Sameer Rau

CSCI 4243W Project Summary (Data Visualization)

Big data has become an increasingly difficult problem for data scientists and others working with very large data sets who want to discover correlations between several variables, and display such trends both computationally and graphically. This is an issue that is not expected to slow down anytime soon. According to Bernard Marr of Forbes, roughly 1.7 megabytes of new data will be created every second for every person on Earth by 2020, and currently only 0.5% of the data has ever been actually studied. One way to cut down this problem to a degree would be to interpret findings for several related data sets at once, and furthermore automate the process as much as possible which would cut down the time it would take to look at each individual spreadsheet. As part of the data visualization project, the proposed input would be fairly simplistic; the user would simply attach all of the data sets they wish to perform analysis on and the process would return extensive output detailing the appropriate tests to run on variables that are related with strong correlation (which could potentially exist between variables in different data sets), the measures of central tendency, and a graphical representation of the correlation.

This Small Business Innovation Research Phase I project would enable the aforementioned ability to cut down the amount of time each person is required to look at the data sheet in order to find trends among the variables at hand. Furthermore, the decision-making of what variables to study in order to find a relationship would no longer be a necessary task for employees as that would be made by the computer itself. With so many data sets at hand to study in a finite amount of time, the efficiency with which the spreadsheets would be looked at would increase and there would therefore be more opportunities provided for conclusions to be drawn and the time spent by people would be better spent by deciding what subsequent actions need to be taken. A primary potential technical hurdle that may be faced during the development of the project would be the format of the spreadsheets and the number of sheets being looked at spontaneously. An assumption early on may be to have formatted data with variable names in the first row and the corresponding values in the each column below the name to make it easier for the computer to differentiate between text and numerical values. The overall goal is to have as many spreadsheets to be studied as possible, although a potential challenge would be size limitations as computing capacity constraints should be taken into consideration as well. It may also be a challenge later on to ensure that the computer can differentiate between categorical (nominal) variables, binary variables, and continuous variables. An incorrect reading of the variable could lead to a suggestion for the wrong test to be conducted. The goal of the research and development stage is to come up with potential constraints that cannot be avoided (such as computing power) as well as errors that could be faced later down the line so when a certain component of the project has been reached a possible work around or solution will have already been developed. For each of these steps, it may be beneficial to just look at many data sets to get a sense for format, size, and the types of variables that are generally studied. It may also be helpful to ensure that files are being read in correctly and no internal errors exist when importing the spreadsheet.

If the Data Visualization project were to be commercialized, it would be a tool that would not only be used by data scientists, but in all fields as data is present everywhere and can be determinant of what can be improved upon in every organizational context. It would be helpful for an analytics team, for example, to just be able to obtain the results of the data at hand and express the significance of the findings to other teams at a company. The project would enable further technological understanding and potentially open up the idea of using computers to optimize time spent on looking at data thus freeing up employees to allocate that time elsewhere. Societal needs can be better met and gauged as more conclusions are drawn from looking at a greater number of data spreadsheets simultaneously.

Project Summary [One (1) page MAXIMUM]. The Project Summary should be written in the third person, informative to other persons working in the same or related fields, and, insofar as possible, understandable to a scientifically or technically literate lay reader. It should not be an abstract of the proposal. Do not include proprietary information in the summary. Proposals that do not contain a complete Project Summary will not be accepted by FastLane or will be returned without review. The Project Summary is completed in FastLane by entering information into 3 text boxes; the aggregate of the 3 text boxes cannot exceed 4,600 characters (including spaces). Please note that the character count function of some word processors excludes hidden characters that WILL be counted by FastLane. Therefore, proposers may consider aiming for fewer characters (e.g., 4000) instead in order to avoid having to slim down the Project Summary later.

Box 1: Overview, Key Words, and Subtopic Name: Describe the potential outcome(s) of the proposed activity in terms of a product, process, or service. Provide a list of key words or phrases that identify the areas of technical expertise to be invoked in reviewing the proposal; and the areas of application that are the initial target of the technology. Provide the subtopic name.

Box 2: Intellectual Merit: This section MUST begin with "This Small Business Innovation Research Phase I project". Address the intellectual merits of the proposed activity. Do not include proprietary information in the summary. Briefly describe the technical hurdle(s) that will addressed by the proposed R&D (which should be crucial to successful commercialization of the innovation), the goals of the proposed R&D, and a high-level summary of the plan to reach those goals.

Box 3: Broader/Commercial Impact: In the short term, the proposed R&D activity is expected to bring the innovation closer to commercialization under a sustainable business model. In this box, describe the potential impacts on society that would be created by the commercialization of the innovation. Examples include generating larger economic impacts, meeting societal needs, and enabling further scientific / technological understanding.

https://capstone.cs.gwu.edu/2015/lib/exe/fetch.php/writing-1-rubric.pdf