Assignment 4

So far you have learnt preprocessing data and applying various classification and regression techniques.

This assignment is divided in 3 parts.

1. **Part 1** – **CLUSTER ANALYSIS**.
   1.1. The purpose of clustering and classification algorithms is to make sense of and extract value from large sets of structured and unstructured data. If you're working with huge volumes of unstructured data, it only makes sense to try to partition the data into some sort of logical groupings before attempting to analyze it.
   1.2. You have to perform KMeans and Hierarchical analysis on the IMDB dataset (refer to 1. In the resources below). The goal is to put all the movies that share some common characteristics in one cluster.
   1.3. For KMeans you will first have to find the optimum number of cluster by plotting the SSE vs # of Clusters (Elbow method) and then proceed with applying Kmeans.
   1.4. For hierarchical clustering, apply single, complete and average link and display the dendogram (the plot that visualizes the hierarchy).

2. **Part 2** – **TEXT MINING**.
   2.1. Text mining is the process of analyzing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends without requiring that you know the precise words or terms that authors have used to express those concepts.
   2.2. To make sense of the text and make it useful for various ML techniques, there has to be a numeric way to express the text.
   2.3. Your task is to create a count vector and a tfidf vector on the given data (refer to 2. In the resources below).
   2.4. Display the count vector and tfidf vector and explain the usage of tfidf.

3. **Part 3** – **ARTIFICIAL NEURAL NETWORK (ANN)**.
   3.1. An ANN is (supposed to be) exactly like a human brain but simulated using a software. Like a brain, a NN consist of various layers of "neurons" that work simultaneously to get the output.

3.2. Your task is to apply ANN on the admission dataset used in assignment 2 using code given in tutorial 6. You will have to make necessary changes to the code to make it work for the admission dataset. Hint: encode the target attribute values to binary values .

4. Resources

4.1. Find the IMDB dataset under Files -> Labs -> data and find the description of the dataset here: https://www.kaggle.com/iarunava/imdb-movie-reviews-dataset

4.2. Dataset for text mining: Please note this is a python list.

['Now for manners use has company believe parlors.',

'Least nor party who wrote while did. Excuse formed as is agreed admire so on result parish.',

'Put use set uncommonly announcing and travelling. Allowance sweetness direction to as necessary.',

'Principle oh explained excellent do my suspected conveying in.',

'Excellent you did therefore perfectly supposing described. ',

'Its had resolving otherwise she contented therefore.',

'Afford relied warmth out sir hearts sister use garden.',

'Men day warmth formed admire former simple.',

'Humanity declared vicinity continue supplied no an. He hastened am no property exercise of. ' ,

'Dissimilar comparison no terminated devonshire no literature on. Say most yet head room such just easy. ']