Lesson Plan 2/22
Tuesday, February 20, 2018  5:15 AM

Admin
- Midterm I next Tuesday
- Questions, clarifications on HW2

HDSC
- Of topical interest: example paper from Facebook modeling behavior to spot illegitimate use:
  https://research.fb.com/publications/copycatch-stopping-group-attacks-by-spotting-lockstep-
  behavior-in-social-networks/

Common Tidying Operations
- Data values as headers (gather)
- More than one entity in a table (normalize)

Cleaning Text
- Regular expressions basics
- Regular expression tools
- The_one_term_per_row_tidy_text_representation

Midterm I discussion
Operations exercise
Midterm I next Tuesday - Questions, clarifications on HW2

HDSC Of topical interest: example paper from Facebook modeling behavior to spot illegitimate use: https://research.fb.com/publications/copycatch

- Stopping group attacks by spotting lockstep behavior in social networks

Common Tidying Operations
- Data values as headers (gather)
- More than one entity in a table (normalize)

Cleaning Text
- Regular expressions basics
- Regular expression tools
- The_one_term_per_row_tidy_text_representation

Midterm I discussion Operations exercise

Lesson Plan 2/22 Tuesday, February 20, 2018 5:15 AM

Tidy Data specific
- One row per entity/relationship
- One column per attribute
- One table: per type of entity/relationship

Problems
1) Headers as values
2) Multiple entity problem

Pew:
entities: group of people of specific religion & income

attr: religion
    income
    number

"gather": fusion performs operation

→

religion income number

religion $10k$-$20k$ $20k$-$40k$ $40k$-$60k$ $60k$-$100k$ $100k$-$200k$ $\notin$ Seattle
Multiple entities in one table (normalize)
Midterm I next Tuesday - Questions, clarifications on HW2 - HDSC

Of topical interest: example paper from Facebook modeling behavior to spot illegitimate use: [https://research.fb.com/publications/copycatch](https://research.fb.com/publications/copycatch)

- STOPPING GROUP ATTACKS BY SPOTTING LOCKSTEP BEHAVIOR IN SOCIAL NETWORKS/

Common Tidying Operations

- Data values as the headers (gather)
- More than one entity in a table (normalize)

Cleaning Text

- Regular Expressions
- Common operations based on RF
- One-token per row representation
Text:

- Document
- Terms

- one_term_per_row

```plaintext

<table>
<thead>
<tr>
<th>Term</th>
<th>document</th>
</tr>
</thead>
<tbody>
<tr>
<td>apic</td>
<td>#apic</td>
</tr>
<tr>
<td>#team</td>
<td>#team</td>
</tr>
</tbody>
</table>
```
Midterm I next Tuesday - Questions, clarifications on HW2

HDSC

Of topical interest: example paper from Facebook modeling behavior to spot illegitimate use: https://research.fb.com/publications/copycatch

- stopping
- group attacks
- by spotting lockstep behavior in social networks/

Common Tidying Operations

Data values as theaders (gather)
- More than one entity in a table (normalize)

Cleaning Text

Regular expressions basics
- Regular expression tools
- The one_term_per_row tidy text representation

Midterm I discussion

Operations exercise

Lesson Plan 2/22

Tuesday, February 20, 2018 5:15 AM