94-775
Unstructured Data Analytics for Policy
Lecture 1: Course Overview, Basic Text Analysis

George Chen
Big Data

We’re now collecting data on virtually every human endeavor

How do we turn these data into actionable insights?
Two Types of Data
Structured Data

Well-defined elements, relationships between elements

Can be labor-intensive to collect/curate structured data

Image source: http://revision-zero.org/images/logical_data_independence/hospital_appointments.gif
Unstructured Data

No pre-defined model—elements and relationships ambiguous

Often: Want to use heterogeneous data to make decisions

Of course, there is structure in this data but the structure is not neatly spelled out for us

We have to extract what elements matter and figure out how they are related!
Example 1: Health Care

Forecast whether a patient is at risk for getting a disease?

Data

- Chart measurements (e.g., weight, blood pressure)
- Lab measurements (e.g., draw blood and send to lab)
- Doctor’s notes
- Patient’s medical history
- Family history
- Medical images
Example 2: Electrification

Where should we install cost-effective solar panels in developing countries?

Data

- Power distribution data for existing grid infrastructure
- Survey of electricity needs for different populations
- Labor costs
- Raw materials costs (e.g., solar panels, batteries, inverters)
- Satellite images
Example 3: Online Education

What parts of an online course are most confusing and need refinement?

Data

• Clickstream info through course website
• Video statistics
• Course forum posts
• Assignment submissions
Unstructured Data Analysis

Not detailed in lecture but addressed by final project

Question
The dead body
This is provided by a practitioner

Data
The evidence
Some times you have to collect more evidence!

Finding Structure
Puzzle solving, careful analysis
Exploratory data analysis

Insights
Answer original question

There isn’t always a follow-up prediction problem to solve!

UDA involves lots of data ➔ write computer programs to assist analysis
Prereq: Python programming

This course is much harder if you don’t already know Python

Part I: Exploratory data analysis

Part II: Predictive data analysis
Part I: Exploratory data analysis

Identify structure present in “unstructured” data

• Frequency and co-occurrence analysis
• Visualizing high-dimensional data/dimensionality reduction
• Clustering
• Topic modeling (a special kind of clustering)

Part II: Predictive data analysis

Make predictions using structure found in Part I

• Classical classification methods
• Neural nets and deep learning for analyzing images and text
Course Goals

By the end of this course, you should have:

• Hands-on programming experience with exploratory and predictive data analysis

• A high-level understanding of what methods are out there and which methods are appropriate for different problems

• A very high-level understanding of how these methods work

• The ability to apply and interpret the methods taught to solve a policy question

I want you to leave the course with practically useful skills solving real-world problems with unstructured data analytics!
Deliverables & Grading

Letter grades are assigned based on a curve.

Percentage Contribution of Different Assignments to Overall Grade

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW1</td>
<td>8%</td>
</tr>
<tr>
<td>HW2</td>
<td>8%</td>
</tr>
<tr>
<td>HW3</td>
<td>4%</td>
</tr>
<tr>
<td>Quiz</td>
<td>35%</td>
</tr>
<tr>
<td>Final Project Proposal</td>
<td>10%</td>
</tr>
<tr>
<td>Final Project</td>
<td>35%</td>
</tr>
</tbody>
</table>

Individual effort

Group effort
Individual Effort Assignments

- If you are having trouble, ask for help!
  - We will answer questions on Piazza and will also expect students to help answer questions!
- Do not post your candidate solutions on Piazza
- In the real world, you will unlikely be working alone
  - We encourage you to discuss concepts/how to approach problems
  - Please acknowledge classmates you talked to or resources you consulted (e.g., stackoverflow)
- For individual effort assignments, do not share your code with classmates (instant message, email, Box, Dropbox, AWS, etc)

Penalties for cheating are severe: 0 on assignment, F in course (=(
Mid-mini Quiz

Format:

- You bring a laptop computer and produce a Jupyter notebook that answers a series of questions (~50% conceptual, ~50% coding)
- Open book, open note, open internet
- No collaboration (obviously)
- You are responsible for making sure your laptop has a compute environment set up appropriately and has enough battery life (or you sit close to a power outlet)
- Late exams will not be accepted
- Thursday 2/7 at usual lecture time/location
Final Project

- Must be done in a group of ~4 students
  - You can choose your own groups
- Final project proposals (2 pages) are due Friday 2/8, 11:59pm & must specify who the group members are
- Required components will be stated in the next slide
- Final project reports are due Thursday 3/7, 11:59pm & consist of:
  - Jupyter notebook (edited down to be clean, concise)
  - Slide deck for your final project presentation
- Last week (3/5 and 3/7): final project presentations!
Final Project Rubric

• **Policy question**: what public policy question are you addressing? Please be clear and concise.

• **Data analysis**: clearly state what part of your data are unstructured (some but not all of the data you are analyzing must be unstructured), and carefully justify every step of your analysis with supporting visualizations/intermediate outputs as needed

• **Code**: your code should actually run!

• **Conclusions**: come up with insights that are based on your quantitative data analysis and that address your original policy question

• **Presentation**: how polished is your final project presentation?
Final Project Rubric II

Contribution of Different Components

Policy Question: 15%
Analysis Justification: 30%
Code Actually Runs: 30%
Conclusions: 15%
Presentation Quality: 10%
Final Project Proposal

- **Policy question:** what public policy question are you addressing? Please be clear and concise.

- **Data:** what data have you found that you want to analyze, and why is at least some portion of it unstructured?

- **Proposed analysis:** what specific methods do you want to try and why? In what way would these address your proposed policy question? Are there specific obstacles you think you will have to address? What would a “successful” analysis look like?
Yes, we will post a few of the best final projects from last year
Main source of material: lectures slides

We'll post complimentary reading as we progress

Check course website
http://www.andrew.cmu.edu/user/georgech/94-775/

Assignments will be posted and submitted on canvas

Please post questions to piazza (link is within canvas)
• The data science/machine learning tools available have changed *drastically* over the last few years

• Working with most of the latest innovations from computer scientists requires some programming (at this point, Python is standard for machine learning research)

• Also good to solidify your programming background by learning more languages

• We will be using **Anaconda (Python 3.6 version)**
  https://www.anaconda.com/
  As of now, do not use the Python 3.7 version!
Late Homework

• You are allotted 2 late days

• If you use up a late day on an assignment, you can submit up to 24 hours late with no penalty

• If you use up both late days on the same assignment, you can submit up to 48 hours late with no penalty

• Late days are not fractional

• This policy is in place precisely to account for various emergencies (health issues, etc) and you will not be given additional late days
Cell Phones and Laptops

Just like what you’d expect in a movie theater

We don’t want your device screens/sounds distracting classmates
Course Staff

Emaad Manzoor  
Teaching Assistant

George Chen  
Instructor

Office hours:  
Check course website  
http://www.andrew.cmu.edu/user/georgech/94-775/
Part 1. Exploratory Data Analysis

Play with data and make lots of visualizations to probe what structure is present in the data!
Basic text analysis: how do we represent text documents?
The **opioid epidemic** or **opioid crisis** is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s. Opioids are a diverse class of very strong painkillers, including oxycodone (commonly sold under the trade names OxyContin and Percocet), hydrocodone (Vicodin), and fentanyl, which are synthesized to resemble opiates such as opium-derived morphine and heroin. The potency and availability of these substances, despite their high risk of addiction and overdose, have made them popular both as formal medical treatments and as recreational drugs. Due to their sedative effects on the part of the brain which regulates breathing, opioids in high doses present the potential for respiratory depression, and may cause respiratory failure and death.[2]
The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.
The epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

**Term frequencies**
- The: 1
- opioid: 3
- epidemic: 1
- or: 1
- crisis: 1
- is: 1
- the: 4
- rapid: 1
- increase: 1
- in: 3
- use: 1
- of: 1
- prescription: 1
- and: 2
- non-prescription: 1
- drugs: 1
- United: 1
- States: 1
- Canada: 1
- 2010s.: 1

**Histogram**

- Frequency of “opioid”:
  - 4/28 = 1/7
  - 3/28
  - 2/28 = 1/14
  - 1/28

**Fraction of words in the sentence that are “opioid”**
increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the

<table>
<thead>
<tr>
<th>Term frequencies</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>The:</td>
<td>1/28</td>
</tr>
<tr>
<td>opioid:</td>
<td>3/28</td>
</tr>
<tr>
<td>epidemic:</td>
<td>1/28</td>
</tr>
<tr>
<td>or:</td>
<td>1/28</td>
</tr>
<tr>
<td>crisis:</td>
<td>1/28</td>
</tr>
<tr>
<td>is:</td>
<td>1/28</td>
</tr>
<tr>
<td>the:</td>
<td>4/28</td>
</tr>
<tr>
<td>rapid:</td>
<td>1/28</td>
</tr>
<tr>
<td>increase:</td>
<td>1/28</td>
</tr>
<tr>
<td>in:</td>
<td>3/28</td>
</tr>
<tr>
<td>use:</td>
<td>1/28</td>
</tr>
<tr>
<td>of:</td>
<td>1/28</td>
</tr>
<tr>
<td>prescription:</td>
<td>1/28</td>
</tr>
<tr>
<td>and:</td>
<td>2/28</td>
</tr>
<tr>
<td>non-prescription:</td>
<td>1/28</td>
</tr>
<tr>
<td>drugs:</td>
<td>1/28</td>
</tr>
<tr>
<td>United:</td>
<td>1/28</td>
</tr>
<tr>
<td>States:</td>
<td>1/28</td>
</tr>
<tr>
<td>Canada:</td>
<td>1/28</td>
</tr>
<tr>
<td>2010s.:</td>
<td>1/28</td>
</tr>
</tbody>
</table>

**Total number of words in sentence:** 28

**Histogram**

- **Frequency**:
  - the: 4/28 = 1/7
  - in: 3/28
  - opioid: 2/28 = 1/14
- **Fraction of words in the sentence that are “opioid”**: 4/28 = 1/7
increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the
Bag of Words Model

Ordering of words doesn’t matter

What is the probability of drawing the word “opioid” from the bag?

increase, drugs, opioid, in, the, States, rapid, crisis, in, Canada, opioid, or, non-prescription, use, 2010s, use, in, United, epidemic.
Handling Many Documents

• We can of course apply this technique of word frequencies to an entire document and not just a single sentence

➔ For a collection of documents (e.g., all of Wall Street Journal between late 1980's and early 1990's, all of Wikipedia up until early 2015, etc), we call the resulting term frequency the **collection term frequency** (ctf)

What does the ctf of "opioid" for all of Wikipedia refer to?

Many natural language processing (NLP) systems are trained on very large collections of text (also called **corpora**) such as the Wikipedia corpus and the Common Crawl corpus
So far did we use anything special about text?
Basic Probability in Disguise

"Sentence": ☀️🌦️🌧️🌧️🌧️气象符号

Nonnegative heights that add to 1

This is an example of a probability distribution

Probability distributions will appear throughout the course and are a key component to the success of many modern AI methods.

Term

Frequency

0 0.1 0.2 0.3 0.4

0 0.1 0.2 0.3 0.4

Term

0 0.1 0.2 0.3 0.4

Term

0 0.1 0.2 0.3 0.4

Term