



TECHNICAL UNIVERSITY OF LIBEREC  
Faculty of Mechatronics, Informatics  
and Interdisciplinary Studies ■

# TECHNOLOGIE PRO BIG DATA CVIČENÍ I.

*Lukáš Matějů*  
29.8.2024 | TPB



# CVIČENÍ

- opakování – Python
- web scraper pro web [iDNES.cz](https://idnes.cz)
  - navrhnete algoritmus, který bude postupně procházet články webového portálu iDNES.cz
  - ke každému článku uloží název článku, obsah článku, kategorii, počet fotografií, datum publikace a počet komentářů
    - pozor na neúplné údaje
  - informace bude ukládat do textového souboru ve formátu JSON
    - doplňte i funkcionalitu pro načtení dat
    - bude potřebná pro další cvičení
- cílem je stáhnout alespoň 250 MB textových dat
- BONUS: stáhněte minimálně 1 GB textových dat

