**CSE 572 DATA MINING**

**Instructor- Prof. Arunabha Sen**

**ASSIGNMENT 2**

Group name: Sriharsha

Group ID: 15

Group members:

1. Sriharsha Uppu (1215318000)
2. Praveen Muruganandam (1215159608)
3. Chandrika Cuddapah Sudhinder (1215321133)
4. Nishant Trivedi (1214072053)

**Results for Task 1**:

Clustering is a type of **unsupervised learning**. This is very often used when you don’t have labeled data. **K-Means Clustering** is one of the popular clustering algorithm. The goal of this algorithm is to find groups (clusters) in the given data.

Uses

K-Means is widely used for many applications. Some are

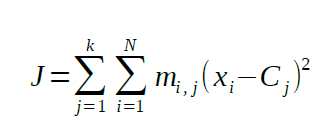
* Image Segmentation
* Clustering Gene Segmentation Data
* News Article Clustering

Assuming we have inputs  *x*1​,*x*2​,*x*3​,...,*xn*​ and value of **K(2<=K<=15), we follow the below steps:**

* Pick K random points as cluster centers called centroids.
* Assign each xi to nearest cluster by calculating its distance to each centroid.
* Find new cluster center by taking the average of the assigned points.
* Repeat Step 2 and 3 until none of the cluster assignments change.

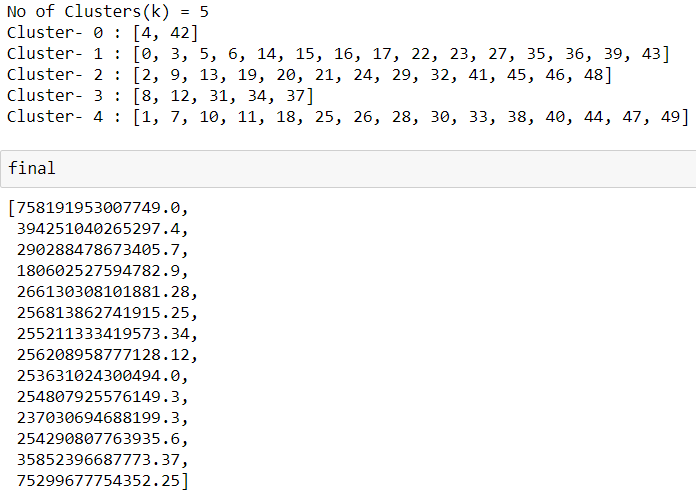
For each value of k, upon convergence (mentioned below) of your K-means algorithm, compute the objective

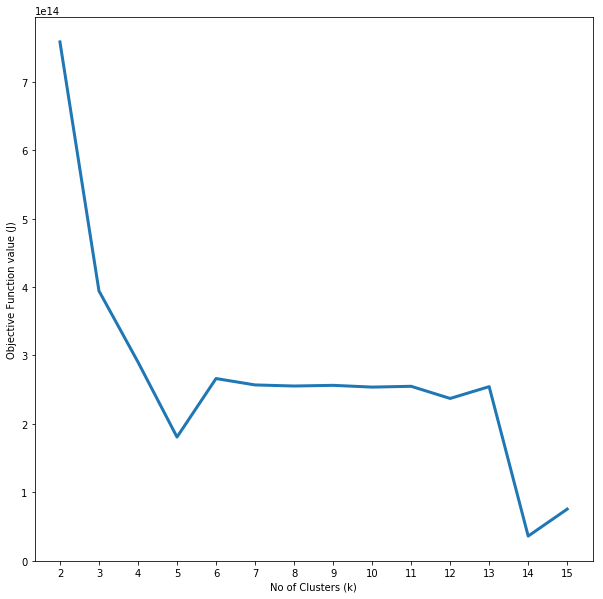
function value :



where N is the number of clusters, mi,j represents the cluster membership and Cj represents the jth cluster center.

We also have to plot a graph between the objective function value(J) and number of clusters(K).





**Results for Task 2**:

Using the Data obtained from overdoses.csv file, we construct a similarity matrix representing the closeness of state pairs with respect to the population and death values with cosine similarity as the metric.

Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. Smaller the angle , higher the similarity.

We apply K-means algorithm devised in the first task to the similarity matrix we constructed.

**Results for task 3**:

For grouping similar literary items together based on subject topic we find that **Cosine Similarity** metric is more accurate. The reasons are given below:

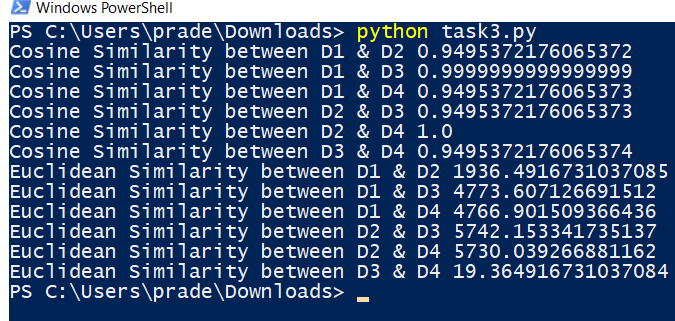
**Two** clusters/groups are formed:

**Book on Data Science and Article on Data Science are similar. (1 and 3).**

**Book on Soccer and Article on Soccer are similar (2 and 4)**

They are evaluated and checked by both **Cosine Similarity** and the **Euclidean Distance** metrics.

* **Cosine Similarity metric:**
* Cosine Similarity is calculated using only the dot product and magnitude of each vector, and is therefore affected only by the terms which the two vectors have in common,
* The cosine similarity is better at catching the semantic of each text, the direction the text points can be thought as its meaning, so texts with similar meanings will be similar.
* Cosine thus has some meaningful semantics for ranking similar documents, based on mutual term frequency.
* An important property of the cosine similarity is its independence of document length.. This happens for example when working with text data represented by word counts. We could assume that when a word occurs more frequent in document 1 than it does in document 2, that document 1 is more related to the topic. This happens when we are working with documents of uneven lengths.



* **Euclidean distance :**

Euclidean has a term for every dimension which is not zero in either vector. Another reason is that when modeling texts as vectors you will have more dimensions, the Euclidean distance metric is not good for high dimensional data.

**Cosine similarity** is generally used as a metric for measuring distance when the magnitude of the vectors does not matter. This happens for example when working with text data represented by word counts. We could assume that when a word (e.g. science) occurs more frequent in document 1 than it does in document 2, that document 1 is more related to the topic of science. However, it could also be the case that we are working with documents of uneven lengths. Then, science probably occurred more in document 1 just because it was way longer than document 2. Cosine similarity provides a correction for this.

You might also apply cosine similarity for other cases where some properties of the instances make so that the weights might be larger without meaning anything different.