wbs
WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

For the
Change
Makers

Programming for Data Analytics

Week 2: Data Collection
Information Systems and Management
Warwick Business School

SQL Exercise

- Please write SQL queries to answer the following questions based the dataset "bigquery-public-data.covid19_nyt".
  - Find top 10 counties with highest fatality rate (#death/#confirmed cases) based on the latest figures, get their names, states and rate. Write another query to find the percentage of people who never or rarely wear for those counties.
  - Find top 10 states with lowest #death based on the latest figures, get their names, #deaths, #confirmed cases. Write another query to find the percentage of people who never or rarely wear for those states.
- *How can you create a dataset that also includes population so your analysis is not biased?

Web Scraping Exercise 1

- Write a script to extract the required information from this webpage (https://www.wbs.ac.uk/about/person/zhewei-zhang):
- Email, phone number, room number, modules taught

Web Scraping Exercise 2

- Check the documentation of BeautifulSoup.
- Write a script to extract the required information for all the research staff of WBS listed on this webpage (https://www.wbs.ac.uk/research/staff/)
- Name, url, job title.
- Save your results into a .csv file.

API Exercise: Retrieving Wikipedia data

- Wikipedia offers a set of APIs for accessing and operating with its data.
- We will use API:query(https://www.mediawiki.org/wiki/API:Query) to retrieve the revision history of specified Wikipedia articles.
- Please write a Python script to collect the most recent 5 revisions to the following pages and save the results into a csv file:
  - University of Warwick
  - Coventry University
  - University of Birmingham
- https://en.wikipedia.org/w/api.php?action=query&format=json&prop=revisions&titles=University%20of%20Warwick&formatversion=2&rvprop=timestamp|user|comment&rvlimit=5

Data collection via Python library

- Various Python libraries available for collecting data.
    - Popular platform: Facebook, Twitter, Wikipedia, Github, etc.
    - API wrapper
    - Bypass platform limitation
    - Not always well-documented.

Further Exercise

- Try to work out the previous exercise with Wikipedia Python library.
- https://pypi.org/project/wikipedia/