Data Collection and After: Miscellaneous Topics

Session 5 @ CBA Batch 12
April 2019

sudhir_voleti@isb.edu

1

DC from Bar & QR codes

Using Py

2

Reading (& Writing) Barcodes & QR codes

- Bar & QR codes are a critical part of Automatic Identification & data Capture (AIDC) systems.
- Why care about bar & QR codes?
- What is the principle behind code reading (and writing)?
- What are 1-D versus 2-D codes?
- Open '**Data extraction from codes.ipynb**'

3

Bar & QR Codes : Review and recap

- A good time to take a step back and review learnings from this exercise.
- What libraries did we call?
- What main inbuilt functions did we use?
- What user-defined functions did we use?
- Any other comment on learnings? Applications? Assignments?

4

Text DC from Images

OCR in R

5

An OCR primer

- We've seen how to read text predefined in the 'character' class via web-scraping.
- But there may be a wealth of text data stored in images as well. Examples?
- Reading this requires optical character recognition (OCR) - which involves serious amounts of machine-training.
- In what follows, we'll see in R how to connect to Google's Tesseract OCR engine & do OCR tasks.
  - Py requires the *pytesseract* module for the same functionality
- Open the OCR folder, and go to file '**An OCR primer.Rmd**'.

6

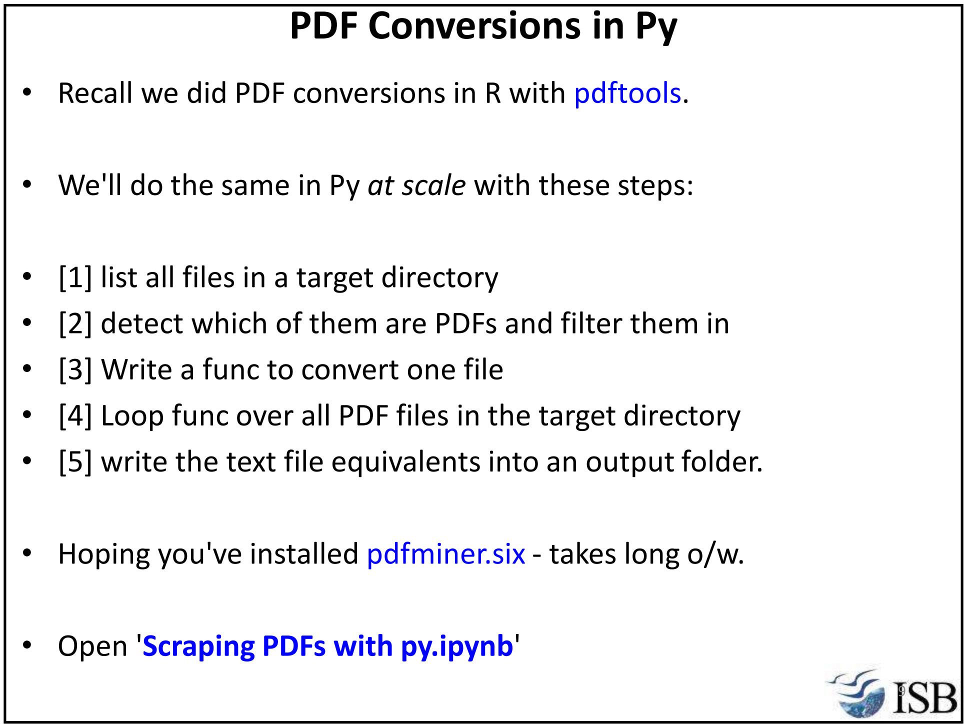
An OCR primer: review and recap

- So how well or poorly did OCR perform based on your initial expectations?
- Some of the places where OCR faced trouble?
  - Options to mitigate the same?
- What is hOCR? How did it help?
- What packages did we see in the primer? What functions do you explicitly recall?

7

Converting PDFs to text with Py

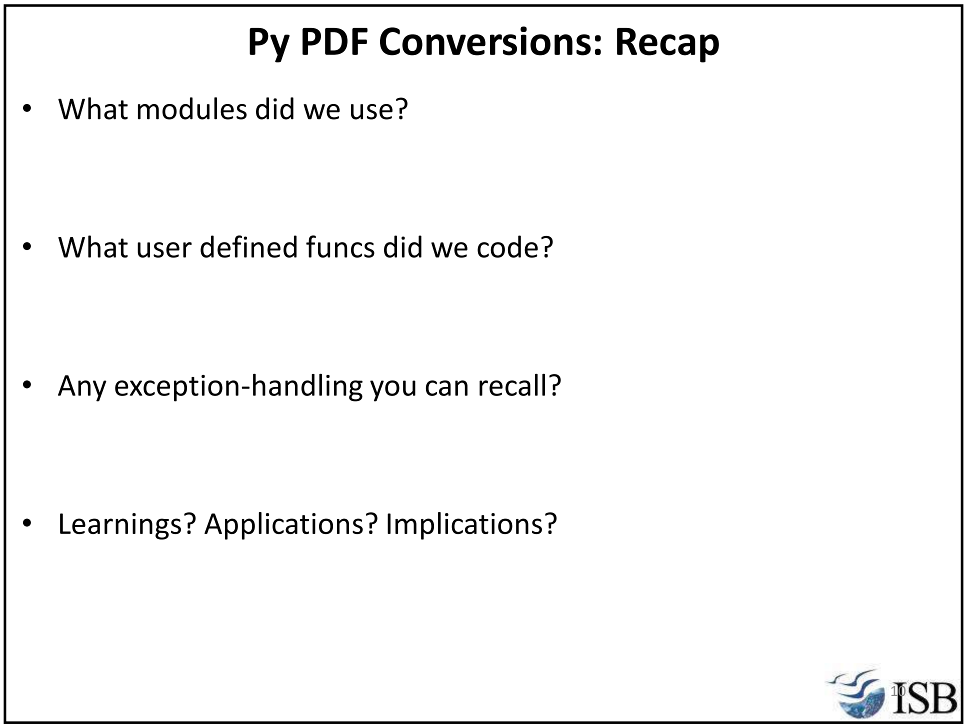
8



PDF Conversions in Py

- Recall we did PDF conversions in R with pdftools.
- We'll do the same in Py *at scale* with these steps:
- [1] list all files in a target directory
- [2] detect which of them are PDFs and filter them in
- [3] Write a func to convert one file
- [4] Loop func over all PDF files in the target directory
- [5] write the text file equivalents into an output folder.
- Hoping you've installed pdfminer.six - takes long o/w.
- Open '**Scraping PDFs with py.ipynb**'

9



Py PDF Conversions: Recap

- What modules did we use?
- What user defined funcs did we code?
- Any exception-handling you can recall?
- Learnings? Applications? Implications?

10

Web-scraping with Py’s Soup

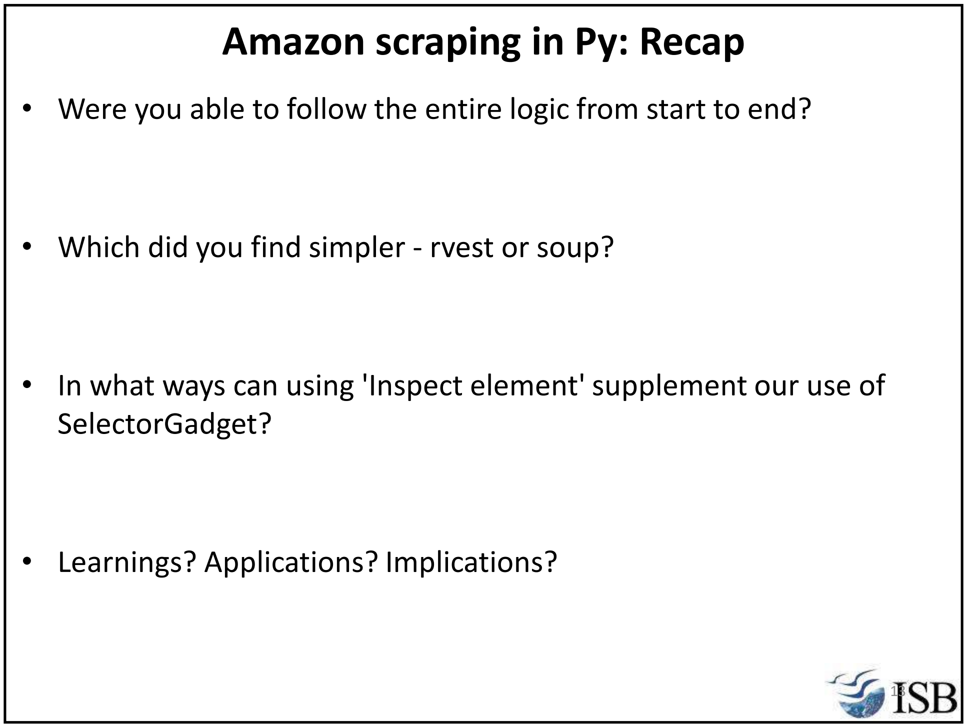
More Examples

11

Revisiting Beaut Soup in Py: Amazon reviews

- Recall scraping Amazon reviews using *rvest*? Let's quickly repeat in Py.
- We'll see a demo of Right-click + Inspect element to ID nodes & CSS elements.
- Here's an illustration:
- Open '**Amazon reviews with beautiful soup.ipynb**'

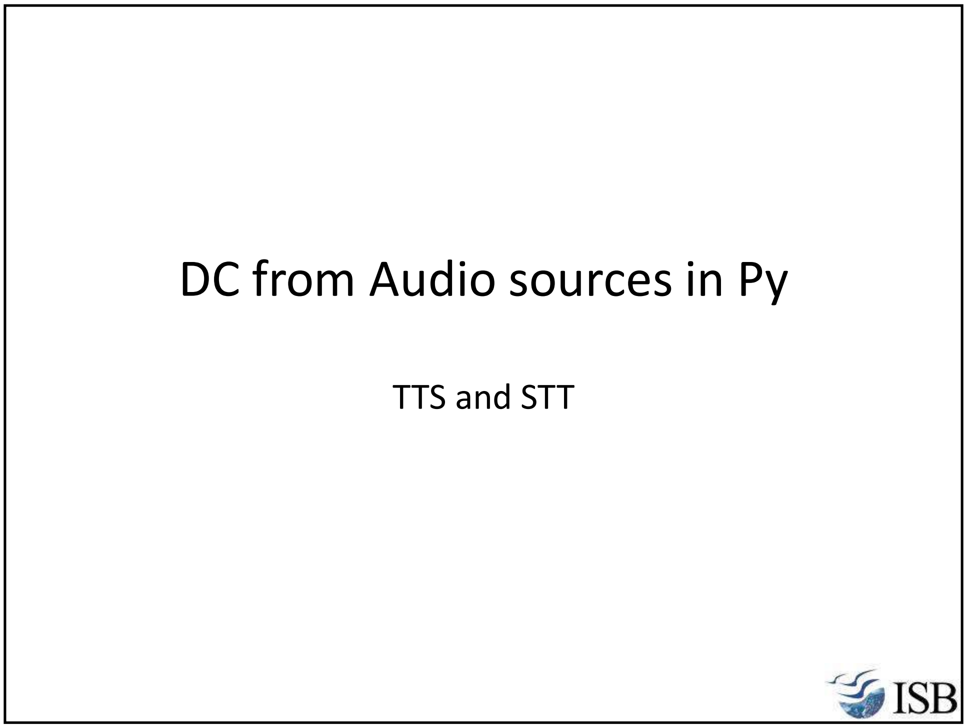
12



Amazon scraping in Py: Recap

- Were you able to follow the entire logic from start to end?
- Which did you find simpler - rvest or soup?
- In what ways can using 'Inspect element' supplement our use of SelectorGadget?
- Learnings? Applications? Implications?

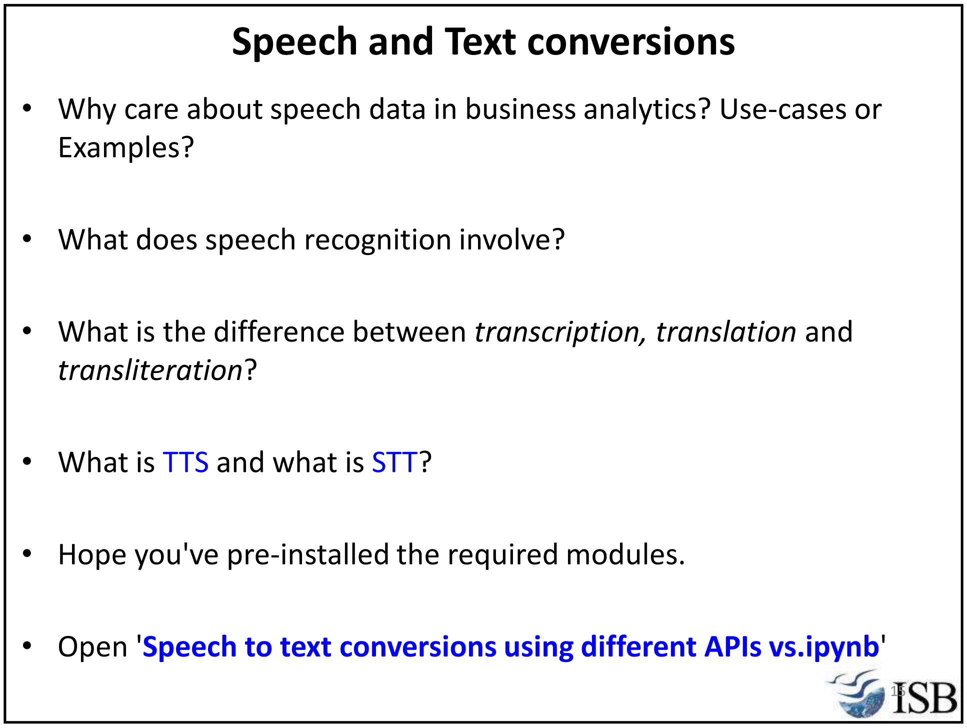
13



DC from Audio sources in Py

TTS and STT

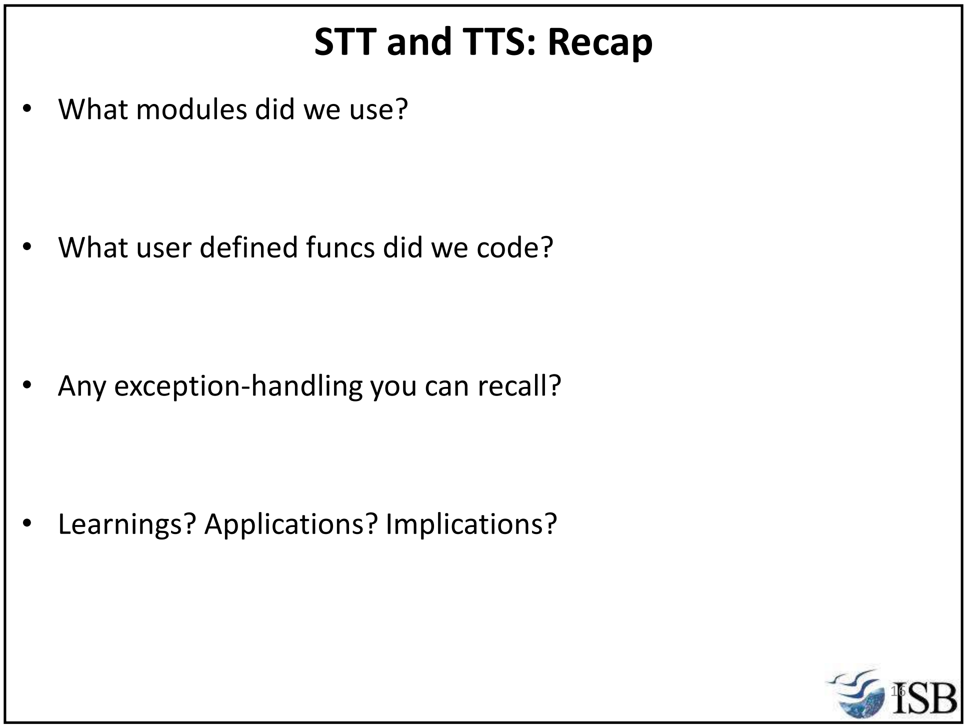
14



Speech and Text conversions

- Why care about speech data in business analytics? Use-cases or Examples?
- What does speech recognition involve?
- What is the difference between *transcription, translation* and *transliteration*?
- What is TTS and what is STT?
- Hope you've pre-installed the required modules.
- Open '**Speech to text conversions using different APIs vs.ipynb**'

15



STT and TTS: Recap

- What modules did we use?
- What user defined funcs did we code?
- Any exception-handling you can recall?
- Learnings? Applications? Implications?

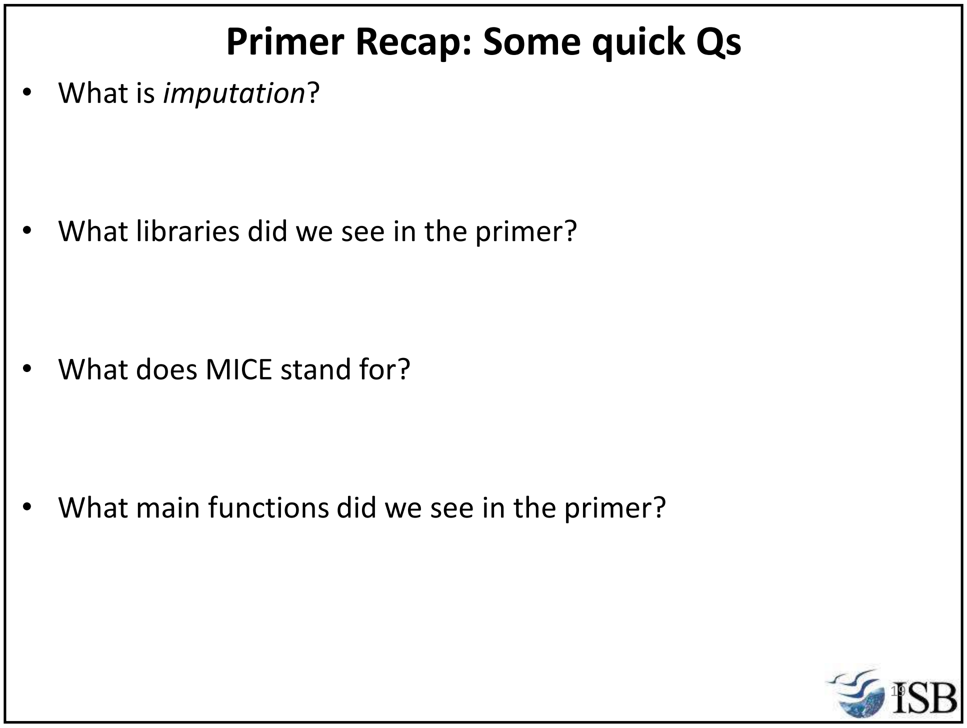
16

Data Quality – Assessment & Improvement

Imputation primer with MICE in R

Missing Data Imputation Primer

- Missing data is a pervasive, nontrivial problem in data handling.
- What are the options available to deal with it? What are some commonplace fixes?
- Two types of missing data - MCAR and MNAR - and why it matters.
- R offers a variety of tools to for missing data handling. So does Py.
- What follows is a quick primer on imputing missing data in R.
  – Open '**Imputation primer.Rmd**'

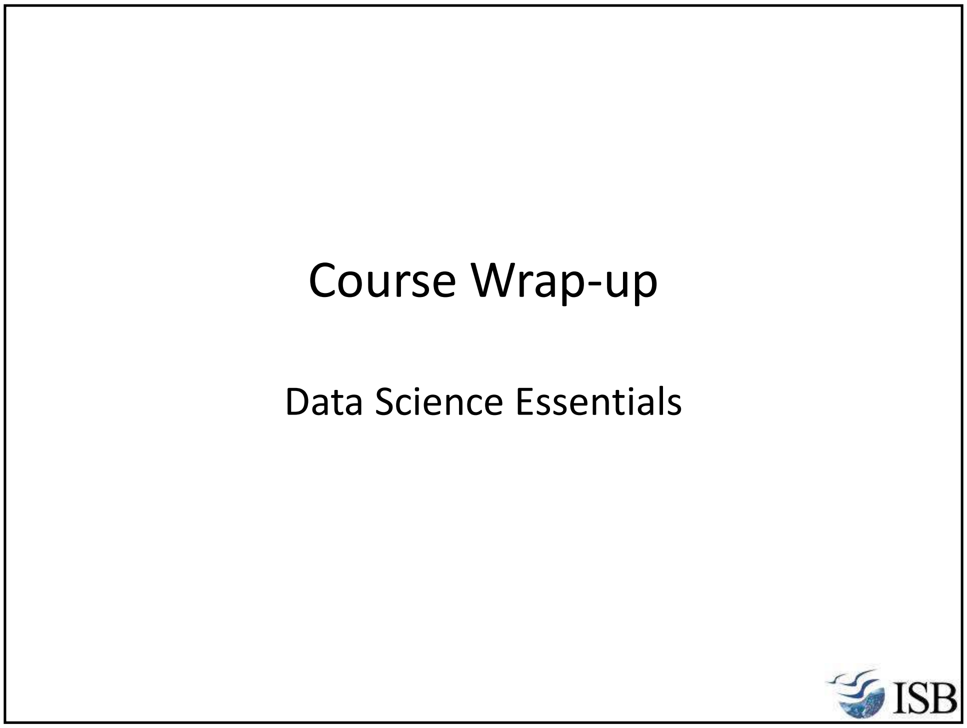


Primer Recap: Some quick Qs

- What is *imputation*?
- What libraries did we see in the primer?
- What does MICE stand for?
- What main functions did we see in the primer?

19

19



Course Wrap-up

Data Science Essentials

20

Who is a Data Scientist?

- An interesting definition goes thus:

- “Someone who is better at Programming than statisticians …
- …. and better at Statistics than programmers.”

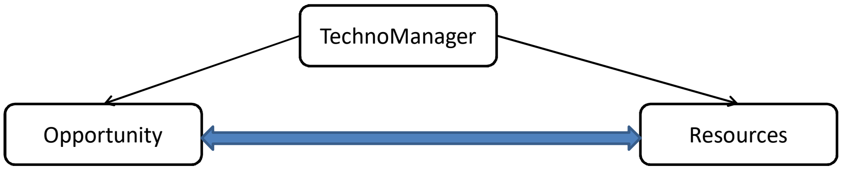
- What should data scientists be good at?
- And what do data scientists spend most of their time doing?

21

Parting Thoughts

22

Course Wrap-up



TechnoManager

Opportunity

Resources

- "A Teacher should show students HOW to think, not WHAT to think." ~ Margaret Mead
- "And above all, be *teachable*." ~ John C Maxwell.
- The business world faces accelerating changes→ Presents both a challenge and an opportunity → E.g., there are now myriad:
  - opportunities for innovative application of core principles from one domain to another +
  - + possibilities to create what didn't exist before within a domain.

ISB

23

Goodbye and Goodluck.

(until your second residency)

ISB

24