**ML/DL Task - Clustering and Title Generation**

You are provided with a set of news articles collected from several websites. The task is to create a python module using machine learning/NLP/Deep learning techniques to cluster the documents with similar content and generate a title for the cluster which represents these news articles. Number of clusters should be customisable by the user and it can contain anywhere from 1 to 100 cluster points. You are free to use any kind of open-source code/model but should give proper justification on why you choose the same. The title generating should not be an extractive one, it should be *generative* i.e any selected sentence among the articles cannot be considered as the title for the cluster.

**Approach:**

Data Cleaning and Pre-processing

Json file into Csv file

Merged Json file

K-Means Clustering

Markov Chain Generator

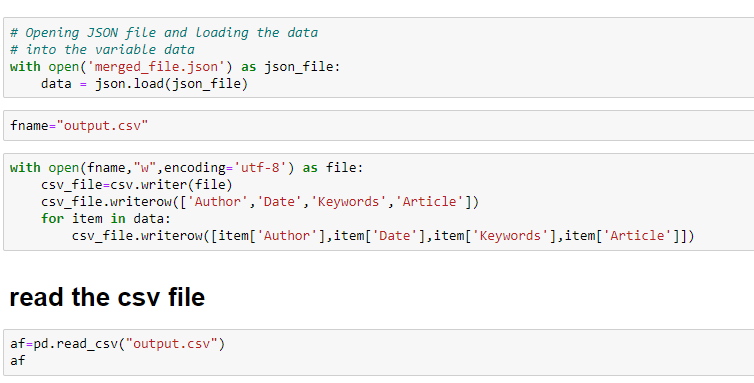
TF-IDF Model

Title Generated

Have we have 1000 document in json format as each files. Firstly just merged those 1000 documents into one common json file.

****

then standard machine learning document reading common json file converted into csv file for simplicity to read and fetch data more effective



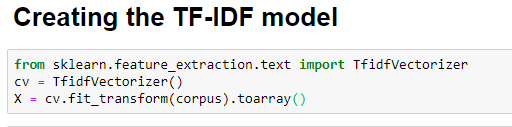
**Data Cleaning and Pre-processing**

Stemming ,lemmatization are the basic NLP pre-processing

**Stemming and Lemmatization** in Python NLTK are text normalization techniques for Natural Language Processing. These techniques are widely used for text preprocessing. The difference between stemming and lemmatization is that stemming is faster as it cuts words without knowing the context, while lemmatization is slower as it knows the context of words before processing.

**TF-IDF Model**

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.

****

Machine learning algorithms cannot work with raw text directly. Rather, the text must be converted into vectors of numbers. In natural language processing, a common technique for extracting features from text is to place all of the words that occur in the text in a bucket.

**Clustering**

It is an Unsupervised machine learning problem. because it is not labelled .here we have to cluster the news articles into different clusters or groups

**K-Means Clustering**

K-Means Clustering is a simple yet powerful algorithm in data science

There is an algorithm that tries to minimize the distance of the points in a cluster with their centroid – the k-means clustering technique*.*

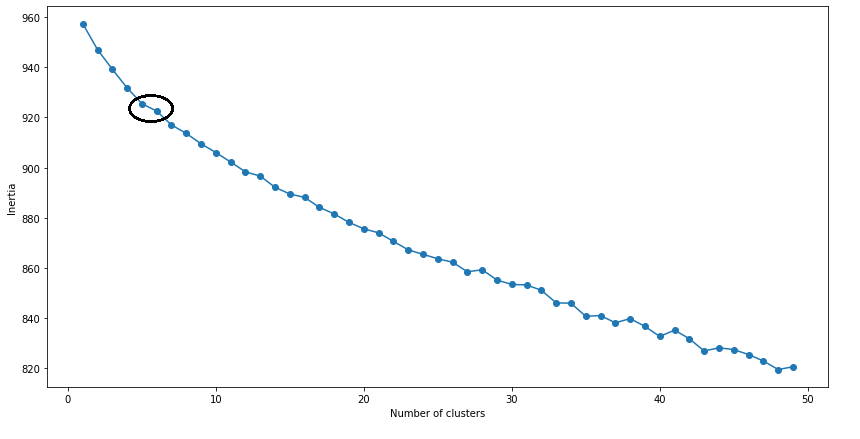
Relatively simple to implement.

Scales to large data sets.

We have to find and choose k value which the no. of cluster we need

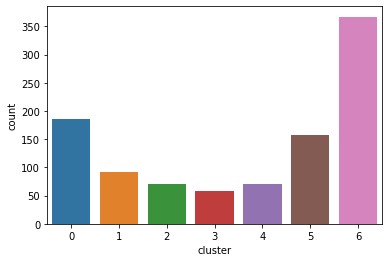
**K=**no. of clusters

we plotting Elbow Curve



From this plot we concluded the no.of clusters. To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion.

Here, we can choose any number of clusters between 7 and 9. We can have 7,8 or even 9 clusters. You must also look at the computation cost while deciding the number of clusters. If we increase the number of clusters, the computation cost will also increase.

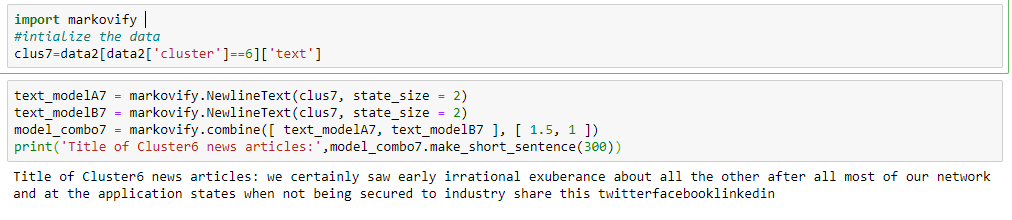


**Title Generation**

for Text generation,here we are using Markov chain generator

Markovify is a simple, extensible Markov chain generator. Right now, its main use is for building Markov models of large corpora of text, and generating random sentences from that

A Markov chain is a **process** that models a sequence of events in which the probability of each event **depends on the state of the previous event**. The model requires a finite set of states with fixed conditional probabilities of moving from one state to another

.