Project Phase #1

Due: 2/26/24 @ 11:59PM

Content Covered

Problem Statements, Data Acquisition, Data Processing, Data Cleaning, EDA

Project Overview

The course project forms the hands-on practical learning component of the course, and will have students putting into practice each step of the data science pipeline (depicted in Figure 1, adapted from [1]). The project will be broken into 3 phases, with Phase 1 covering the first 5 steps of the data science pipeline shown below. The project is expected to be motivated by issue(s) in an application domain of your interest, and addressing these issues using data gathered from the domain.

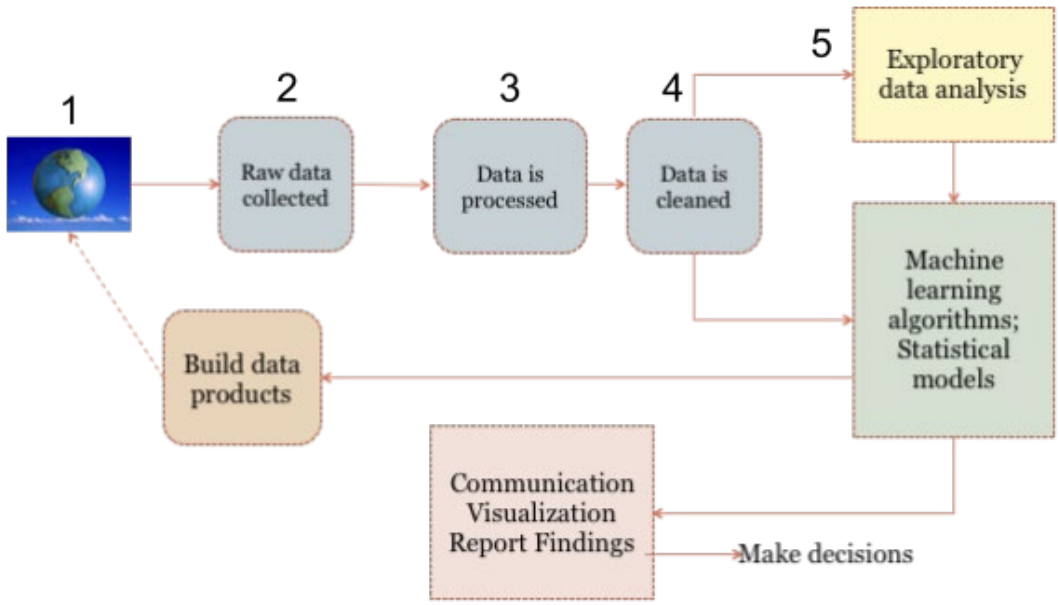

Figure 1: The Data Science Pipeline

Learning Outcomes for Phase 1:

1. Identify problems prevalent in public application domains. (Task 1 of figure 1)
2. Research and identify data sets (preferably structured data) to address the problems and collect the relevant data sets. (Task 2)
3. Clean and provision the data for downstream explorations and analytics. (Tasks 3, 4)
4. Understand the basic characteristics of the data by performing John Tukey's exploratory data analysis (EDA) [2]. (Task 5)

Description

An important and critical phase of the data-science process is data collection. Several organizations including the federal government (data.gov, pew research) have their data available to the public. Social network applications such as Twitter and Facebook collect enormous amounts of data contributed by their numerous and prolific users. For other businesses such as Amazon and NYTimes data is a significant and valuable byproduct of their main business. Nowadays everybody has data. Most of these data generator businesses make a subset of their data available for use by registered users for free. Some of them, as downloadable data files (.csv, .xlsx), others as a database (.db, .db3). Sometimes the data that needs to be collected is not in a specific format but is available as web page content. In this case, typically a web crawler is used to crawl the web (pages) and scrape the data from these web pages to extract the information needed. Data generating organizations have realized the need to share at least a subset of their data with users interested in developing applications. Entire data sets are sold as products. You'll research these organizations that deal with data and data solutions, and decide on a domain and set of data, and the goals of your analysis of this data. In particular, what are the questions you want to answer by analyzing this data. For this project you'll work with structured data sets.

General Project Requirements

1. **Work Environment:** Required language for the project is Python. You can use any Python environment of your choice: Jupyter, IPython, etc.
2. **Programming:** Prepare yourself to program by learning from the course textbooks and online resources.
3. **Academic Integrity: You will get an automatic F for the course if you violate the academic integrity policy. See the course syllabus for more detail.**
4. **Project Phases:** This project will span three separate phases, each building on the last. Each phase has its own due date, and must be completed before you can move onto the next phase. During Phase 1 you will be forming a problem statement, getting your data, and doing initial EDA. During Phase 1 you may change your problem, or the data you choose to use. Once Phase 1 is complete you will no longer be allowed to change, which makes it critical that you carefully complete Phase 1.
5. **Teams:** For the duration of the project you may work in groups of two or three. Please make sure that you have registered your team via Piazza, which must be completed before asking for project guidance. Project discussion should only occur between you

and your teammate, or you and course staff. Each team member must contribute each part of the project. There will be **one submission per team.**

6. **487 vs 587:** In certain instances, 587 students will be required to complete additional work, and in general their projects will be held to higher standards. Instances of additional work will be clearly identified in the deliverables section.

Submission Requirements

1. **Deadlines:** Your submission is due by 11:59 PM on Monday, 2/26/2024. Only one-day late submission will be considered, and there will be a 20% penalty. You must submit Phase 1 to begin work on Phase 2. Please start the project as soon as possible.
2. **Submission**: Project deliverables should be submitted via Brightspace. There should be one final submission per group. You can submit multiple times before the deadline but we will grade the final submission. For the final submission you are required to submit a **zip** file containing all the required deliverables. The zip file must be named: **`member1_member2_member3_phase_1.zip`**. It should contain a PDF for your project report named `report.pdf` and a `src/` directory with your commented code files.

Deliverables **[100 marks total]**

1. **Problem Statement:** Form a title and problem statement that clearly state the problem and questions you are trying to answer. Additionally:
    a. **[10 marks]** Discuss the background of the problem leading to your objectives. Why is it a significant problem?
    b. **[10 marks]** Explain the potential of your project to contribute to your problem domain. Discuss why this contribution is crucial?

2. **Data Sources [10 marks]:** Collect your data. Your data can come from multiple sources. For example, Medical, Bank, sports, health, Kaggle, Amazon reviews, Twitter, Youtube, Reddit, etc. This data has to be large enough for the data analysis to yield significance. At least 2000 rows/records. Although there is no requirement for the number of columns, as this will vary based on project, your data should contain enough columns to successfully complete the deliverables for Phase 1. If your data is already clean, or does not have a sufficient number of columns to complete Phase 1, you must find different/additional data sources. **You must cite and link your data sources in the report.**
3. **Data Cleaning/Processing [35 Marks]:** Your dataset has to be cleaned and properly processed. Please submit a report where you explain each processing/cleaning step properly. We expect to see comments and markup for this step. In order to get full marks you must clearly document **7 (10 for 587 students)** distinct processing/cleaning operations.
4. **Exploratory Data Analysis (EDA) [35 Marks]**: Perform exploratory data analysis as defined in the NIST publication [2] and as originally described by John Tukey [3]. Record

the outcomes and what you learned and how you will use this information. For example, in choosing features (columns) and dropping columns, and in short feature engineering. You need to demonstrate **7 (10 for 587 students)** different, significant and relevant EDA operations and describe how you used these to process the data sets further to provision them for downstream modeling and analytics. Figures and tables should be included where relevant.

Additional Information and References

Here are some places you can explore datasets for your problem statement:

- Pew Research Center: https://www.pewresearch.org/download-datasets/
- Kaggle: https://www.kaggle.com/datasets
- Google Dataset Search: https://datasetsearch.research.google.com/
- UCI ML Repository Datasets: https://archive.ics.uci.edu/ml/datasets.php

Here are some example cleaning/EDA steps which you can keep for reference.

Univariate or Multivariate Non-Graphical Techniques

- Make sure to remove duplicates
- Set proper precision
- Show the general characteristics of the data (center, spread, modality, shape, and outliers)
- You can calculate measures of spread including variance, standard deviation, and interquartile range
- If you are working with text data: Normalize cases (depends on your requirement), Remove punctuations, Remove emoticons and unwanted text (NOTE: Text data needs a lot more cleaning, refer to https://www.analyticsvidhya.com/blog/2022/01/text-cleaning-methods-in-nlp/)

Univariate or Multivariate Graphical

- Use various plots (Popular graph types include line graphs, bar graphs, pie charts, scatter plots, box plots, and histograms. To learn more: https://matplotlib.org/stable/gallery/index.html, https://seaborn.pydata.org/)

References

[1] C. O'Neill and R. Schutt. Doing Data Science., O'Reilly. 2013.

[2] NIST on EDA, https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm, last viewed February, 2021.

[3] John Tukey Biography, https://mathshistory.st-andrews.ac.uk/Biographies/Tukey/, last viewed 2021.