Class 4: Homework

Realtime and Big Data Analytics

Homework

Class 4

---

1. **Read the material in the Analytics Project Description packet provided on NYU Classes in file:**
   BigData_ProjectInfo.pdf

2. **Reading:** Please read the analytics paper provided in the Resources tab (also available at: http://online.liebertpub.com/doi/pdfplus/10.1089/big.2014.0061): "Targeting Villages for Rural Development Using Satellite Image Analysis" by Kush R. Varshney, George H. Chen, Brian Abelson, Kendall Nowocin, Vivek Sakhrani, Ling Xu, and Brian L. Spatocco.

3. **Analytics project – Propose an Analytic! (This is being assigned early so you can start to think about possible projects.)**
   This is an individual homework assignment. **It is not a team assignment**. Please develop this analytic idea independently.

   An analytic provides actionable insights through analysis of one or more datasets. Think about a problem you'd like to solve or study - think big!

   Write a project proposal describing an analytic that you would like to build. Record information about your proposal in the provided template file: **BigData_IndividualProjectProposal. pptx** for Windows Powerpoint users, or **BigData_IndividualProjectProposal.keys** for Mac Keynote users.

   Identify at least 2 dissimilar data sources that you would use in your proposed analytic.

   Some things to consider about your data sources:

   - How can you obtain a copy of the data? Who owns the data? Is it open data? Is it public or private?
   - How much data is there in each source – just magnitude – is it MB? GB? TB?
   - If the data is very large, where can you store it?
   - How do you gain access to the data? How long will it take to get access/be approved?
   - Who must you ask for permission to access the data?
   - Is the data static, periodic, near realtime?

     If it's static, you will copy it once, the source data will not grow.

     If it's periodic, you will copy it periodically, which means your files will be growing over time.

     If it's near realtime, you will be continually reading in the data, which means your data will be growing over time.
   - If the data is being collected in near real-time, what is the velocity of the data and the volume per unit of time? Can your Hadoop environment support this?

   This may or may not turn out to be the analytic that you ultimately build with your team. Use your imagination, take a risk.

   You will form teams in the coming week. Teams will select one of the analytics proposed by team members, or design an entirely different analytic to implement for the team analytics project.

Homework

Class 4

4. Try out Twitter

a. Access Twitter by downloading and installing the appropriate library for your programming language. For example, if you program in Java, install a stable version of Twitter4J (documentation is at http://twitter4j.org/en/index.html).

Information about Twitter's APIs is here: https://dev.twitter.com and https://dev.twitter.com/overview/documentation

Libraries are here: https://developer.twitter.com/en/docs/developer-utilities/twitter-libraries.html

Code examples (for Twitter4J) are here: http://twitter4j.org/en/code-examples.html

b. Develop a program that gets recent tweets (either through the Streaming API or the REST API) and outputs them to a file. **Upload only your program (your source code) to NYU Classes - remove any keys or private information. Do not upload any input or output data (no tweets).**

Note: This is not a Hadoop assignment, but it has proven helpful in the past. By completing this assignment, you will have a reliable data source, and potentially a near realtime one, as a backup in case other data sources that you planned to use turn out to be unavailable for your analytic.