Source Selection: Focused Web Crawling

- Why use focused web crawling?
- How do focused web crawlers work?
- What are the benefits and disadvantages of focused web crawling?
- Example toolkits:
  - Python: scrapy
  - Perl: WWW::Mechanize
- *Improving the Performance of Focused Web Crawlers*, Satiris Batsakis, Euripides Petrakis, Evangelos Milios, Data and Knowledge Engineering Journal, 2009.
- MUST BE IN ENGLISH

Source Selection: Wrappers

- Wrappers are used to extract tuples (database entries) from structured web sites
- Discuss the different ways to create wrappers
  - Advantages and disadvantages
  - How do wrappers deal with changing websites?
- Give some examples of different wrapper creation software package and discuss their pros and cons.
- *Automatic Wrappers for Large Scale Web Extraction*, Nilesh Dalvi, Ravi Kumar Mohamed Soliman, VLDB, 2010.
- MUST BE IN ENGLISH

Source Selection: Mechanical Turk

- Explain the mechanism of mechanical turk (focusing on the difference from focused web crawling/wrappers).
- Discuss about the data quality issues in Mechanical Turk (refer to the paper below).
- Amazon Mechanical Turk: `https://en.wikipedia.org/wiki/Amazon_Mechanical_Turk`.
- *Inside the Turk: Understanding Mechanical Turk as a Participant Tool*, Gabriele Paolacci and Jesse Chandler, 2014 (Downloadable from `https://www.researchgate.net/publication/275640107_Inside_the_Turk`).
- MUST BE IN ENGLISH