95-865 Unstructured Data Analytics

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

<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>

<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>

<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

Question: The dead body
This is provided by a practitioner

Data: Some times you have to collect more evidence!

Finding Structure: Puzzle solving, careful analysis
Exploratory data analysis

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
95-865

Prereq: Python programming

Part I: Exploratory data analysis

Part II: Predictive data analysis
95-865

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:

• Lots of 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 and what their limitations are

• The ability to apply and interpret the methods taught to solve problems faced by organizations

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/95-865/

Assignments will be posted and submitted on canvas

Please post questions to piazza (link is within canvas)
Deliverables & Grading

All assignments involve coding in Python (popular amongst machine learning/computer science community)

HW3 uses Google Colab for cloud computing (many real datasets too large to either fit or process on personal machine)

Letter grades are assigned based on a curve

All assignments involve coding in Python (popular amongst machine learning/computer science community)

HW3 uses Google Colab for cloud computing (many real datasets too large to either fit or process on personal machine)
Collaboration & Academic Integrity

• 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)
• 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 =(
Pittsburgh/Adelaide Weirdness

• Piazza is shared across Pittsburgh and Adelaide sections

• Homework due dates are *almost* identical across sections

• CMU Pittsburgh campus has break days on April 5, 15, 16
  • Adelaide Wed/Fri sections unaffected by these dates
  • Pittsburgh sections will have a lecture on April 2 during the recitation slot to keep up with the pace

• CMU Adelaide finishes 1 week earlier and has their final exam period during CMU Pittsburgh’s last week of class…
The Two Quizzes

Format:

• You produce a Jupyter notebook that answers a series of questions
• Each quiz is 80 minutes
• Open notes, open internet, closed to collaboration
• You are responsible for making sure your laptop has a compute environment set up appropriately and has working internet
• Late exams will not be accepted
• Quiz 1: Friday April 16 (recitation slot)
• Quiz 2: TBA
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)
Zoom Etiquette

• I will **not** require that you have your video on if you don’t want to have it on

• However, please have your Zoom account show a photo of you

  • There’s only so much I can do to make this more like a real classroom experience

• If you have questions, or want to answer a question I asked, you can either raise your hand in Zoom to ask or type your question in Zoom chat
Course Staff

Teaching Assistants

Jingbo Jiang
Erick Rodriguez
Xuejian Wang

Instructor

George Chen

Office hours:
Check course webpage
http://www.andrew.cmu.edu/user/georgech/95-865/