Big Data 2017 - Project Overview

Deadlines:
March 25th: Submit group information
April 10th: Part 1 is due
May 1st: Part 2 and final report due

Project Mechanics

You will form a group of 3 people. Groups with more than 3 members are not allowed. If you want to have a smaller group, you need to ask permission to do so (contact instructors through NYU Classes).

You must create a:
- github repo for your project
- Google Document that describes the goals of your project, the questions you are investigating, and what you have done so far. This document should be updated as your work progresses, and will serve as the basis for the final project report. The document should be created using your NYU account and should be readable by anybody that has a link to the document

Submit your group information in the following google form:
https://goo.gl/forms/Q7IHZDYQ0q3AyIoG2
Note that you should only submit *one* form per group.

Project Goals

The goal for this project is to give you hands on experience with analyzing real, big data and apply the concepts you have learned. Each group will select one data collection from a list (see Data Collections below). For the first part of the project, you will analyze the data and generate a descriptive summary of their contents as well as a list of data quality issues. For the second part, you will integrate your selected collection with one or more data sets and look for interesting relationships between the data sets. A list of potentially related data sets is provided below and you are encouraged to look for other data.

Hadoop Stack: A key component of your project is to get hands-on experience with the techniques you are learning in the course. Therefore, you must use the NYU Cluster and the Hadoop stack. You can choose to write map-reduce, Spark, SparkSQL programs, or any other tools that works on Hadoop or Spark.

Reproducible results: All components of your project must be reproducible! Your code must be available on github, and all your analysis results and plots need to be accompanied by the appropriate scripts you used to derive them. Others should be able to reproduce your results. You should include a README with instruction to run your code and generate the results, plots and visualizations.

Data Collections

1. *Transportation*
   Yellow taxi (2013-2016)
   Green taxi (2013-2016)
   Download: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

2. *Complaints*
   311 (2009-2016)
   Download: (2010-2016) https://data.cityofnewyork.us/Social-Services/311/wpe2-h2i5
   (2009) https://data.cityofnewyork.us/Social-Services/new-311/9s88-aed8

3. *Crime*
   NYPD (2006-2016)
   Download:
   https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

Part I

To understand the data collection you have selected, you will first explore it and generate a summary of the data. The summary should provide an overview of the data and its contents. For example, a good summary for the Transportation collection can be found at: http://www.nyc.gov/html/tlc/downloads/pdf/2016_tlc_factbook.pdf
(If you select the Transportation collection, your summary must go beyond what is provided in the TLC Fact Book!)

You will also look for data quality issues as well as anomalies. Some questions you should ask include but are not limited to:

- Are there non-empty values that represent missing data (e.g., NULL, N/A, UNSPECIFIED, TBA, (999)999-9999) If so, how many in each column?
- Are there different kinds of values in the same column (e.g., integers and strings)?
- Are there suspicious or invalid values in columns? (e.g., a negative value in a price field; for spatial data, coordinates outside the city perimeter; an invalid zipcode)
- Are there surprising (or suspicious) events in the data? (e.g., a day with too few taxi trips or too few noise complaints). Note that you may have to aggregate the data in different ways to search for these events.

Some techniques that may be useful for finding unusual features and outliers include:

- Regression Analysis:
  Java/Scala: https://commons.apache.org/proper/commons-math/userguide/stat.html
  Python: http://scikit-learn.org/stable/modules/linear_model.html

- Box Plot:
  Python: http://matplotlib.org/examples/statistics/boxplot_demo.html

- Correlation:
  - Pearson's Correlation
    Java/Scala: https://commons.apache.org/proper/commons-math/userguide/stat.html
    Python: http://docs.scipy.org/doc/scipy/reference/stats.html
  - Mutual Information
    Java/Scala: https://github.com/jlizier/jidt
    Python: http://scikit-learn.org/stable/modules/classes.html

Note: These are just recommendations. You are allowed (and encouraged) to play with other techniques and implementations.

Deliverable: A report with the data summary and data quality issues submitted to NYU Classes. For each column in the data set, you should provide a script that assigns to each value:
1) a base type (i.e., INT/LONG, DECIMAL, TEXT, maybe DATETIME)
2) a semantic data type (e.g., phone, address, city, state, zipcode) -- here, your script could use, e.g., a regular expression to identify phone numbers, a dictionary to check whether a given string is a city.
3) a label from the set [NULL -> missing or unknown information, VALID -> valid value from the intended domain of the column, INVALID/OUTLIER -> suspicious or invalid values]
Note that some columns may have values of multiple types (e.g., telephone, email).

You must include a summary for (1), (2) and (3) in your report. You should include in your summary other data quality issues covered during the 3/27 lecture on data cleaning.

The code/scripts you use in your project must be available in the github repo you provided. You must include a README file with detailed instructions on how to run your code/scripts. Any figures and results in the report must point to the script you used to derived the result. See below for details on the structure of the report.

Part II

Now that you understand your data collection, you will start by generating a series of hypotheses about the data. In particular, you should try to explain any issues you found in Part I. You will then identify data set(s) that can help you prove of disprove your hypotheses, and analyze the integrated data.

Examples of hypotheses:

- Suppose you found days when there were fewer taxi trips than expected. You may posit that this was due to bad weather. To test the hypothesis about bad weather leading to the reduction in number of taxi trips, you could look at the distribution of the number of taxi trips per day and the distribution of precipitation amounts.
- Suppose you found out that the crime rate has been increasing in a region of Manhattan. You may wonder what the reason could be, and try to look how the census data of the same region changed over the years.
- Suppose you found out that, in some days, the number of 311 noise complaints is significantly larger for a specific region. You may posit that this is due to a high number of vehicle collisions or traffic in the same region.
- Suppose you found out that, for some days, the number of 311 heating complaints was more than 10 times higher than expected. You may posit that this is associated to abrupt temperature drops, and to test this hypothesis you could compare the distributions of temperature values and of number of 311 heating complaints per day.
- Suppose you found a single day where the number of 311 water system complaints is abnormally high. You might posit that this reflects a data quality issue, and to test this hypothesis you can look at additional distributions around the same time interval -- for example, number of 311 plumbing complaints, number of 311 traffic complaints, temperature values.

Note that:

1) relationships among data sets can occur at different spatio-temporal resolutions, thus, you must aggregate the data at different resolutions while searching for these relationships; and
2) sometimes, they only materialize for values that are higher than the norm. For example, only high precipitation values may affect the number of taxi trips. Therefore, besides checking for correlations over the whole data, you should also consider looking for correlations between extreme or unusually high (or low) values.

A useful reference on this topic is Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets, Chirigati et al., ACM SIGMOD, 2016.

A good source for data sets about NYC is https://opendata.cityofnewyork.us.

Some data sets that may be useful include:

- Traffic speed (2009 - 2016)
    - Each GPS location corresponds to the center point of a road segment
    - Link: https://figshare.com/articles/Traffic_Speed_Data_2009_-_2016_/4765759
- Vehicles collisions: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
- Census data
    - Demographics: http://www.nyc.gov/html/dcp/html/census/demo_tables_2010.shtml
    - Income information: http://www.nyc.gov/html/dcp/html/census/socio_tables.shtml
    - Shape files for census tracts: http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml (search for "tract")
- Weather information
    - http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets
    - http://www7.ncdc.noaa.gov/CDO/dataproduct
    - Select either "Surface Data, Global Summary of the Day", or "Surface Data, Hourly Global" for a more detailed analysis.
    - You can then choose NY state, and select the "John F Kennedy International Airport" station (or all the stations, Central Park, JFK and LaGuardia).
- Property and Construction data
    - ACRIS (sales data): https://data.cityofnewyork.us/City-Government/ACRIS-Real-Property-Master/bnx9-e6tj
    - Multi Agency Permits (including all applications for construction activity): https://data.cityofnewyork.us/City-Government/Multi-Agency-Permits/xfyi-uyt5

Deliverable: A revised version of the report you submitted for Part I, submitted to NYU Classes, that includes:

- Updates to the summary and data quality issues in case you had new findings about these during your analysis
- A new section entitled Part II that includes the results for the data analysis you carried out. Part II must include:
    - Description of the experimental setup (cluster configuration, number of mappers/reducers, tools you used) as well as any optimizations you applied to speed up your code.
    - List of hypotheses you set out to investigate.
    - For each hypothesis, describe the analyses you carried out to prove or disprove it. You will consider different data sets and attributes while investigating your hypotheses. An important deliverable is a list of relationships/correlated attributes and an associated score (depending on the technique they use). For example:

  - (taxi_trips, precipitation, 0.8); (taxi_income, precipitation, 0.7) The report should include a discussion on why each hypothesis is true or false.
  - The use of insightful visualizations is highly encouraged. The code for your analyses must be available in the github repo you provided. You must also include a README file with instructions on how to run your code. Any figures and results in the report must point to a script that can be used to reproduce the result.
- You should describe the individual contributions of each of the project's members.

Notes on Report Structure

The project report should have the following information:

• Title

• Authors

• Abstract: The idea of the abstract is to provide a brief summary of
the report. It should give the reader an idea about the work that was undertaken
and what you findings are.

• Introduction: This is where you need to outline the motivation for your work, state the problem you are addressing, why is important, the underlying concepts, and justify if and how big data infrastructure was needed (or useful) for your project.

• Part I: Data summary and data quality issues

• Part II: Data Exploration

  - Experimental techniques and methods: You should provide details about the methodology and tools you used. You should also describe your experimental setup, including the data you used, and the cluster configuration (e.g., node configuration, number of nodes, mappers and reducers).
  - Results and discussion: Discuss your findings as well as any issues/challenges you encountered and how you addressed them. For Part I, you will have subsections for Data Summary and Data Quality Issues. For Part II, you will have subsections for the different analyses you performed.

• Individual Contributions: describe the contributions each member of group made to the project

• Summary/conclusions

• References: List any references you have used

Suggestions

- To aggregate data over space, you can use pre-defined shapefiles, such as
  - NYC Neighborhoods shapefile in geojson: http://nycdata.pediacities.com/dataset?tags=neighborhoods
  - NYC ZIP: http://nycdata.pediacities.com/dataset/nyc-zip-code-tabulation-areas