Kenneth Luna

Independent Research Proposal

The purpose of this independent study is to enhance and gain exposure to one of the most state of art data processing frameworks, Apache Spark. Gaining exposure to Spark will further improve my goal of knowing how to manage and efficiently process massive amounts data. In Prof. Freire’s Big Data class, we mastered using Hadoop Clusters; dedicating my summer to using Spark will only complement my ability to work with large distributed networks to process massive amounts of information. Furthermore, my job at Two-Sigma will require me to work with Spark often and I’d love to start my new career with a certain level of know-how with this framework. I have also spoken to faculty at NYU CDS to explore potential datasets we could use with Spark.

The learning objectives of this course are to master the Spark framework and develop working knowledge of Java and/or Scala. Much of our programming at CUSP has been mainly in Python and R; I’d like to have a solid working knowledge of Java and Scala by the time I finish my summer semester in late June. This will also provide me the opportunity to improve my exposure to Object-Orientated Programming (OOP).

The faculty member who will be advising my project is Huy Vo. We spoke about this independent research project on May 5th 2015. He is currently reaching to IT to explore ways we can leverage CUSP’s current IT infrastructure to run Spark on the Hadoop Cluster. What is great about this independent research project is that it will be a learning experience for both of us as the framework is very nascent.

I expect the final deliverable to be a research project comparing the performance differences between Hadoop and Spark on large urban datasets. Huy and I are currently exploring dataset we believe will be provide interesting insights for this case study. We will also like to apply specific types of ML algorithms to see if we can see improved performance between both frameworks.