Stat 474: Project Guidelines, Fall 2018

Overview:

The course project is your opportunity to explore a topic of your academic or personal interest. You will be randomly selected in a group of 2-3. Your project should focus on data analysis of "Big Data" in its broad meanings (i.e., its 3 V's - Volume, Variety, Velocity). Your project is supposed to be interesting for you: choose a topic you like or either of the two projects described below. Your final report on your project should be submitted by **5:00 P.M. on Wednesday, December 12**.

Requirements and Deadlines:

- **Presentation:** Each group will prepare and present their project on **Friday, December 7 at 11:00 A.M.**
- **Final report:** Submit an electronic version of your paper in PDF or MS. Word format and your R code to *Blackboard* by **Wednesday, December 12, 5:00 PM**.

General Advice:

- Your should start your project immediately. Preprocessing and cleaning up the data set usually take 80% of the total time spent on the project.
- Your report should be written in R Markdown and then knitted into PDF format. You can use `output: pdf_document` in your markdown YAML to configurate the knit output format,
- The text of your paper (not including tables, graphics, appendices (R code, output, or mathematical derivations), etc. . .) should be about 4-6 pages in length if you choose single-spaced (there is no required spacing option. . . choose your favorite). The length may vary widely from project to project.
- Thoughtful graphics are often more illuminating than tables or pages of R output.
- Your grade does not depend on whether your original hypotheses are correct - you will probably learn more if your hypotheses are incorrect!
- Your research question, motivation, hypotheses, methods, assumptions, results, limitations, and conclusions and further discussions should be presented clearly in your paper, if applicable (you may not need a separate section for each of these topics, and not all of these topics are relevant for all projects). Your paper should also include a short discussion of any challenges you faced. Be sure to cite sources, if applicable, and include a reference list.

Details on the project write-up:

The standard project is to analyze a "big" data set that is of interest to you. Your report should contain the following:

1. *Abstract:* A one paragraph summary of what you set out to learn, and what you ended up finding. It should summarize the entire report.

2. *Introduction:* A discussion of what questions you are interested in.
3. *Data Set:* Describe details about how the data set was collected and the variables in the data set.
4. *Analysis:* Describe how you analyze the data set. Specifically, you should discuss how you carried out the steps in analysis discussed in class, i.e., data exploration using graphs, modeling or data mining

techniques used to answer your questions of interest.

5. *Results:* Explain in details how the graphs and/or model results provide the answers to your questions of interest.
6. *Limitations of study and conclusion:* Describe any limitations of your study and how they might be overcome in future research and provide brief conclusions about the results of your study.

Details on the project topic:

1. Choose your own project:

You are asked to analyse a "big" data set in its broad sense. That is,

- The data set can be very large in volume (should be roughly 1 GB or more)
- The data set is mainly unstructured, e.g., text data
- The data set is obtained by incorporating data from multiple sources (must be 3 or more sources)
- The online streaming data set (e.g., Twitters data)

Before searching for the interesting data set, you should first come up the your research questions. Your questions should be relatively narrow in scope, so that you will be able to complete the project within 4 weeks.

2. Airline on-time performance

The goal of this project is to use graphs to analysis a very large data set on flight arrival and departure details for all commercial flights within the USA from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up to 1.6 GB of space compressed and 12 GB when uncompressed. You can download the data set from http://stat-computing.org/dataexpo/2009/the-data.html.

You are free to come up with your questions of interest, but here are a few ideas to get you started:

- When is the best time of day/day of week/ time of year to fly to minimise delays?
- Do older planes suffer more delays?
- How well does weather predict plane delays?
- Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?

Note: You must come up with one question that is not one of the above suggested questions above.

Some supplemental data on airport, carrier codes, plans and weather can be reached from links on http://stat-computing.org/dataexpo/2009/supplemental-data.html

3. Hillary Clinton's Emails

Throughout 2015, Hillary Clinton has been embroiled in controversy over the use of personal email accounts on non-government servers during her time as the United States Secretary of State. Some political experts and opponents maintain that Clinton's use of personal email accounts to conduct Secretary of State affairs is in violation of protocols and federal laws that ensure appropriate recordkeeping of government activity.

There have been a number of Freedom of Information lawsuits filed over the State Department's failure to fully release the emails sent and received on Clinton's private accounts. On Monday, August 31, the State Department released nearly 7,000 pages of Clinton's heavily redacted emails (its biggest release of emails to date).

The documents were released by the State Department as PDFs. Kaggle community have cleaned and normalized the release documents. You can download the SQLite database file from the Blackboard. It contains 4 relational tables: `Aliases`, `EmailReceivers`, `Emails`, and `Persons`. The raw email text is stored in `Emails` with the following attributes:

- Id - unique identifier for internal reference
- DocNumber - FOIA document number
- MetadataSubject - Email SUBJECT field (from the FOIA metadata)
- MetadataTo - Email TO field (from the FOIA metadata)
- MetadataFrom - Email FROM field (from the FOIA metadata)
- SenderPersonId - PersonId of the email sender (linking to Persons table)
- MetadataDateSent - Date the email was sent (from the FOIA metadata)
- MetadataDateReleased - Date the email was released (from the FOIA metadata)
- MetadataPdfLink - Link to the original PDF document (from the FOIA metadata)
- MetadataCaseNumber - Case number (from the FOIA metadata)
- MetadataDocumentClass - Document class (from the FOIA metadata)
- ExtractedSubject - Email SUBJECT field (extracted from the PDF)
- ExtractedTo - Email TO field (extracted from the PDF)
- ExtractedFrom - Email FROM field (extracted from the PDF)
- ExtractedCc - Email CC field (extracted from the PDF)
- ExtractedDateSent - Date the email was sent (extracted from the PDF)
- ExtractedCaseNumber - Case number (extracted from the PDF)
- ExtractedDocNumber - Doc number (extracted from the PDF)
- ExtractedDateReleased - Date the email was released (extracted from the PDF)
- ExtractedReleaseInPartOrFull - Whether the email was partially censored (extracted from the PDF)
- ExtractedBodyText - Attempt to only pull out the text in the body that the email sender wrote (extracted from the PDF)
- RawText - Raw email text (extracted from the PDF)

You are asked to perform text analysis to uncover the political landscape in Hillary Clinton's emails, e.g., sentiment analysis on the email content, sentiments associated wirh foreign coutries, clustering and topic modeling on all emails.