Title of Your project

Your names and Roll Numbers

March 25, 2018

1 General Information about data

Discuss here briefly.

- Source of data. Not the source where you downloaded from, but the actual source where data is generated.
- Give your source of data, where you downloaded it form, e.g. a URL where you downloaded or crawled data from.
- Name some of the other work done on this paper, e.g. which paper you read to get idea about this datasets, or is the dataset provided by Kaggle, (how many people competed).
- What type of data is it? e.g. graph, text, tabular.
- What does a data object represent (a user, a book, a node in a graph, a student)
- Have you downloaded all datasets? If answer is "NO", give the reason.
- What is/are the size(s) of data on disk?
- What are the dimensions of data?
- Other size related properties to explain?

2 Data Format, Types and Description

Without repetition, give the answer of following questions regarding above mentioned heading. You can use table format.

- What is/are file types? like `.csv` etc.
- What are the attributes in each dataset?

- Detailed description of data:
  - What is size of data (number of instances),
  - dimensionality of data,
  - data types of each feature, description of features.
  - If you are working on graph data, then for instance number of nodes, number of edges, directed (or undirected) etc.
  - You have to do it for each data set; if you are using more than one type of data
  - You may use a table here if the description takes too long.
- Also give the answer of any other (missing) questions falling in this section, that you think is appropriate

3 Data preprocessing

- Does your data set require preprocessing?
- If you have done preprocessing on your data sets, mention what were those.
- Did your data set contain missing values?
- For instance if you are working on text data, you must have removed stop words, if the data is in XML format, you must have converted into relational data etc.
- Which technique(s) did you apply for filling the missing values?

4 EDA

- You should report average, median, minimum, maximum, variance of numerical features.
- If you are working on graphs you should report for each graph the average degree, density, maximum degree, minimum degree etc.
- You have to plot histograms/bar graphs, box-and-whisker diagrams, five-number summaries, scatter plots too see correlation between attributes.
- If you are using multiple data sets, you have to repeat the above points for each dataset.