Reflection Paper CS457—Data Mining

Phoebe Priscilla Amoako, Student

**Class Overview**

Before the Spring Semester, I was enthused about taking a class in Data Mining because I decided to make that my area of research for my capstone project. I loved the following things about the class:

* **Topics Covered**: The class covered the concept of regression, community detection and clustering, graphs etc. which have widened my scope of data mining in general. The result is that I am able to read articles and online forums under data mining and understand them well enough.
* **Assignments and Quizzes**: These assessments achieved their aim of getting us to understand class concepts previously taught, especially the quizzes.
* **Class Material**: Wiley’s textbook on Data Mining was beneficial to the class and enabled us to get a good grasp of excel because of the many step-by-step examples used. Also, the sense of humor in the book made it an easy read.

Below are some of the things that could be improved or included in subsequent sessions of the class:

* **Class Material**: Maybe another textbook on the R programming language should be used alongside with Wiley’s. This will help students get up to speed with R.
* **Assessments and Quizzes**: Most of the assignments should be based more on R than excel. This will be a good way to get students up to speed with the R language and understand how to go about analyzing huge chunks of data with it.
* **Research and New Developments**: Data mining is a relatively new field and there are new explorations springing up. I am not sure about how feasible this will be but students from the onset of the class could pick up areas of interest like health, education, economics etc. and find out how data mining is being used in such areas. This will be beneficial in helping students apply whatever concepts are being studied in class and their relevance to the real world. It could even be a good way to get students involved in class discussions by sharing what new insights in their allocated areas.

**My project**

After very long deliberation, I decided to analyze Facebook group data on the following institutions: Ashesi University College, Wheaton College and Olin School of Engineering. My choice was influenced by the fact that all three colleges were liberal, small and partnered whether in terms of curriculum or study abroad programs. I wanted to find out how gender influenced user activity on the Facebook pages for these universities and the extent to which status types influenced user activity.

In the course of my project here are some challenges that I had:

* **Data Cleaning**: This topic was not emphasized in class and did not prepare me for the amount of work that I had to do. Facebook provided me with a text file which I had to transport into a csv format and eventually clean up to be able to use the data to answer my research questions. After trying to clean my data manually took about 70% of my time and still proved futile, I decided to learn how to use R to do the cleaning which was more difficult than I anticipated.
* **Getting Around R**: My project made me realize how I had not gotten a good enough grasp in R. After my project, I look back and realize how much quicker I could have gotten around R if I had practiced with it sooner so that learning new R techniques would not take as much time**.**
* **Limited Data:** I felt that the data I had did not have as much information as I would have preferred. I would have liked to know which active users on the page were students of the selected colleges but due to Facebook privacy issues, that was impossible to do. In the long run, it made my analysis weaker than intended.
* **Time**: I would have appreciated it if the project was began at the beginning of the semester to enable more time for learning about new tools and exploration.

Fortunately some of the R scripts that the class used gave me a good grasp of how to get started with writing my very own R script.

In general, I have found the class really interesting and exciting. For the first time this class has been run, I believe that we have covered enough to make me have data mining discussions with an expert and equip me to get started with data mining on my own. If I am around during the next time Data Mining is offered as a course, I will be glad to sit in the class!