Big Data Management
Data Cleaning

Björn Þór Jónsson

Big Data + Data Quality

From Week 6

- Goal: to extract significant **value** from big data
- Key issue: data quality
  - Raw data is often of questionable **veracity**
  - How do we obtain high quality information?
- Data error types:
  - Validity
  - Accuracy
  - Completeness
  - Currency
  - Consistency

Today's Goal

- See real data
  - From the volleyball site blak.is
- Mostly uncontrolled input
  - Limited input checking
  - No proper identifier
  - Lazy (and sometimes funny) users

Approximate Record Matching

- Typically used to merge records
  - From different systems
  - Without common identifiers
  - Uncontrolled input
- Here
  - Merge records over time…

Approximate String Matching

- What are typical errors?
  - Case/capitalisation
  - Character missing/inserted
  - Character misspelled
  - Characters swapped
  - Abbreviations
- String similarity
  - Levenshtein (edit) distance

Wagner–Fischer Algorithm

kitten
sitting

		k	i	t	t	e	n
	0	1	2	3	4	5	6
s	1	1	2	3	4	5	6
i	2	2	1	2	3	4	5
t	3	3	2	1	2	3	4
t	4	4	3	2	1	2	3
i	5	5	4	3	2	2	3
n	6	6	5	4	3	3	2
g	7	7	6	5	4	4	3

The Exercise (on Piazza)

- Understand the data
  - Look at data values (extremes, averages)
- Clean the data (as needed)
  - Given: A simple type conversion
- Create clusters of person records
  - Given: A simple join based on phone numbers
- Create a mapping: FromID → ToID
  - Given: A basic method for connected clusters

If you need "local domain knowledge" call me ☺

But First…

- Course Evaluation…