Winter Workshop on Data Science and Machine Learning

26th - 30th December, 2017

Task Sheet - Day 2

DataSet Preparation:

Extend the Twitter Task 1 in Day 1 as below.

For each tweet of username X (SrBachchan) collected, find the following. Consider only 100 tweets.

a. Length of tweet

b. Number of hashtags in tweet

c. Number of @ mentions in tweet

d. Likes received by the tweet

e. Retweets received by the tweet

f. Sentiment expressed in the tweet (refer TextBlob API)

g. Hour when tweet was posted, eg. If a tweet is posted at 7:35 pm, then hours = 19

Construct a CSV file (data matrix) whose each row contains tweetid as first column, following by seven columns which represents information about that tweet as above. In each row, the values of above features (a, b, c, d, e, f, g) are stored in comma separated format.

Note: For now, the above CSV file contains features of tweets belonging to only one user (X)

Data Analysis:

Uni-variate:

1. Find mean, median and standard deviation of all the above features of a tweet using R.
2. Compare and analyze the ranges of features (b,c,d,e) using BoxPlots. Plot all boxplots on same graph using R.
3. Compare and analyze probability mass functions (PMFs) of features (b,c,d) through PMF plots. Draw all PMF plots in same graph. You should draw the PMF plot using matplotlib module of python.
4. Compare and cumulative distribution functions (CDFs) of features (d,e,f,g) through CDF plots. Draw all CDF plots in same graph. You should draw the CDF plot using matplotlib module of python.

Bi-variate:

1. Find correlation value between all pairs of features (a,b,c,d,e,f,g) in the above dataset constructed using R.
2. Draw scatter plots between features (b,d), (c,d), (b,e) and (c,e) using R.

Graph Construction:

Each row in the CSV file constructed above contains features of a single tweet. Until now, we have assumed that all these tweets are not related to each other.

Now, let us change this assumption and we say that two tweets are related with each other based on some criteria. The 'relatedness' can be captured as an *edge* of graph. Each of the following relation can be used to construct a separate 'undirected' graph.

G1. Tweet 'a' is related to Tweet 'b' iff they share a common hashtag.
G2. Tweet 'a' is related to Tweet 'b' iff they share atleast m common words. (Construct different graphs for different values of 'm') Use at least following values of m = 1,2,3 (lets call the graphs as G2.1, G2.2, G2.3)

Graph Analysis:
Visually draw graphs G1 and all G2.x using either graph-tools module or networkX module of python. Compute the following metrics for each node (tweet) in graphs and save these features in CSV file where the first column represents nodeID followed by computed values of the following features:-

a. node degree
b. eccentricity of node
c. betweenness centrality of node
d. clustering coefficient of node

1. Compare and analyze the ranges of above node features (a,b,c,d) using BoxPlots. Plot all boxplots on same graph using R.
2. Compare and analyze probability mass functions (PMFs) of features (a,b,c,d) through PMF plots. Draw all PMF plots in same graph. You should draw the PMF plot using matplotlib module of python.
3. Compare and cumulative distribution functions (CDFs) of features (a,b,c,d) through CDF plots. Draw all CDF plots in same graph. You should draw the CDF plot using matplotlib module of python.

Compute following values (at graph level) for each of the graphs G1 and all G2.x

a. Average Path Length of graph
b. Radius and Diameter of graph
c. Number of Connected Components in graph
d. Average Clustering Coefficient of graph