Week 6: Daily Morning Challenge

Day 1: Tuesday 28th January 2020

Question 1: How would you optimise a web crawler to run much faster, extract better information, and better summarise data to produce cleaner databases?

Answer:

The web being a vast ocean of data, the possibilities it opens to the business world are endless. However, extracting this data in a way that will make sense for business applications remains a challenging process. Web crawling and data extraction is something that can be carried out through more than one route. In fact, there are so many different technologies, tools and methodologies you can use when it comes to web scraping. However, not all of these deliver the same results.
A good web scraper requires very robust, multi-component architecture that is fault tolerant. The retrieval logic can be complicated since the data can be in different format. A typical application based on web scraper requires regular maintenance in order to function smoothly. By Optimizing the DB query for improved time complexity of the whole system, we essentially reduce the fetch time to merely a fraction of seconds.

Question 2: Briefly explain how to use the five number summary to perform descriptive analysis on a dataset?

Answer:

The five-number summary is a set of descriptive statistics that provides information about a dataset. It consists of the five most important sample percentiles. There are a variety of descriptive statistics such as the mean, median*, mode,* skewness*,* kurtosis, *and* standard deviation and so on. Rather than looking at these descriptive statistics individually, combining them helps to give us a complete picture. The five-number summary is a convenient way to combine five descriptive statistics. The five-number summary consists of the following:

- The minimum – this is the smallest value in our data set.
- The first quartile – this number is denoted Q_1 and 25% of our data falls below the first quartile.
- The median – this is the midway point of the data. 50% of all data falls below the median.
- The third quartile – this number is denoted Q_3 and 75% of our data falls below the third quartile.
- The maximum – this is the largest value in our data set.

The median, first quartile, and third quartile are not as heavily influenced by outliers.

Question 3: Outline 5 graphical methods that can be use understand and describe relationships between attributes in a dataset?

Answer:

(A) The histogram is another popular choice for displaying continuous data. A histogram looks similar to a bar chart, but in a histogram, the bars (also known as bins because you can think of them as bins into which values from a continuous distribution are sorted) touch each other, unlike the bars in a bar chart. Histograms also tend to have a larger number of bars than do bar charts.

(B) Line graphs are also often used to display the relationship between two variables, usually between time on the x-axis and some other variable on the y-axis. One requirement for a line graph is that there can only be one y-value for each x-value

(C) Scatterplots define each point in a data set by two values, commonly referred to as x and y, and plot each point on a pair of axes. Scatterplots are a very important tool for examining bivariate relationships among variables.

(D) The Countplot is kind of like a histogram or a bar graph for some categorical area. It simply shows the number of occurrences of an item based on a certain type of attribute. It show the counts of observations in each categorical bin using bars

(E) The Pie chart presents data in a manner similar to the stacked bar chart: it shows graphically what proportion each part occupies of the whole. Pie charts, like stacked bar charts, are most useful when there are only a few categories of information and the differences among those categories are fairly large.