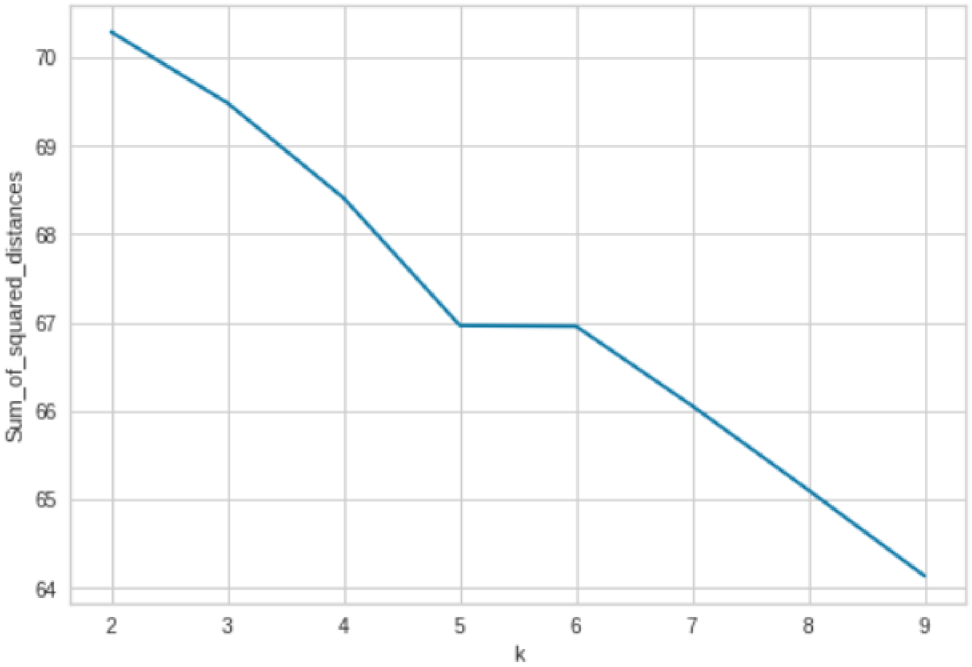
IT350 LAB 4

Shashikant kumar

191IT249

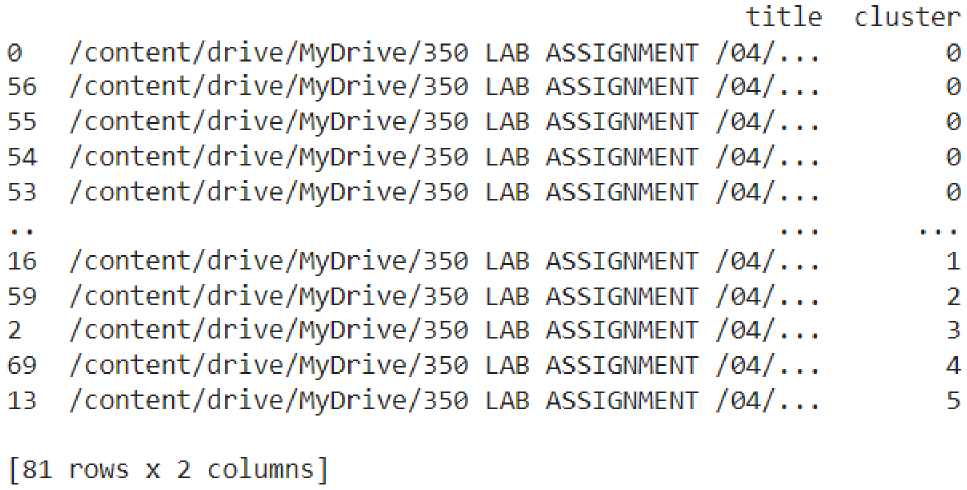
1) Find the clusters in the given dataset based on the content similarity and image similarity using,

Choosing appropriate k:



Using elbow method, we take k = 6;

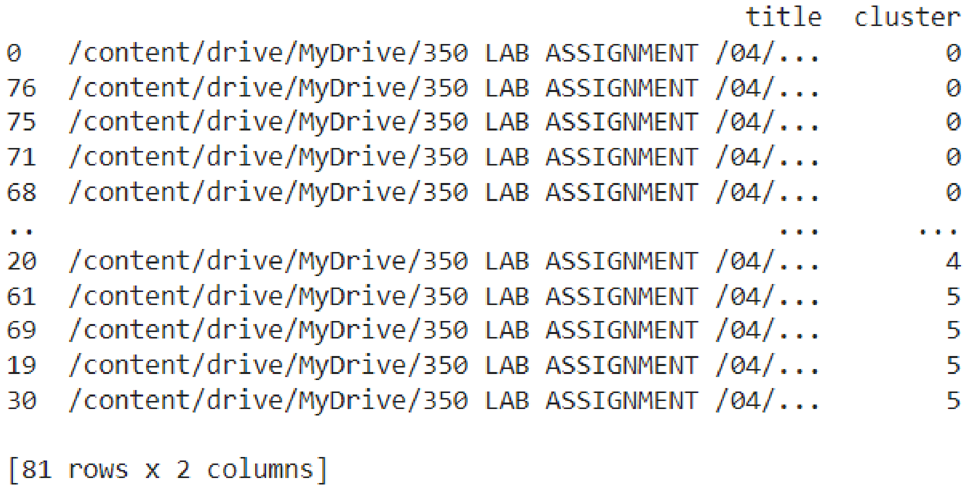
K-means clustering:



```
                                               title  cluster
0   /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
56  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
55  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
54  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
53  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
..                                                ...      ...
16  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        1
59  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        2
2   /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        3
69  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        4
13  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        5

[81 rows x 2 columns]
```

Hierarchical clustering methods:



```
                                               title  cluster
0   /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
76  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
75  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
71  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
68  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        0
..                                                ...      ...
20  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        4
61  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        5
69  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        5
19  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        5
30  /content/drive/MyDrive/350 LAB ASSIGNMENT /04/...        5

[81 rows x 2 columns]
```

Text similarity analysis

Text similarity has to determine how ‘close’ two pieces of text are both in surface closeness [**lexical similarity**] and meaning [**semantic similarity**].

To consider **semantic similarity** we need to focus on **phrase/paragraph levels** (or lexical chain level) where a piece of text is broken into a relevant group of related words prior to computing similarity.

Since **differences in word order** often go hand in hand with **differences in meaning** , we'd like our sentence embeddings to **be sensitive to this variation**.

K-means and Hierarchical Clustering Dendrogram:

With K-mean related algorithms, we first need to convert sentences into vectors.k-means (and minibatch k-means) are very sensitive to feature scaling

Reducing the dimensionality of our document vectors by applying latent semantic analysis

Image similarity

Similar to text similarity. Basic process description.

Using k-means/hierarchical clustering

Pre-trained model

Assemble dataset

Pre-process data : similar sizes for images

Train on model

Feature extraction

Dimensionality Reduction using t-SNE

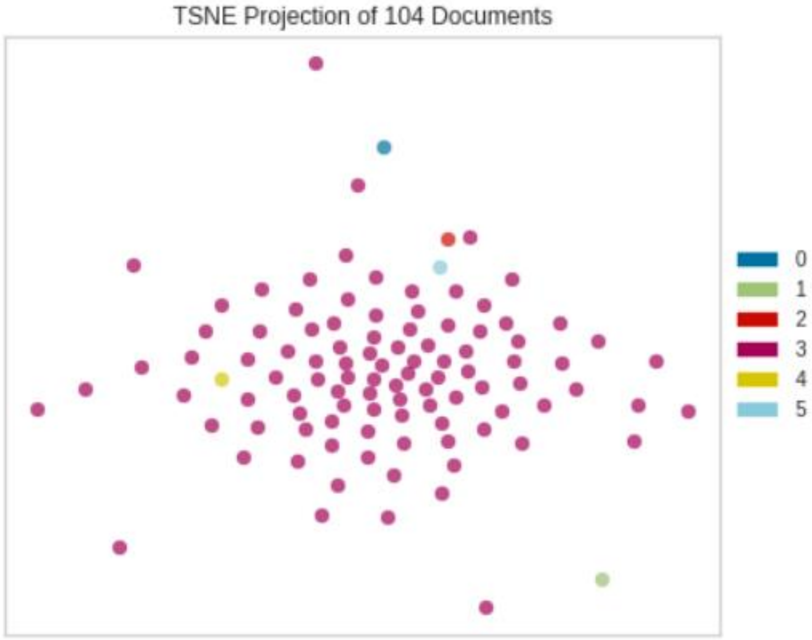
k-means clustering/Hierarchical clustering

2) Plot t-SNE visualization for derived clusters.

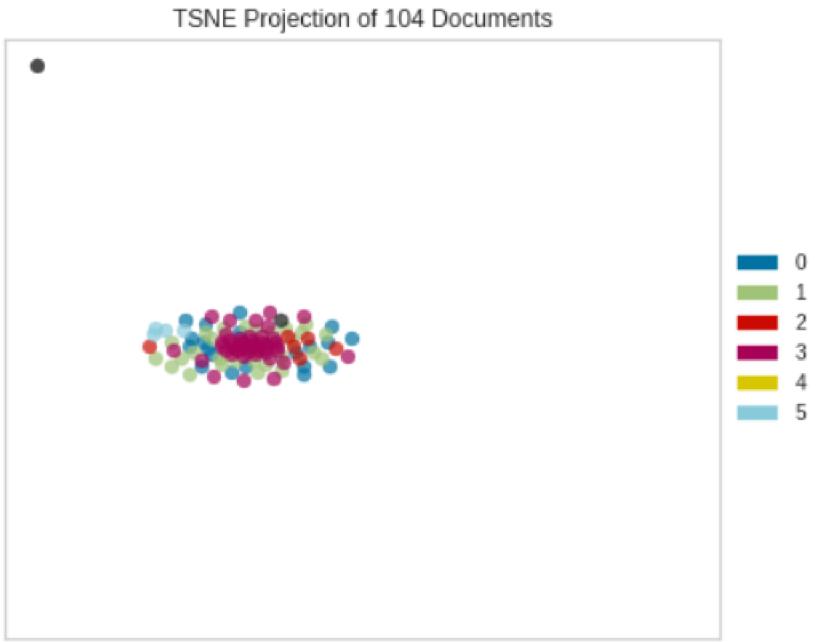
t-SNE is something called **nonlinear dimensionality reduction**.

t-SNE is mostly used to understand high-dimensional data and project it into low-dimensional space

For K-means



For Hierarchical



Conclusion:

t-SNE is a great tool to understand high-dimensional datasets. It might be less useful when you want to perform dimensionality reduction for ML training (cannot be reapplied in the same way). It's not deterministic and iterative so each time it runs, it could produce a different result.

3) Evaluate the clusters that are obtained using appropriate methods.

The Dataset is clustered into 6 clusters:

Namely, 0,1,2,3,4,5

For K-means:

The number of papers in following cluster are:

0 – 76
1 – 1
2 – 1
3 – 1
4 – 1
5 – 1

For Hierarchical:

The number of papers in following cluster are:

0 – 15
1 – 18
2 – 4
3 – 37
4 – 3
5 – 4

Some important factors for evaluating clusters:

*(a) Clustering tendency (b) Number of clusters, **k** (c) Clustering quality*

(a) Clustering tendency

Null Hypothesis and Alternate Hypothesis

(b) Number of optimal clusters, k

If *k* is too high, each point will broadly start representing a cluster and if *k* is too low, then data points are incorrectly clustered. Finding the optimal number of clusters leads to granularity in clustering.
The approach:

Domain knowledge, Data driven approach
Elbow method could be used to find the optimal k. Within-cluster variance is a measure of compactness of the cluster. Lower the value of within cluster variance, higher the compactness of cluster formed.

Sum of within-cluster variance, *W*, is calculated for clustering analyses done with different values of k. *W* is a cumulative measure how good the points are clustered in the analysis. Plotting the *k* values and their corresponding sum of within-cluster variance helps in finding the number of clusters.

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

Google Collab Drive Link (Code):

https://colab.research.google.com/drive/1ZoXGr8pWUbvuXX2_hV1RJB6Il560FQSY#scrollTo=w99OqDF56n_5