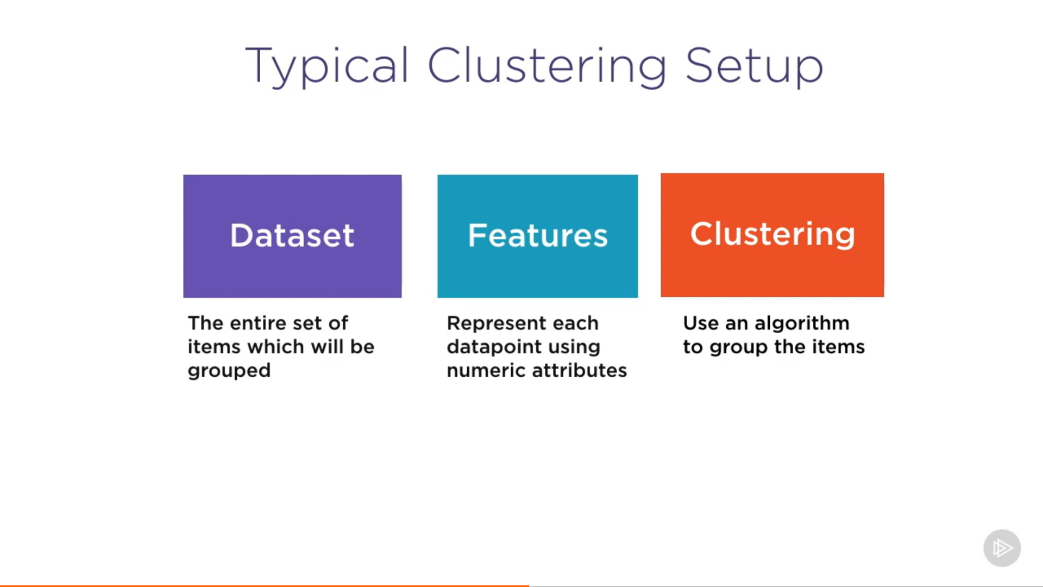
**Steps:**

* Take a large number of instances
* Divide them into groups
* The groups are unknown beforehand

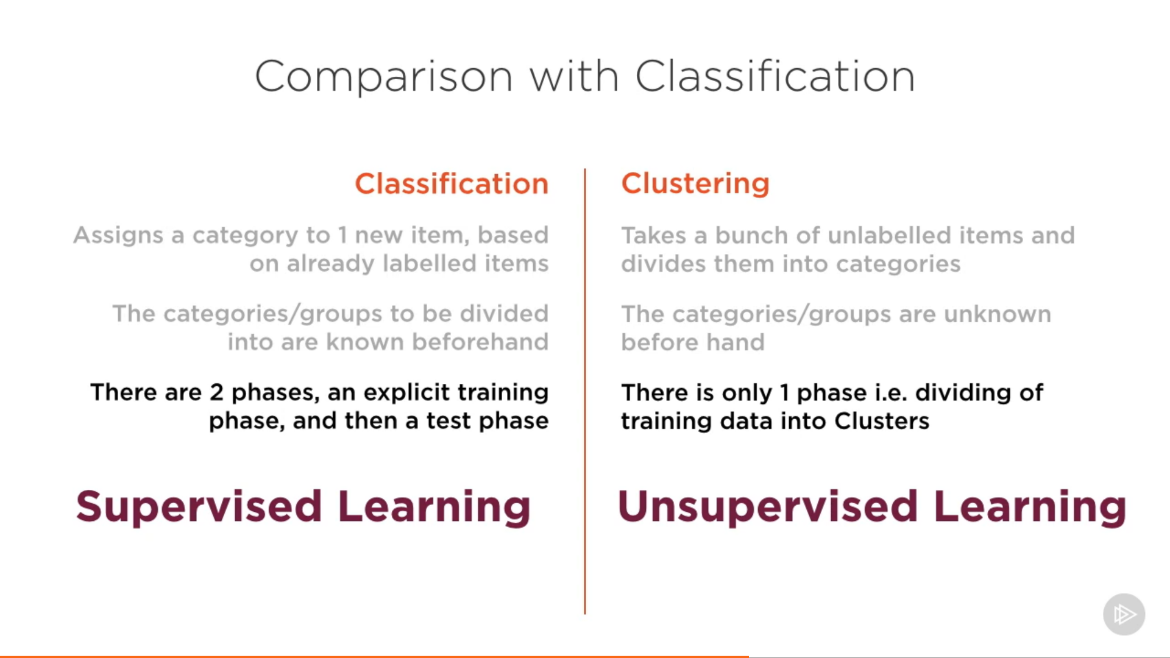
What kind of groups can these users be divided into?

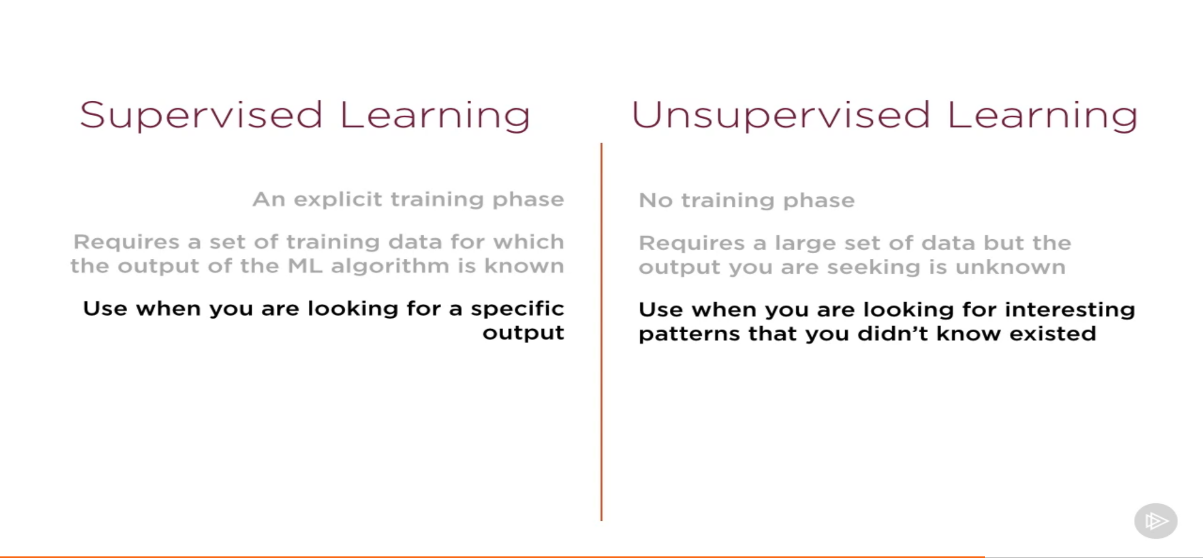
What kind of themes are present in this set of articles?



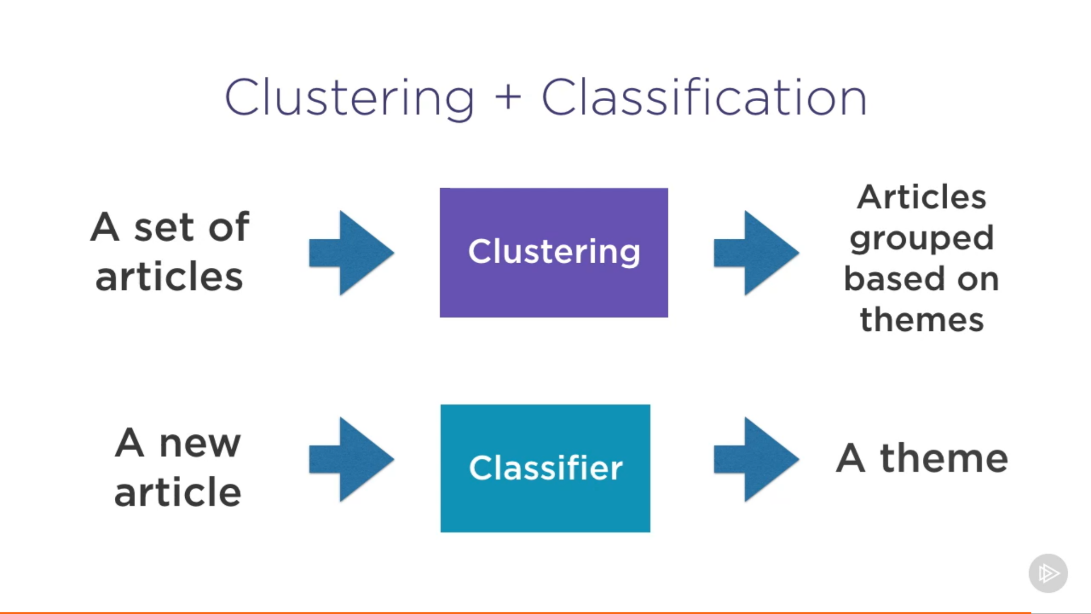


Comparison with Classification





Clustering and Classification go hand in hand



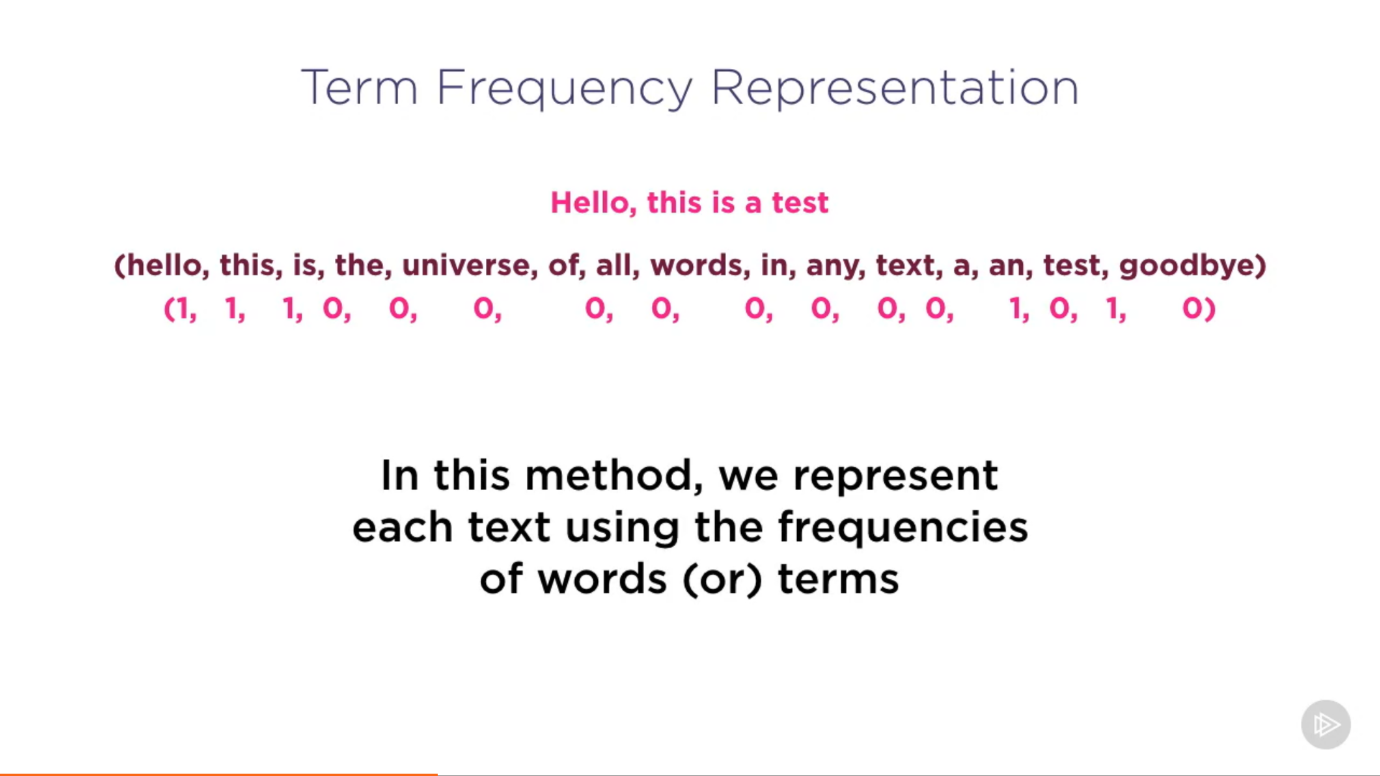
**Document Clustering with K mean:**

Given any set of Documents, Group them based on similarity of content

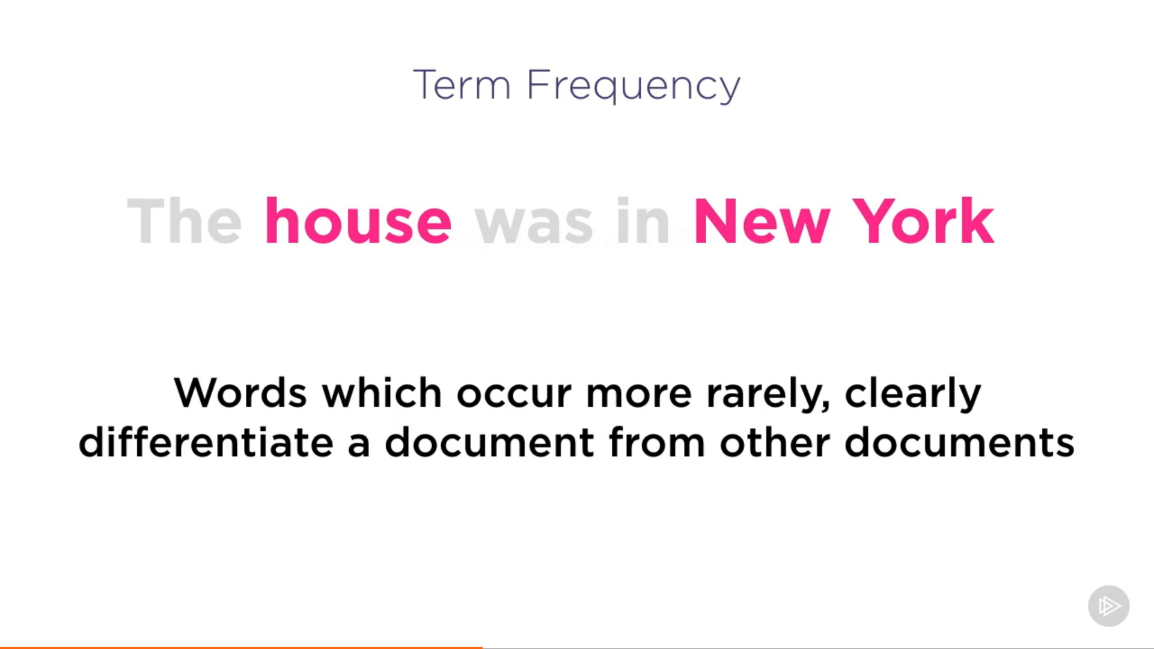
Study the identified clusters to find interesting themes.

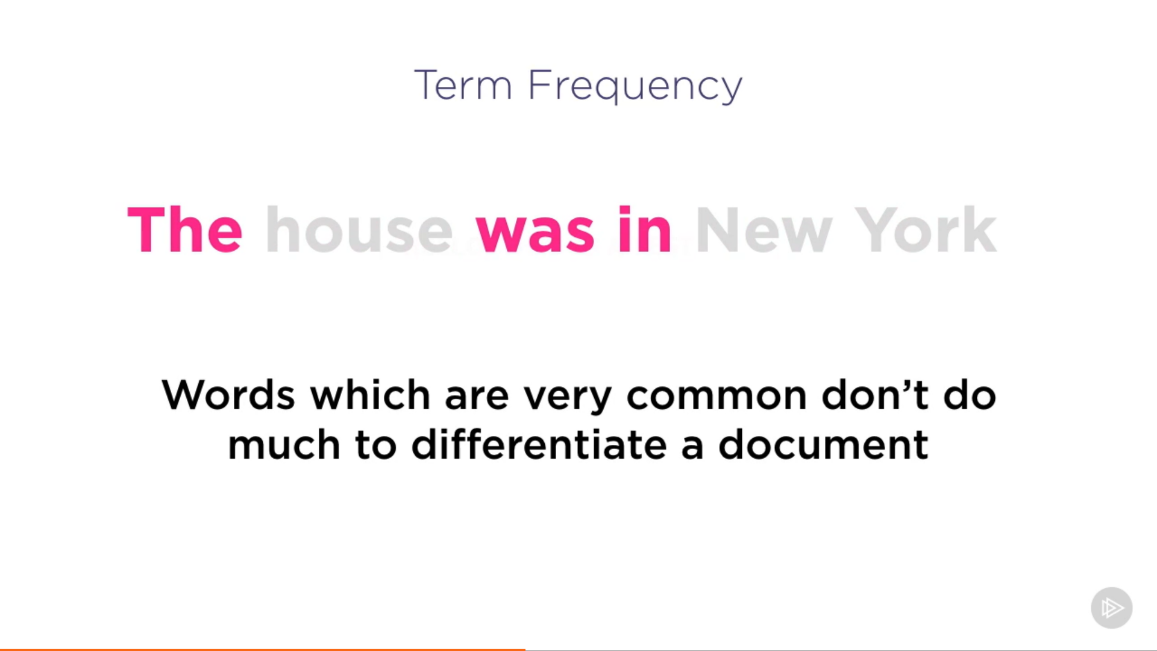
**Representing Text using features:**

Term Frequency representation



Some words characterize a document more than others



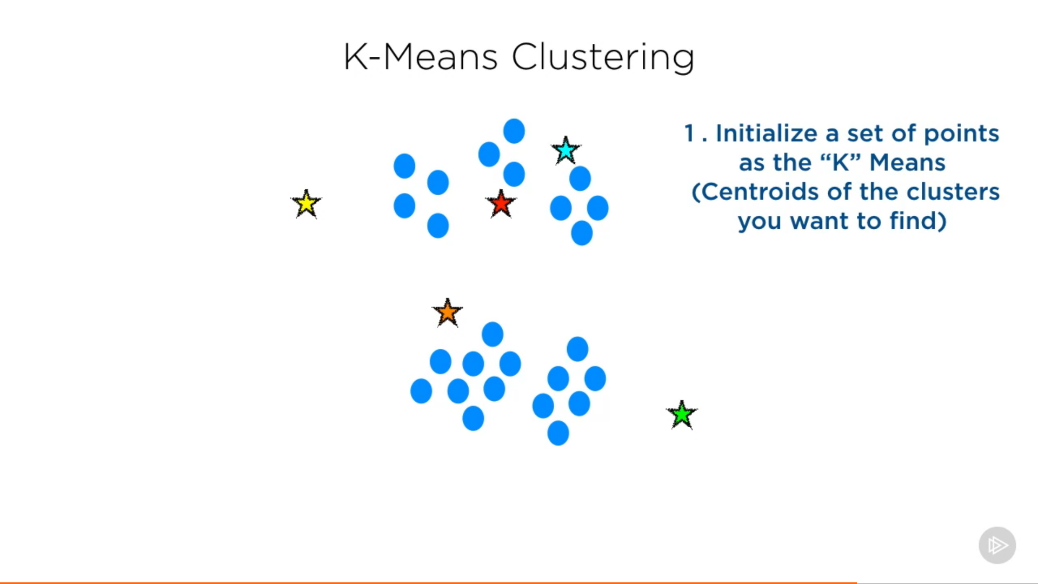


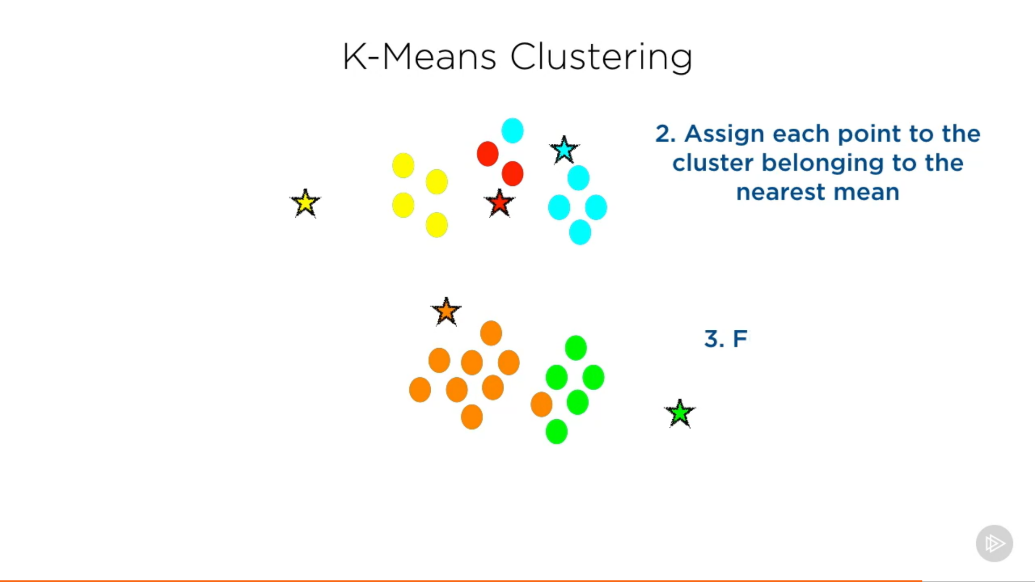
Weight the term frequencies to take the rarity of a word into account

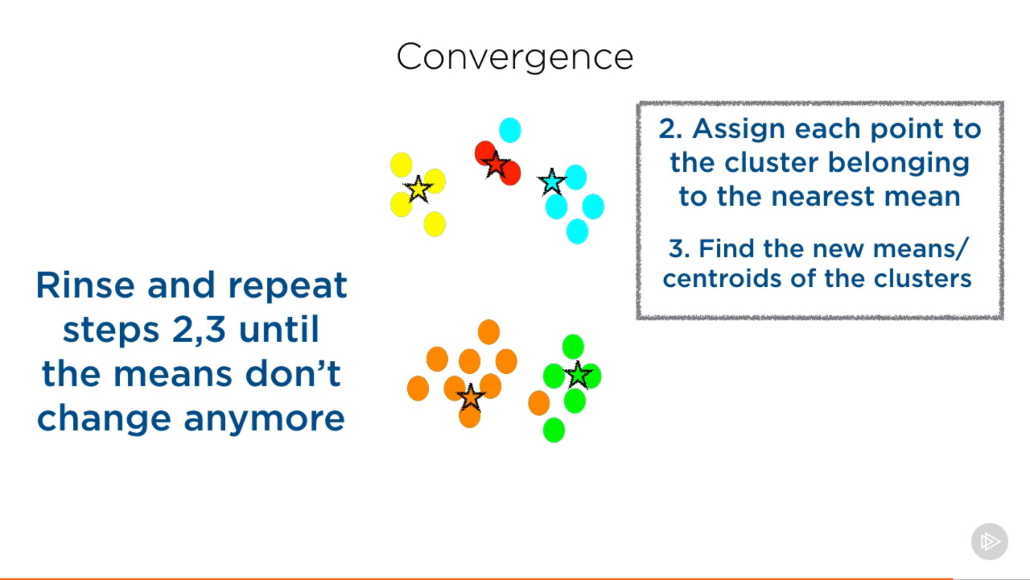
Weight = 1 / # documents the appears in

Term frequency – Inverse Document Frequency

K Mean Clustering – This algorithm will divide data into K clusters.(K is specified by clusters)







**Implementation:**

We will take the IMDB TV review dataset and see if there are some movie reviews that can be clustered to different groups.