COSCS494/594 Fundamentals of Digital Archeology Conclusions

Audris Mockus
University of Tennessee
audris@utk.edu

Data Discovery: mining deep web via REST API's, http, search

- No documentation
- No definition on rate limits: being banned
- Poor organization of APIs

Data retrieval

- Retrieve data and application commands (git/hg clone)
- Many architectures in the cloud
- Numerous troubles with large datasets (e.g., passwords, etc)
  - use expect
  - use timeout
- Network bottlenecks: need to verify
- Need to keep track what was accomplished
- Takes weeks

Data storing

- JSON-specific databases: mongodb
- flat files
- keep data compressed

Data analysis

- Distributions/Transformations/Outliers
- Correlations
- Logistic/Linear regression
- Negative binomial, zero-inflated models
- PCA/Factor analysis
- Steming, stopword removal, tfidf, LDM, other text analysis methods

Software engineering

- Version control
- Work in teams
- Planning/Scheduling (proposals)
- Python
- R
- shell script
- managing virtual machines

Key lessons

- Can not trust data unless you have retrieved yourself and understand how it comes to be
  - what events are not observed
  - what attributes are not available
  - a lot of it is incorrect
    - inaccurate sensors, inaccurate entry
    - problems retrieving
  - Never totally sure if the results are accurate
- Why many languages/technologies
  - no one technology best for the entire process
  - technologies come and go, problems remain, need to keep adapting to new technology all the time
- It all takes a lot of work...
  - every class should be a lesson on how to live