**BIG DATA – LESSONS LEARNED**

**Diana Amador**

The most challenging part of this lesson was using Amazon Web Services and running the Job streaming necessary to generate de Big Data Matrix.

In task 1 I experienced a series of problems to identify the Common Crawl Data that was suitable for the task. Some of the segments I chose generated an input error when I ran my mapper-reducer scripts. I was able to identify new ones and resubmit the job streaming.

In task 2, it was essential to understand how hadoop and AWS work, and learn how to read the logs and find solutions to errors using Stack Overflow. The proper use of Stack Overflow is most if one really wants to hack a problem.

For task 2, I have learned some python on my own and that allow me to modify the mapper and reducer scripts. That saved me a lot of time and trouble.

In task 3, I was familiar with the attribute selection and model testing; however, it was important to use an heuristic approach to avoid excessive testing that would have led me to over testing. I learned to pay more attention to balance the weight of all of the performance metrics. That prevented me from going into exhaustive testing in the wrong direction.

Task 4 put together everything. It provided a perfect example of what a data mining / predictive analysis would be in the real world, and in that sense I learned how to deal with real problems and to find solutions that may not be academic at all but solved my problem and allowed me to go forward.

My best advise to anyone taking over this kind of tasks is to understand the data structures and metrics when measuring performance. Understand that it is not the same working with a sample than working on Big Data. A small difference will have a huge impact, for example the time to build a model.

Another consideration is dealing with parallel processing. It is essential to know what the scripts are doing. I think it is a mistake to run scripts you don’t understand. Learning python or javascript is a most to control what the mapper is doing with the data. I think this lesson demands python knowledge.

Last, but not least: pay attention to details. A letter, a dot, pretty much a small detail can and will derail the job.