95-865
Unstructured Data Analytics
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

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!
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

The evidence
Some times you have to collect more evidence!

Finding Structure

Puzzle solving, careful analysis
Exploratory data analysis

Insights

When? Where?
Why? How?
Perpetrator catchable?

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

Students who ignore this prereq do poorly in the course

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:

• 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!
Deliverables & Grading

Assignments involve coding in Python (we use popular packages such as scikit-learn and keras)

Some problems require cloud computing (we use Amazon Web Services)

Letter grades are assigned based on a curve

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework</td>
<td>20%</td>
</tr>
<tr>
<td>Mid-mini Quiz</td>
<td>35%</td>
</tr>
<tr>
<td>Final Exam</td>
<td>45%</td>
</tr>
</tbody>
</table>
1 Grading Exception

If you do better on the final exam than the mid-mini quiz:

Contribution of Different Assignments to Overall Grade

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>Homework</td>
</tr>
<tr>
<td>80%</td>
<td>Final Exam</td>
</tr>
</tbody>
</table>
Homework vs Exam Grading

- We will only grade part of your homework for accuracy, and the rest on effort

- Warning: you must have run your code already in Jupyter notebooks (more about this later)

- Exams are graded entirely on accuracy
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**

- 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)

- **Do not share your code with classmates** (instant message, email, Box, Dropbox, AWS, etc)

**Penalties for cheating are severe**

* e.g., 0 on assignment, F in course  =(*
Programming and Cloud Computing

- The data science/machine learning tools available have changed drastically over the last few years
- Working with most of the latest innovations requires some programming (Python is common)
- Datasets encountered by many organizations are now often massive
- Datasets often either won’t fit or won’t be processed fast enough on your personal machine but renting compute resources is now cheap (e.g., Amazon Web Services, Google Compute)
Course Textbook

No existing textbook matches the course… =(  

Main source of material: lectures slides
We'll post complimentary reading as we progress

Check course website
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)
Computing Environment

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

• We will give instructions for any third party packages to install and how to set up Amazon Web Services for cloud compute

• You will be submitting assignments in the form of Jupyter notebooks
Mid-mini Quiz and Final Exam

Format:

• You have to bring a laptop computer and produce a Jupyter notebook that answers a series of questions

• 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

• Quiz: Friday Feb 15 at usual recitation time/location

• Final: Friday Mar 1 at usual recitation time/location
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
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

Teaching Assistants

- Emaad Manzoor
- Yucheng Huang

Instructor

- George Chen

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

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

Total number of words in sentence: 28

Histogram

- Frequency
  - 4/28 = 1/7
  - 3/28
  - 2/28 = 1/14
  - 1/28

Fraction of words in the sentence that are “opioid”
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

opioid: 3
epidemic: 1
or: 1
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

Frequency

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

Ordering of words doesn’t matter.
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?
This is an example of a probability distribution.

Nonnegative heights that add to 1.

Probability distributions will appear throughout the course and are a key component to the success of many modern AI methods.
Now let's take advantage of properties of text

In other words: natural language humans use has a lot of structure that we can exploit
Some Words Don't Help?

How helpful are these words to understanding semantics?

Bag-of-words models: many frequently occurring words unhelpful

We can remove these words first (remove them from the "bag")

→ words that are removed are called stopwords

(determined by removing most frequent words or using curated stopword lists)
Example Stopword List (from spaCy)

'a', 'about', 'above', 'across', 'after', 'afterwards', 'again', 'against', 'all', 'almost', 'alone', 'along', 
'already', 'also', 'although', 'always', 'am', 'among', 'amongst', 'amount', 'an', 'and', 'another', 'any', 
'anyhow', 'anyone', 'anything', 'anyway', 'anywhere', 'are', 'around', 'as', 'at', 'back', 'be', 'became', 
'because', 'become', 'becomes', 'becoming', 'been', 'before', 'beforehand', 'behind', 'being', 'below', 
'beside', 'besides', 'between', 'beyond', 'both', 'bottom', 'but', 'by', 'ca', 'call', 'can', 'cannot', 'could', 
'did', 'do', 'does', 'doing', 'done', 'down', 'due', 'during', 'each', 'eight', 'either', 'eleven', 'else', 
'elsewhere', 'empty', 'enough', 'etc', 'even', 'ever', 'every', 'everyone', 'everything', 'everywhere', 
'except', 'few', 'fifteen', 'fifty', 'first', 'five', 'for', 'former', 'formerly', 'forty', 'four', 'from', 'front', 'full', 
'further', 'get', 'give', 'go', 'had', 'has', 'have', 'he', 'hence', 'her', 'here', 'hereafter', 'hereby', 'herein', 
'hereupon', 'hers', 'herself', 'him', 'himself', 'his', 'how', 'however', 'hundred', 'i', 'if', 'in', 'inc', 'indeed', 
'into', 'is', 'it', 'its', 'itself', 'just', 'keep', 'last', 'latter', 'latterly', 'least', 'less', 'made', 'make', 'many', 
'may', 'me', 'meanwhile', 'might', 'mine', 'more', 'moreover', 'most', 'mostly', 'move', 'much', 'must', 
'my', 'myself', 'name', 'namely', 'neither', 'never', 'nevertheless', 'next', 'nine', 'no', 'nobody', 'none', 
'nore', 'nor', 'not', 'nothing', 'now', 'nowhere', 'of', 'off', 'often', 'on', 'once', 'one', 'only', 'onto', 'or', 
'other', 'others', 'otherwise', 'our', 'ours', 'ourselves', 'out', 'over', 'own', 'part', 'per', 'perhaps', 'please', 
'put', 'quite', 'rather', 're', 'really', 'regarding', 'same', 'say', 'see', 'seem', 'seemed', 'seeming', 'seems', 
'serious', 'several', 'she', 'should', 'show', 'side', 'since', 'six', 'sixty', 'so', 'some', 'somehow', 
'someone', 'something', 'sometimes', 'sometimes', 'somewhere', 'still', 'such', 'take', 'ten', 'than', 'that', 
'the', 'their', 'them', 'themselves', 'then', 'thence', 'there', 'thereafter', 'thereby', 'therefore', 'therein', 
'thereupon', 'these', 'they', 'third', 'this', 'those', 'though', 'three', 'through', 'throughout', 'thru', 'thus', 
'to', 'together', 'too', 'top', 'toward', 'towards', 'twelve', 'twenty', 'two', 'under', 'unless', 'until', 'up', 
'upon', 'us', 'used', 'using', 'various', 'very', 'via', 'was', 'we', 'well', 'were', 'what', 'whatever', 'when', 
'whence', 'whenever', 'where', 'whereafter', 'whereas', 'whereby', 'wherein', 'whereupon', 'whenever', 
'whether', 'which', 'while', 'whither', 'who', 'whoever', 'whole', 'whom', 'whose', 'why', 'will', 'with', 
'within', 'without', 'would', 'yet', 'you', 'your', 'yours', 'yourself', 'yourselves'
Is removing stop words always a good thing?

“To be or not to be”
Some Words Mean the Same Thing?

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

Should capitalization matter?

What about:
• walk, walking
• democracy, democratic, democratization
• good, better

Merging modified versions of "same" word to be analyzed as a single word is called lemmatization
(we'll see software for doing this shortly)
What about a word that has multiple meanings?

Challenging: try to split up word into multiple words depending on meaning (requires inferring meaning from context)

This problem is called word sense disambiguation (WSD)
Treat Some Phrases as a Single Word?

**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

First need to detect what are "named entities": called **named entity recognition**
(we'll see software for doing this shortly)

Treat as single 2-word phrase “United States”?
Some Other Basic NLP Tasks

- **Tokenization**: figuring out what are the atomic "words" (including how to treat punctuation)

- **Part-of-speech tagging**: figuring out what are nouns, verbs, adjectives, etc

- **Sentence recognition**: figuring out when sentences actually end rather than there being some acronym with periods in it, etc
Bigram Model

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.

Ordering of words now matters (a little)

“Vocabulary size” (# unique cards) dramatically increases!

If using stopwords, remove any phrase with at least 1 stopword

1 word at a time: unigram model
2 words at a time: bigram model
$n$ words at a time: $n$-gram model