94-775 Unstructured Data Analytics for Policy

Lecture 1: Course overview, analyzing text using frequencies

George Chen
What is this course about?
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

**Patients Table**

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>First Name</th>
<th>Last Name</th>
<th>Middle Initial</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Doctors Table**

<table>
<thead>
<tr>
<th>Doctor ID</th>
<th>First Name</th>
<th>Last Name</th>
<th>Middle Initial</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Appointments Table**

<table>
<thead>
<tr>
<th>Appointment ID</th>
<th>Patient ID</th>
<th>Doctor ID</th>
<th>Start time</th>
<th>End time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Can be labor-intensive to collect/curate structured data
Unstructured Data

No pre-defined model—elements and relationships ambiguous

Common examples:

- Text
- Images
- Videos
- Audio

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!_

Just because something can be stored as any of these doesn’t mean that it must be unstructured!
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
  - Sometimes 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

Part I: Exploratory data analysis

Part II: Predictive data analysis

We’re now also requiring 95-791 Data Mining
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

Part II: Predictive data analysis

*Make predictions using known structure in data*

- 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!
No existing textbook matches the course… =(

Main source of material: lectures slides

We'll post complimentary reading as we progress

Check course webpage
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)
Letter grades are assigned based on a curve.

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
  • For code: post smallest snippet, how you know it’s wrong (error message/etc), & what you’ve tried to resolve the issue
• In the real world, you will unlikely be working alone
  • We encourage discussing concepts
  • 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 produce a Jupyter notebook** that answers a series of questions (a mix of conceptual & 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
• Late exams will *not* be accepted
• **Thursday 4/22** during usual lecture time
Final Project

- Must be done in a group of ~4-5 students
- You can choose your own groups
- Final project proposals (2 pages) are due Tuesday 4/20, 11:59pm & must specify who the group members are
- Required components will be stated in the next slide
- Final project reports are due Friday 5/7, 11:59pm & consist of:
  - Jupyter notebook (edited down to be clean, concise)
  - Slide deck for your final project presentation
- Last week (May 6 & 7): final project presentations!
Final Project Rubric I

- **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

- Analysis Justification: 30%
- Code Actually Runs: 30%
- Conclusions: 15%
- Presentation Quality: 10%
- Policy Question: 15%
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?
Some final projects from the past years will be posted on 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
  https://www.anaconda.com/
Late Homework Policy

- 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
- There is no need to tell us if you’re using a late day or not (we’ll figure it out from submission timestamps)
Course Staff

Teaching Assistants
- Jingbo Jiang
- Xuejian Wang

Instructor
- George Chen

Office hours:
Check course webpage
http://www.andrew.cmu.edu/user/georgech/94-775/