Big Data -- Fall 2019

Final Project

Code and report due at **9am on Monday, December 9**
Project presentation **during class on Monday, December 15**

The goal of the project is to give you hands-on experience with multiple steps of the data lifecycle that benefit from big data infrastructure.

The project consists of the three tasks described below. Specific details about data sets and output formats for the different tasks will be provided in a separate document.

Grading and Submission Instructions

The project is worth 100 points, or 30% of the overall course grade. The project is to be done in teams of three. The project consists of three tasks, each worth 30 points. Each project team will present their work during the final class, with project presentation worth 10 points.
You will submit your project report and presentation slides on NYU Classes, and in your report, you should include a link to the github repository where we can find your code.

The final report must follow the ACM Proceedings Format, using the **sample-sigconf.tex** template provided at https://www.acm.org/publications/proceedings-template for LaTeX (version 2e), or https://www.acm.org/binaries/content/assets/publications/word_style/interim-template-style/interim-layout-.docx for Word.

Final project presentations will be done in a poster format: instructors and your classmates will be your audience. All teammates must be present on the day of their team's presentation. Plan for a 10-minute presentation, including time for questions.

Task 1: Generic Profiling (30 points)

Open data often comes with little or no metadata. You will profile a large collection of open data sets and derive metadata that can be used for data discovery, querying, and identification of data quality problems.

The data sets you will work with can be found on Dumbo, in the HDFS folder **/user/hm74/NYCOpenDat**

For each column in the dataset collection, you will extract the following metadata

1. Number of non-empty cells
2. Number of empty cells (i.e., cell with no data)
3. Number of distinct values
4. Top-5 most frequent value(s)
5. Data types (a column may contain values belonging to multiple types)

Identify the data types for each distinct column value as one of the following:

- INTEGER (LONG)
- REAL
- DATE/TIME
- TEXT

For each column count the total number of values as well as the distinct values for each of the above data types.

For columns that contain at least one value of type INTEGER / REAL report:

- Maximum value
- Minimum value
- Mean
- Standard Deviation

For columns that contain at least one value of type DATE report:

- Maximum value
- Minimum value

For columns that contain at least one value of type TEXT report:

- Top-5 Shortest value(s) (the values with shortest length)
- Top-5 Longest values(s) (the values with longest length)
- Average value length

Extra credit: For each table T, indicate columns that are candidates for being keys of T.

Output: For each table, you will output the corresponding metadata for each column. The precise format of the output can be found in NYU Classes, under the Project resources.

Report: In your project report, you should discuss the challenges you have faced while designing and implementing your solution and (e.g., skewed data), time to run the profiling tasks, any optimizations you have implemented to speed up your code, and data quality issues you have identified (e.g., too many missing values, heterogeneous columns -- columns with values of multiple types). You should summarize your results using plots/visualizations, e.g., a histogram that shows for each data type, how many columns contain that type. Also report the most common types that co-occur in a column -- this can be done by applying frequent itemsets: you can find the 2-, 3-, 4-frequent itemsets

Useful References

- Profiling relational data: a survey. Abedjan et al., VLDB Journal 2015 (https://link.springer.com/article/10.1007/s00778-015-0389-y)
- Mining Database Structure; Or, How to Build a Data Quality Browser. Tamraparni Dasu et al., ACM SIGMOD 2002. (http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=FEBFD4B33516758381A9FE6E2DA38BDC?doi=10.1.1.123.3496&rep=rep1&type=pdf)

Task 2: Semantic Profiling (30 points)

For this task you will extract more detailed information about the semantics of columns. We will work for a subset of the data sets used in Task 1.

For each column, identify and summarize semantic types present in the column. These can be generic types (e.g., city, state) or collection-specific types (NYU school names, NYC agency). For each semantic type T identified, enumerate all the values encountered for T in *all columns* present in the collection.
You will look for the following types and add one or more semantic type labels to the column metadata together with their frequency in the column:

- Person name (Last name, First name, Middle name, Full name)
- Business name
- Phone Number
- Address
- Street name
- City
- Neighborhood
- LAT/LON coordinates
- Zip code
- Borough
- School name (Abbreviations and full names)
- Color
- Car make
- City agency (Abbreviations and full names)
- Areas of study (e.g., Architecture, Animal Science, Communications)
- Subjects in school (e.g., MATH A, MATH B, US HISTORY)
- School Levels (K-2, ELEMENTARY, ELEMENTARY SCHOOL, MIDDLE)
- College/University names
- Websites (e.g., ASESCHOLARS.ORG)
- Building Classification (e.g., R0-CONDOMINIUM, R2-WALK-UP)
- Vehicle Type (e.g., AMBULANCE, VAN, TAXI, BUS)
- Type of location (e.g., ABANDONED BUILDING, AIRPORT TERMINAL, BANK, CHURCH, CLOTHING/BOUTIQUE)

- Parks/Playgrounds (e.g., CLOVE LAKES PARK, GREENE PLAYGROUND)

For example: for a column called Areas_study, you would include (Architecture,10), ('Animal science', 25). Also include the count for any values of type not listed above, e.g., (other, 100).

You will devise different strategies for identifying these types. For example:

- To identify columns with the same type, given a column you can search for other columns with similar content (see chapter on Similarity in our textbook) and/or with similar attribute names.
- Some types can be detected using regular expressions (e.g., phone numbers)
- External resources such as ontologies or dictionaries can also be useful to detect some types (e.g., the list of all states in the USA)

Use your creativity and also *Google* to look for additional strategies. Hint: Think about how you can combine multiple strategies to improve the effectiveness of your semantic type detection.

Output: For each column, assign one or more labels from the above list. If you encounter a type that is not in this list, suggest a label that describes the type. The precise format of the output can be found in NYU Classes, under the Project resources.

Report: Describe strategies you used to detect the different types, their benefits and limitations. Report the precision and recall for the different strategies. To compute precision and recall, you need the true type of each column: the team members will collaboratively and manually label the columns with their ***true type****. Note that a given column may have values of different types, therefore, its true type may consist of multiple labels. Include only labels for types that occur frequently, and omit any outliers. Include visualizations that summarize your findings, e.g., a histogram showing for each type, the number of columns in which the type appears; a visualization that shows the prevalence of heterogeneous columns, i.e., columns that have values belonging to multiple types.*

Useful References

Yeye He and Dong Xin. SEISA: set expansion by iterative similarity aggregation. International conference on World Wide Web (WWW '11), 2011.

Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Renée J. Miller. Table union search on open data. Proc. VLDB Endow. 11, 7 (March 2018), 813-825.

Meihui Zhang, Marios Hadjieleftheriou, Beng Chin Ooi, Cecilia M. Procopiuc, and Divesh Srivastava. Automatic discovery of attributes in relational databases. ACM SIGMOD International Conference on Management of data (SIGMOD '11), 2011.

Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çagatay Demiralp, and César Hidalgo. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), 2019.

Andrew Ilyas, Joana M. F. da Trindade, Raul Castro Fernandez, Samuel Madden. Extracting Syntactical Patterns from Databases. ICDE 2018.

Task 3: Data Analysis (30 points)

For this task, you can choose one of the following options:

1) You will use what you learned while profiling the data to identify potential quality issues, including:
    a) Null values: identifying values is challenging since they have different representations, e.g., NULL, N/A, 999-999-99999
    b) Find outliers and explain them: your goal will be to determine whether outlier values are data features or if they represent data quality problems

2) Spatial coverage. You will study the spatial coverage of the NYC open data sets. For example, are there particular regions that are data poor (i.e., that have comparatively fewer data sets or data points)? Is there missing data for certain regions? How does spatial coverage vary over time? Is there a correlation between spatial coverage and economic indicators such as income per capita and education level?

3) Identify the three most frequent 311 complaint types by borough. Are the same complaint types frequent in all five boroughs of the City? How might you explain the differences? How does the distribution of complaints change over time for certain neighborhoods and how could this be explained?

4) Compute demographic disparities in high school graduation rates by zip code, for White, Hispanic, Black, and Asian demographic groups. List five zip-codes with the highest disparity between any pair of demographic groups listed above.

5) How quickly does NYC address curb and sidewalk service requests? Are these requests addressed more promptly in some zip codes than in others?

6) You will explore problems with restaurants in NYC. You will analyze data from multiple data sets, e.g., 311 food poisoning complaints, restaurant inspections, and attempt to better understand the problem. For example, is there a correlation between FOOD POISONING complaints in 311 and restaurant inspections? Does the weather contribute to an increase or decrease in complaints? Are there areas in the City that receive more

complaints? Is there a correlation between complaints in different areas and the indicators for these areas (e.g., income, education level)?

Note: For any of these choices, you are expected to add questions of your own, formulate and test hypotheses. You should also look for references on the topic that can help guide your analysis.

Report: For your report, describe the questions you asked, the data you used and your findings (both positive and negative).

Useful references:

1. Efficient Algorithms for Mining Outliers from Large Data Sets. Ramaswamy et al., SIGMOD 2000. (ftp://ftp10.us.freebsd.org/users/azhang/disc/disc01/cd1/out/papers/sigmod/efficientalgorisrrak.pdf);
2. Anomaly Detection: A Survey. Chandola et al, CSUR 2009 (http://www.dtc.umn.edu/publications/reports/2008_16.pdf)
3. Ming Hua, Jian Pei: Cleaning disguised missing data: a heuristic approach. KDD 2007: 950-958
4. Ming Hua, Jian Pei: DiMaC: a system for cleaning disguised missing data. SIGMOD Conference 2008: 1263-1266
5. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (KDD'96)
6. https://www.kdnuggets.com/2018/12/four-techniques-outlier-detection.html
7. Doraiswamy et al., Topological Analysis to Support Event-Guided Exploration in Urban Data. IEEE TVCG, 20(12): 2634-2643, 2014