Welcome to the 2017 Columbia Data Science Hackathon!

We're looking forward to spending the next 20+ hours with you, working on real-world data science challenges, making friends, and having fun!

Google Cloud Platform Setup

To help make the hackathon go smoothly, we are proud to have Google Cloud Platform as a sponsor. Using their technology, you'll be able to do all of your data science in your browser with a high powered computer supporting all of your computations! There will be a talk on how to use GCP at 7:15 PM with more details on how to set this up.

Economic + Financial + Energy Data (Bloomberg)

Dive into the timeseries data that Bloomberg customers can't live without!

- Think you can beat the market? Backtest your model on 10 years of stock prices from thousands of tickers.
- Think green tech is poised to challenge the convent slaional energy industry? Find the relationship between Bloomberg New Energy Finance data (released for the first time at this hackathon!) and our other datasets that back up your claim.

Above all, have fun, learn something new, and tell a story that interests **you**!

Enron Data (Digital Reasoning)

Introduction

The Enron email corpus contains five hundred and seventeen thousand emails from the corporate email servers of the Enron Corporation. The corpus was released to the public domain by the Federal Energy Regulatory Commission post the bankruptcy of Enron on December 2, 2001. Enron employed over twenty thousand employees and was a major force in the American energy sector. Despite being named "America's Most Innovative Company" by Fortune magazine for years, it was discovered that senior Enron employees had engaged in systematic accounting fraud to inflate corporate assets and defraud investors. Its subsequent failure remains one of the most infamous events of the twenty-first century.

Files

<table>
<tr><td>enron_emails_csv.zip</td><td>517,402 emails from Enron corpus (May 7, 2015 version)</td></tr>
<tr><td>enron_sentences_json_zip</td><td>3,153,441 sentences extracted from the email body text</td></tr>
<tr><td>schema:</td><td><table>
<tr><td>id</td><td>sentence id</td></tr>
<tr><td>messsageId</td><td>message hash</td></tr>
<tr><td>vertexId</td><td>id of message node (unused)</td></tr>
<tr><td>originalText</td><td>text extracted from email body</td></tr>
<tr><td>preprocessedText</td><td>normalized text</td></tr>
<tr><td>previous</td><td>previous sentence</td></tr>
<tr><td>next</td><td>next sentence</td></tr>
<tr><td>originalHash</td><td>hash of original text</td></tr>
<tr><td>preprocessedHash</td><td>hash of preprocessed text</td></tr>
</table></td></tr>
<tr><td>enron_employee_list_partial.csv</td><td>partial list of Enron employee emails & departments</td></tr>
</table>

Data Science Questions

1. What patterns of unethical behavior can be identified from the emails?
2. Who were the most unethical individuals from the emails?
3. How can we build profiles of individuals from the email corpus?
4. What range of topics were discussed in the email corpus?
5. Can we analyze and map the sender/receiver or mention network of communications?
6. Can we predict the roles of individuals from the partial list of employees?
7. How can we discern if emails pertain to business or personal matters?
8. Were the Enron corporate spam filters successful?
9. Can we identify suspicious emails that potentially pose as an insider threat?
10. Can we find instances of rumor language?
11. Can we find instances of secrecy language?
12. How can we use the Enron sentences corpus to find patterns of bad behavior?
13. What is the type and frequency of gifts that enron employees gave or received?
14. Were Enron employees in the corpus mostly male or female? married or single? over 30?
15. Can we use sources like DBpedia or Freebase to enrich profiles of Enron employees?

Other References

- Carnegie Mellon CALO Project: https://www.cs.cmu.edu/~./enron/
- Kaggle Enron Dataset: https://www.kaggle.com/wcukierski/enron-email-dataset
- Enron Key Figures: http://www.corpwatch.org/article.php?id=457
- Enron Chronology of Events: http://www.corpwatch.org/article.php?id=2278

Government Spending Contracts (Enigma)

What's Enigma?

Enigma is a data operations and intelligence startup based in New York City. We build data infrastructure, industry-specific solutions and provide data for commercial workflows. We have also built a free platform to connect a wider world of people to public data, Enigma Public.

What's the dataset?

In 2016 alone the U.S. government spent over $3 trillion dollars. Government procurement is a sizeable part of the U.S. economy — as procurement overall is a sizeable part of the global economy. This Government Spending Contracts dataset includes over 15 years of contracts, from 2000-2017, with the federal government. The data is quite granular, including not only the vendor who won the contract, the value of the contract, the agency within the federal government relevant to the contract and 200+ other columns of details.

Digging into government procurement data, i.e. where the government is spending money, reveals a lot about government priorities, as well as the types of businesses that are receiving contracts. For the past few decades the federal government has instituted supplier diversity programs whereby they have attempted to diversify the types of businesses winning government contracts. To what extent have they been successful?

A recent project of note on a similar topic was Steve Ballmer's USAFacts site, though that project aimed at covering the topical impact of spending, rather than examining the companies with which the money was being spent.

Some things you might ask:

1. Where are supplier companies across the U.S. located?
    - How does that vary by industry and how has that changed over time?
2. What factors make a company more likely to win a contract with a given agency?

Explore the data in Enigma's online portal here

Data use policies

- If you're using the Bloomberg dataset, please agree to their data policies
- If you're using the Digital Reasoning and Enigma datasets, you're allowed to share the data and results as long as you credit the data sponsor

---

Project evaluation

Everyone will participate in an initial round of short presentations, then teams that are selected as finalists will give longer presentations of their work. The hackathon will be open-ended — we will provide suggested data problems to work on, but this is no Kaggle competition — so you'll get to work on whatever you find most interesting. Project evaluation is not based on achieving 95% model accuracy but rather how you choose to credit an impactful model and analytical framework to answer a question about data.

Successful projects usually have a mix of modeling and data visualization, and good presentations always clearly communicate the analysis, results, and impact. The judges want to see that you're thoughtful about the problem at hand, effective in executing your analysis, and compelling when talking about why your project matters.

Event timeline

Monday 9.18.2017	TBD	Pre-hackathon workshops
Thursday 9.21.2017		
Friday 9.22.2017	5pm – 6pm	Student check-in
	6pm – 8pm	Ramp-up workshops and tech talks by sponsors
	8pm	Dinner is served
	9pm	Hacking starts
Saturday 9.23.2017	8am	Breakfast is served
	11am	Hacking ends
	11am – 12pm	Teams present to judges, science fair style
	12pm – 12:30pm	Finalists are chosen and lunch is served
	12:30pm – 1:30pm	Finalists give presentations
	2:00pm	Winners are announced

Housekeeping / Reminders

- All hackathon updates and announcements will be made on-site as well as in the Facebook group: https://www.facebook.com/groups/2017ColumbiaDSHackParticipants
- Please wear your black hackathon lanyard if you leave the auditorium, so that security knows you are part of the hackathon
- Lerner Hall is officially closed between 3 - 8 am, so if you leave the auditorium during this time, you will not be able to re-enter until 8 am on Saturday
- If you have any questions, please feel free to ask one of our volunteers (in green t-shirts) or post in the Facebook group

---

Sponsors

Google Cloud Platform

Bloomberg

NBCUniversal

Digital Reasoning

Honeywell

TWO SIGMA

facebook

enigma

Organized By

Columbia Data Science Society