Data Processing

Its graph o’clock
Housekeeping

- Midterm on Wednesday (tomorrow)! You will do great!
- No Wednesday review session this week!
- Breakout was due yesterday, Crypto (shorter assignment) is out!
- Short lecture today!
Today

- Midquarter feedback breakdown
  - TLDR: We are slowing down

- Art show!

- Data visualization
  - Recap processing CSVs
  - Do some fun data visualization!
Midquarter Feedback - speed

- Many of you feel we are going very fast

- You are not behind! This class is very front-loaded. You are learning a new language and a new way of thinking very quickly!

- We will start slowing down - the second half of the course reinforces what we’ve learned so far. There are way fewer new concepts in the second half of the course

- You are not missing anything if you don’t understand everything in one lecture - we expect that! That’s why we have section, review sessions, LAIR and homeworks!
Midquarter Feedback - time on Assignments

- 85% of you take 10-15 hours on an assignment or less - This is our goal!

- Spending more than that does not mean you are behind (I took 20+ hours on 106 assignments) - **we want to see you in LAIR/OH/Ed more.** Ask for help early and often!

- According to most reports, OH have been helpful, yay!
We are going to keep up the interactive lectures!

Would you prefer lectures be less interactive (more lecture, fewer prompts to class), the same, or more interactive than they are now?

41 responses

- 65.9% The same
- 12.2% Less
- 9.8% More
- 9.8% Significantly more
- Significantly less
Midquarter feedback - lectures

- You really liked animation - yay! That's still my favorite lecture demo

- You haven’t loved lists/strings - these are confusing topics

- We will get some more hands-on experience with lists and strings today, no “here are a bunch of functions” slides
Midquarter feedback - Challenges

- Lots of you find the assignments very challenging. This is a good thing, but remember not to struggle through them alone! Come to LAIR/OH

- Keeping up with the speed/amount of content was another big challenge - you are learning a TON really really fast! We are slowing down in the second half of the quarter

- Breakout was a major challenge, but many of you reported it’s been a good challenge - we hope so!
Midquarter feedback - wins

- You all inspire me! This was my favorite part to read :) Take a second to appreciate all the hard work you’ve put in!

  completing the assignments

  Learning coding for the first time

  Completing assignments

My biggest win in the course so far is overcoming the above challenge every week and not giving up.

  Being able to figure out the homework after much trial and error.

  Being more and more comfortable with coding
Today

Midquarter feedback breakdown
  TLDR: We are slowing down :)

- Art show!
  - Remember Bluescreen?

- Data visualization
  - Recap processing CSVs
  - Do some fun data visualization!
A ~selection~
A ~selection~

\[ E = MC^2 \]
Flat Earth Gang
Winners....

Drumroll please!
Best Artistic

Cedric Heidt

Sihan Liu
Best Humor

Phoebe Deimler

Sam Scott
Best use of background

Marshall Schneider

Chase Mullens
Best Famous Person

Elena Yu

Bill Klimczak
Today

- Midquarter feedback breakdown
  - TLDR: We are slowing down :)

- Art show!
  - Remember Bluescreen?

- Data visualization
  - Recap processing CSVs
  - Do some fun data visualization!
Recall: read_csv.py

- Write a program that can take in a filename of a .csv file and a column number

- Print out every row in the csv file as a list (implement this first)

- Then, print just the data in each row from the column provided

- Example: Run
  python3 read_csv_soln.py staff_info.csv 1
def main():
    args = sys.argv[1:]
    filename = args[0]

    #print everything in file
    for line in open(filename):
        line = line.strip()  
        row = line.split(',','
        print(row)
Recall: read_csv.py

```python
def main():
    args = sys.argv[1:]
    filename = args[0]

    #print everything in file
    for line in open(filename):
        line = line.strip()
        row = line.split(',','
        print(row)

filename = 'staff_info.csv'
```
Recall: read_csv.py

```
staff_info.csv

name,degree,concentration,title
Frankie,BS MS,CS,Lecturer
Ecy,BS,SymSys,Head TA

def main():
    args = sys.argv[1:]
    filename = args[0]

    #print everything in file
    for line in open(filename):
        line = line.strip()
        row = line.split(',,')
        print(row)

line = 'name,degree,concentration,title\n'
```

Terminal
```
python3 read_csv_soln.py
staff_info.csv 1
```
Recall: read_csv.py

```python
def main():
    args = sys.argv[1:]
    filename = args[0]

    line = 'name,degree,concentration,title'
    for line in open(filename):
        line = line.strip()
        row = line.split(',,')
        print(row)
```

Terminal

```
python3 read_csv_soln.py staff_info.csv 1
```
```
staff_info.csv
name,degree,concentration,title
Frankie,BS MS,CS,Lecturer
Ecy,BS,SymSys,Head TA

def main():
    args = sys.argv[1:]
    filename = args[0]

    #print everything in file
    for line in open(filename):
        line = line.strip()
        row = line.split(',','
        print(row)
```

Recall: read_csv.py

### staff_info.csv

<table>
<thead>
<tr>
<th>name</th>
<th>degree</th>
<th>concentration</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankie</td>
<td>BS</td>
<td>MS</td>
<td>CS, Lecturer</td>
</tr>
<tr>
<td>Ecy</td>
<td>BS</td>
<td>SymSys</td>
<td>Head TA</td>
</tr>
</tbody>
</table>

```python
def main():
    args = sys.argv[1:]
    filename = args[0]

    # print everything in file
    for line in open(filename):
        line = line.strip()
        row = line.split(',', ')
        print(row)
```

### Terminal

```
python3 read_csv_soln.py
staff_info.csv 1
['name', 'degree', 'concentration', 'title']
```

```
line = 'name,degree,concentration,title'
row = ['name', 'degree', 'concentration', 'title']
```
```python
def main():
    args = sys.argv[1:]
    filename = args[0]

    # print everything in file
    for line in open(filename):
        line = line.strip()
        row = line.split(',','
        print(row)
```

```python
line = 'Frankie,BS MS,CS,Lecturer\nrow = [
    'name',
    'degree',
    'concentration',
    'title',
]```
Recall: read_csv.py

```python
staff_info.csv

name,degree,concentration,title
Frankie,BS MS,CS,Lecturer
Ecy,BS,SymSys,Head TA

def main():
    args = sys.argv[1:]
    filename = args[0]

    line = line.strip()  #print everything in file
    for line in open(filename):
        row = line.split(',')
        print(row)

line = 'Frankie,BS MS,CS,Lecturer'
row = [row[0], row[1], row[2], row[3]]
```

Terminal

```
python3 read_csv_soln.py
staff_info.csv 1
['name', 'degree', 'concentration', 'title']
```
Recall: read_csv.py

```python
def main():
    args = sys.argv[1:]
    filename = args[0]

    # print everything in file
    for line in open(filename):
        line = line.strip()
        row = line.split(',')
        print(row)
```

```
line = 'Frankie,BS MS,CS,Lecturer'
row = ['Frankie', 'BS MS', 'CS', 'Lecturer']
```

Terminal

```
python3 read_csv_soln.py staff_info.csv 1
['name', 'degree', 'concentration', 'title']
```

staff_info.csv

```
name,degree,concentration,title
Frankie,BS MS,CS,Lecturer
Ecy,BS,SymSys,Head TA
```
Recall: read_csv.py

```python
def main():
    args = sys.argv[1:]
    filename = args[0]

    #print everything in file
    for line in open(filename):
        line = line.strip()
        row = line.split(',

        print(row)
```

Terminal:
```
python3 read_csv_soln.py
staff_info.csv 1
['name', 'degree', 'concentration', 'title']
['Frankie', 'BS MS', 'CS', 'Lecturer']
```
Recall: read_csv.py

```python
def main():
    ... code from before ...
    col_num = int(args[1])
    #print everything in column
    for line in open(filename):
        line = line.strip()
        row = line.split(',' , ')
        print(row[col_num])
```

Terminal

```
python3 read_csv_soln.py
staff_info.csv 1
['name', 'degree', 'concentration', 'title']
['Frankie', 'BS MS', 'CS', 'Lecturer']
['Ecy', 'BS', 'SymSys', 'Head TA']
```
Recall: read_csv.py

```
staff_info.csv
name, degree, concentration, title
Frankie, BS MS, CS, Lecturer
Ecy, BS, SymSys, Head TA
```

```
def main():
    ... code from before ...
    col_num = int(args[1])
    # print everything in column
    for line in open(filename):
        line = line.strip()
        row = line.split(',', ')
        print(row[col_num])
```

```
Terminal
python3 read_csv_soln.py
staff_info.csv 1
['name', 'degree', 'concentration', 'title']
['Frankie', 'BS MS', 'CS', 'Lecturer']
['Ecy', 'BS', 'SymSys', 'Head TA']
```

```
line = ‘name, degree, concentration, title
'
def main():
    # code from before ...
    col_num = int(args[1])
    #print everything in column
    for line in open(filename):
        line = line.strip()
        row = line.split(',,')
        print(row[col_num])

line = 'name,degree,concentration,title'

python3 read_csv_soln.py
staff_info.csv 1

['name', 'degree', 'concentration', 'title']
['Frankie', 'BS MS', 'CS', 'Lecturer']
['Ecy', 'BS', 'SymSys', 'Head TA']
Recall: read_csv.py

```python
def main():
    ... code from before ...
    col_num = int(args[1])
    # print everything in column
    for line in open(filename):
        line = line.strip()
        row = line.split(',', ',')
        print(row[col_num])
```
def main():
    ... code from before ...
    col_num = int(args[1])
    # print everything in column
    for line in open(filename):
        line = line.strip()
        row = line.split(',
        print(row[col_num])

line = 'name,degree,concentration,title'
row =

<table>
<thead>
<tr>
<th>name</th>
<th>degree</th>
<th>concentration</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankie</td>
<td>BS MS</td>
<td>CS</td>
<td>Lecturer</td>
</tr>
<tr>
<td>Ecy</td>
<td>BS</td>
<td>SymSys</td>
<td>Head TA</td>
</tr>
</tbody>
</table>
### Key Idea:

**staff_info.csv**

<table>
<thead>
<tr>
<th>name</th>
<th>degree</th>
<th>concentration</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankie</td>
<td>BS, MS</td>
<td>CS</td>
<td>Lecturer</td>
</tr>
<tr>
<td>Ecy</td>
<td>BS</td>
<td>SymSys</td>
<td>Head TA</td>
</tr>
</tbody>
</table>

```python
for line in open(filename):
    line = line.strip()
    row = line.split(', , )
```
Let’s do some data processing!

Graphs graphs graphs graphs
data_processing.py

- Write a program that allows the user to specify the filename of a CSV, a column number in that CSV, a min_frequency and a max_frequency, and any number of string values
- Display a bar chart representing the frequency with which each string value appears in the specified column in the dataset
- (Demo in the started code)
- Use the pre-made `make_bar_chart` function
- Decompose logic to process the file
- Use it on our anonymized Assn0 answers!
data_processing.py
Milestones

1. Understand provided code

2. Write function that returns a list of frequencies for each string in the given list of values

3. Test above function on small dataset

4. Call `make_bar_chart`

5. Run on Assn0 dataset and explore!
Key idea!

Test your data-processing functions on small files, where you can manually tally expected output!
Recap

- You all are excellent programmers and artists

- We can use string functions to interpret files as rows of data - each row is a list!

- Once we have our list rows, we can use our graphics abilities to make graphs!