Compsci 101
Sorting, CSV

Susan Rodger
November 10, 2022

S is for ...

- Software
  - Joy, sorrow, fun, changing the world
- System and sys
  - Connecting to the machine at different levels
- Sorting
  - From hat to tim to more

Barbara Liskov

- Among first women in US to earn Ph.D. in Computer Science: 1968
- Turing Award 2008, Software Engineering and Programming Languages
- Object-Oriented
  - CLU
- Liskov Substitution Principle

“Every time you exchange e-mail with a friend, check your bank statement online, or run a Google search, you are riding the momentum of her research” – MIT President Rafael Reif about Liskov

Announcements

- APT 5 due today!
- Assignment 5 due Thurs, Nov 17
- No Lab this week
- Reading and Sakai Quizzes due next week
- APT Quiz 2 – today through Monday
APT Quiz 2 Nov 10-14

• Opens 10/14 Noon
• Closes at 11pm 11/14 – must finish all by this time
• There are two parts based on APTs 1-5
  • Each part has two APT problems
  • Each part is 3 hours – more if you get accommodations
  • Each part starts in Sakai under tests and quizzes
  • Sakai is a starting point with countdown timer that sends you to a new apt page just for each part
  • Could do each part on different day or same days
• Old APT Quiz so you can practice (not for credit) – on APT Page

APT Quiz 2

• Is your own work!
  • No collaboration with others!
  • Use your notes, lecture notes, your code, textbook
  • DO NOT search for answers!
  • Do not talk to others about the quiz until grades are posted
• Post private questions on Ed Discussion
  • We are not online between 9pm and 9am!
  • We are not on all the time, especially weekends
  • Will try to answer questions between 9am – 9pm
    • About typos, cannot help you in solving APTs
• See 101 APT page for tips on debugging APTs

Don't go to Sakai to start APT Quiz until you are ready to start
If you click on it, you start it!
PFTD

• Sorting
  • Sorting using standard Python APIs
• CSV Library
  • How to read data using standard Python APIs
• Lambda
  • Language construct to make sorting simpler (next week)

Why Sort Data?

• Help understand data
  • Great American Eclipse, August 21, 2017
  • Spotify tracked the playing of the song

Why Sort Data?

• Every field needs to visualize and understand data
  • Sorting helps with this from movies to policy to sports to location of infections to

How your GIS department can respond to COVID-19

A staggering wealth of geospatial information has emerged regarding the COVID-19 outbreak. Dashboards, near real-time services, and GitHub repositories have built the foundation for an extraordinarily transparent response effort.
How To Sort: Algorithms

• Does scale matter? It depends!

• You're playing Spades, Hearts, Bridge, Go-Fish
  • How you sort doesn't really matter, but whether you sort makes play more efficient? Better?

• Many ways to sort
  • Bubblesort, Insertion sort, Selection sort
  • Quicksort, Mergesort, Timsort, Bogo sort
  • Python uses Timsort

Solve a Larger Problem

• Suppose I were to give you the top 1000 artists
  • Top 1,000 songs, find top 10 artists
  • How many songs per artist?

Scale

• As the size of the problem grows we want ...
  • The algorithm to still work and be fast!
  • What to do?

• Search example
  • Google search results work
  • SoundHound/Shazam results work
  • ContentID on YouTube results work
Python to the Rescue

- Using `.sort(...)`, `sorted(...)`, and `lambda`
- Using CSV library and its API
  - CSV – Comma Separated Values
- Why use the CSV library?
  - How to handle the song “Hello, I Love You”?
  - Row 166 in spreadsheet

Hits by Artists: SongReader.py

- What is returned by this function?
- details of csv: `next` and no `split` and ...

```python
def countByArtist(name):
    csvf = open(name, 'r', encoding='utf-8')
    reader = csv.reader(csvf)
    header = next(reader)
    print('header row labels', header)
    data = {}
    for row in reader:
        artist = row[2]
        if artist not in data:
            data[artist] = 0
        data[artist] += 1
    csvf.close()
    return data
```

WOTO-2 countByArtist

Two APIs: CSV and Sorting

- CSV Library to read and process data
  - Comma-separated, but can by ":" separated, or any character as we'll see later

- Similar to reading a file – returned by `open`
  - Iterable is returned by `csv.reader`
  - The `next` function advances iterable
  - Don't call `split`, we can access by index
    - Also by header-row label with `csv.DictReader`
CSV API

- `freader = csv.reader(file)` – returns an iterable
  - Every line from the file in a form ready for you
- `line = next(freader)`
  - Gives you next row as list of strings
- `for row in freader:`
  - Gives you the rest of rows as list of strings

What does this do? `freader` an iterable
Where name is a filename

```python
csvf = open(name, 'r', encoding='utf-8')
freader = csv.reader(csvf)
print("freader", freader)
header = next(freader)
print("header", header)
for row in freader:
    print("row", row)
```

What if you call `next` one extra time?
Where name is a filename

```python
csvf = open(name, 'r', encoding='utf-8')
freader = csv.reader(csvf)
print("freader", freader)
header = next(freader)
print("header", header)
nextline = next(freader)
print("next", nextline)
for row in freader:
    print("row", row)
```

Sorting to Print/Visualize

- Dictionary is ('Beatles', 51) tuples
  - But tuples not in order, so we must ...

```python
if __name__ == '__main__':
counts = countByArtist("data/top1000.csv")
print("\nFirst 5 artists:")
for artist in sorted(counts.items())[:5]:
    print(artist)

print("\nTop 5 artists:")
sortedCount = sorted([(a[1], a[0]) for a in counts.items()])
sortedArtists = [(a[1], a[0]) for a in sortedCount]
for artist in sortedArtists[-5:]:
    print(artist)
```
Sorting API and Sorting Concepts

- What is `counts.items()` – how is it sorted?
  ```python
  print('\nFirst 5 artists:')
  for artist in sorted(counts.items())[:5]:
    print(artist)
  ```

- What does `sorted` return?
  - A list, you can slice a list, look for clues!
  - What can be sorted? A sequence
  - `sorted(counts.items())`

Sorting by Number of Songs

- Sort by first value vs sort by second value
  - Need to put sequence back to original format
  ```python
  print('\nFirst 5 artists:')
  for artist in sorted(counts.items())[:5]:
    print(artist)
  
  print('\nTop 5 artists:')
  sortedArtists = sorted([(a[1], a[0]) for a in counts.items()])
  sortedArtists = [(a[1], a[0]) for a in sortedArtists]
  for artist in sortedArtists[-5:]:
    print(artist)
  ```

Python Sorting API

- We'll use both `sorted()` and `.sort()` API
  - How to call, what options are
  - How to sort on several criteria
  ```python
  print('\nFirst 5 artists:')
  for artist in sorted(counts.items())[:5]:
    print(artist)
  
  print('\nTop 5 artists:')
  sortedArtists = sorted([(a[1], a[0]) for a in counts.items()])
  sortedArtists = [(a[1], a[0]) for a in sortedArtists]
  for artist in sortedArtists[-5:]:
    print(artist)
  ```

- Creating a new list, modifying existing list
  - `sorted()` creates list from .. iterable
  - `x.sort()` modifies the list `x`, no return value!