Lecture 20 - sed

DSE 511

Drew Schmidt
2022-11-03

Announcements

- Homework due Monday Nov 7
  - Problem 6:
    1. You may assume the input is an integer
    2. IF YOU USE STRINGS: treat a string as a vector of chars
- Questions?

```
as.integer(2^31)
```

```
## Warning: NAs introduced by coercion to integer range
```

```
## [1] NA
```

Content

- Background
- Basic sed
- Examples

Background

sed

- CLI tool
- What does it do?
  - Frankly, a lot…
  - Commonly: find/replace
- Very powerful!

The Name

- Stream EDitor
- It edits text
- An extremely powerful "find/replace" (and more)

Common Uses for sed

- Software engineering
  - Find/replace
- Data science
  - Find/replace
    - Convert tsv to csv
    - Substitute client/model/... in script
    - ...

My Personal Tier List

1. `git`
2. `grep`

- Big drop off from here!
- A notable contender: `sed`
  - Probably top 5
  - Definitely top 10

Basic sed

Basic sed

- `sed <flags> pattern file(s)`
- `sed -e` - expression
- `sed -i` - in-place editing
- And more...

Chaining sed Commands

- Normal command

```
cmd arg1 | cmd arg2 | cmd arg3 | ...
```

- `sed`

```
sed -e arg1 -e arg2 -e arg3 ...
```

The Typical sed Commands

- Replace the first instance of `a` per line with `b`

```
sed -e 's/a/b/' some_file.txt
```

- Replace the every instance of `a` with `b`

```
sed -e 's/a/b/g' some_file.txt
```

- More complicated things are possible...

Basic sed

```
echo "hello world" > /tmp/example.txt
cat /tmp/example.txt
```

```
## hello world
```

```
sed -e s/h/H/ /tmp/example.txt
```

```
## Hello world
```

```
sed -e s/hello/goodbye/ /tmp/example.txt
```

```
## goodbye world
```

Basic sed

```
sed -e s/o/0/ /tmp/example.txt
```

```
## hell0 world
```

```
sed -e s/o/0/g /tmp/example.txt
```

```
## hell0 w0rld
```

```
sed -e s/^h/H/ -e 's/ w/ W/' /tmp/example.txt
```

```
## Hello World
```

Basic sed

```
echo -e "aa\nab\naa" > /tmp/new_example.txt
cat /tmp/new_example.txt
```

```
## aa
## ab
## aa
```

```
sed -e s/a/x/ /tmp/new_example.txt
```

```
## xa
## xb
## xa
```

```
sed -e s/a/x/g /tmp/new_example.txt
```

```
## xx
## xb
## xx
```

Basic sed

```
sed -e '3s/a/x/' /tmp/new_example.txt
```

```
## aa
## ab
## xa
```

```
sed -e '2,3s/a/x/' /tmp/new_example.txt
```

```
## aa
## xb
## xa
```

```
sed -e '1s/a/x/g' -e '3s/a/x/g' /tmp/new_example.txt
```

```
## xx
## ab
## xx
```

Basic sed

```
sed -e 's/a/x/2g' /tmp/new_example.txt
```

```
## ax
## ab
## ax
```

```
sed -e 's/a/x/3g' /tmp/new_example.txt
```

```
## aa
## ab
## aa
```

Neat Tricks

```
sed -e 's/\(\b[a-z]\)/\(\1\)/g' /tmp/example.txt
```

```
## (h)ello (w)orld
```

```
sed -e 's/\(\b[a-z]\)/\(\1\)/g' /tmp/new_example.txt
```

```
## (a)a
## (a)b
## (a)a
```

Other Uses for sed

```
sed -n -e 1,2p /tmp/new_example.txt
```

```
## aa
## ab
```

```
sed -n -e 1p -e 3p /tmp/new_example.txt
```

```
## aa
## aa
```

Other Uses for sed

```
sed -e '2d' /tmp/new_example.txt
```

```
## aa
## aa
```

```
sed -e '1,3d' /tmp/new_example.txt
```

```
sed -e '2,3d' -e 's/a/x/' /tmp/new_example.txt
```

```
## xa
```

Other Uses for sed

```
sed -n -e 1p -e 3p -e s/a/x/ /tmp/new_example.txt
```

```
## aa
## aa
```

```
sed -e s/a/x/ -n -e 1p -e 3p /tmp/new_example.txt
```

```
## xa
## xa
```

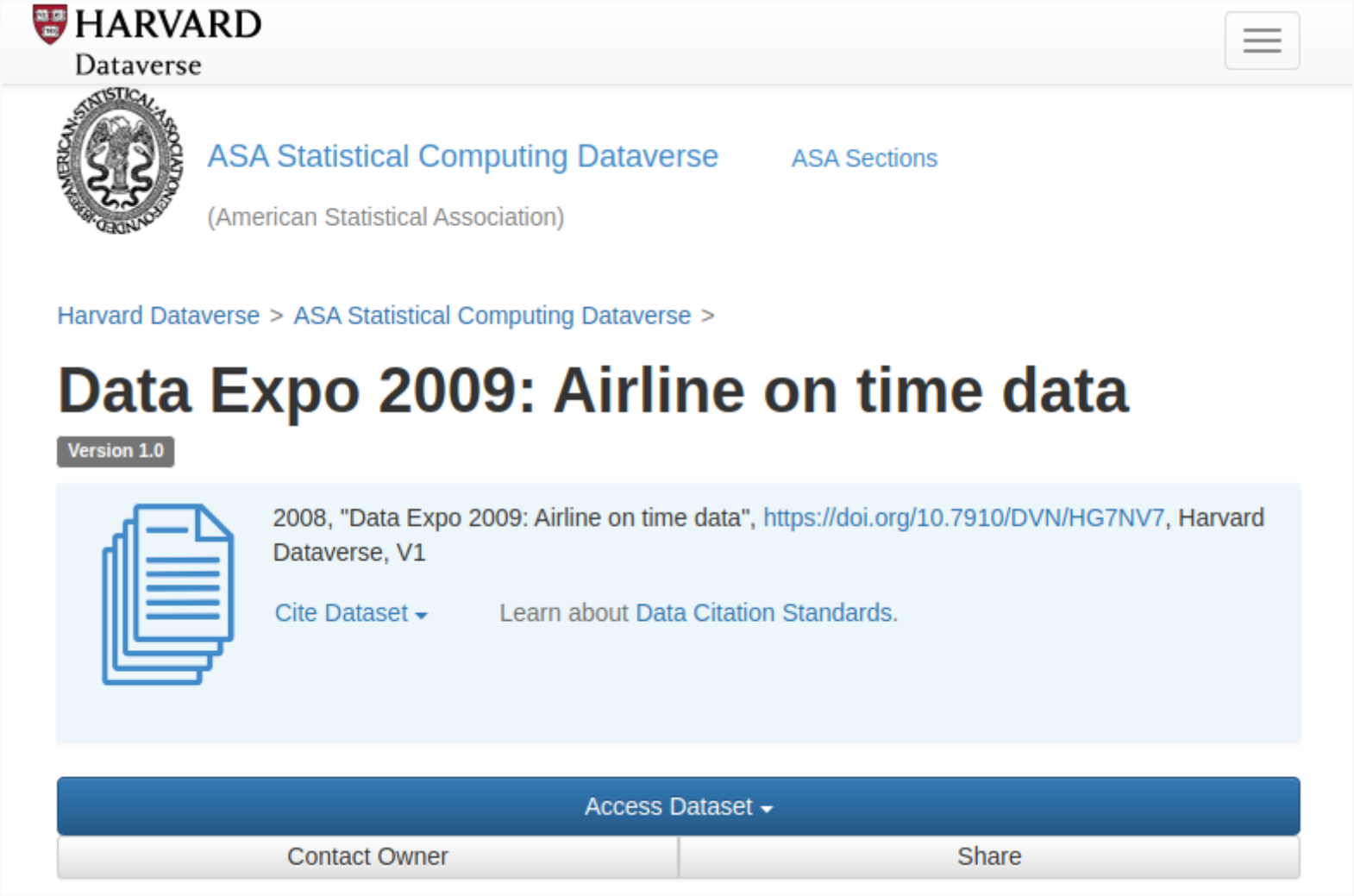
```
sed -i -e s/a/x/ -n -e 1p -e 3p /tmp/new_example.txt
```

```
cat /tmp/new_example.txt
```

```
xa
xa
```

Examples

Example: The Airlines Dataset



Example: The Airlines Dataset

```
head ~/sw/data/airlines/csv/1987.csv | sed -e 's/,/\t/g'
```

```
## Year     Month     DayofMonth     DayOfWeek     DepTime     CRSDepTime     ArrTime     CRSArrTime     Uni
## 1987     10     14     3     741     730     912     849     PS     1451     NA     91     79     NA     23
## 1987     10     15     4     729     730     903     849     PS     1451     NA     94     79     NA     14
## 1987     10     17     6     741     730     918     849     PS     1451     NA     97     79     NA     29
## 1987     10     18     7     729     730     847     849     PS     1451     NA     78     79     NA     -2
## 1987     10     19     1     749     730     922     849     PS     1451     NA     93     79     NA     33
## 1987     10     21     3     728     730     848     849     PS     1451     NA     80     79     NA     -1
## 1987     10     22     4     728     730     852     849     PS     1451     NA     84     79     NA     3     -
## 1987     10     23     5     731     730     902     849     PS     1451     NA     91     79     NA     13
## 1987     10     24     6     744     730     908     849     PS     1451     NA     84     79     NA     19
```

Example: The Airlines Dataset

```
head ~/sw/data/airlines/csv/1987.csv | sed -e 's/,/\t/g' -e 's/NA/__NA__/g'
```

```
## Year    Month    DayofMonth    DayOfWeek    DepTime    CRSDepTime    ArrTime    CRSArrTime    Uni
## 1987    10    14    3    741    730    912    849    PS    1451    __NA__    91    79    __NA__
## 1987    10    15    4    729    730    903    849    PS    1451    __NA__    94    79    __NA__
## 1987    10    17    6    741    730    918    849    PS    1451    __NA__    97    79    __NA__
## 1987    10    18    7    729    730    847    849    PS    1451    __NA__    78    79    __NA__
## 1987    10    19    1    749    730    922    849    PS    1451    __NA__    93    79    __NA__
## 1987    10    21    3    728    730    848    849    PS    1451    __NA__    80    79    __NA__
## 1987    10    22    4    728    730    852    849    PS    1451    __NA__    84    79    __NA__
## 1987    10    23    5    731    730    902    849    PS    1451    __NA__    91    79    __NA__
## 1987    10    24    6    744    730    908    849    PS    1451    __NA__    84    79    __NA__
```

Example: The Airlines Dataset

```
head -n 3 ~/sw/data/airlines/csv/1987.csv | sed -e 's/,/  /g'
```

```
## Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  ArrTime  CRSArrTime  UniqueCarrier  Flig
## 1987  10  14  3  741  730  912  849  PS  1451  NA  91  79  NA  23  11  SAN  SFO  447  NA  NA  0
## 1987  10  15  4  729  730  903  849  PS  1451  NA  94  79  NA  14  -1  SAN  SFO  447  NA  NA  0
```

```
head -n 3 ~/sw/data/airlines/csv/1987.csv | sed -e 's/,/  /g' -e 's/-[0-9]/NA/g'
```

```
## Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  ArrTime  CRSArrTime  UniqueCarrier  Flig
## 1987  10  14  3  741  730  912  849  PS  1451  NA  91  79  NA  23  11  SAN  SFO  447  NA  NA  0
## 1987  10  15  4  729  730  903  849  PS  1451  NA  94  79  NA  14  NA  SAN  SFO  447  NA  NA  0
```

```
head -n 3 ~/sw/data/airlines/csv/1987.csv | sed -e 's/,/  /g' -e 's/-[0-9]/0/g'
```

```
## Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  ArrTime  CRSArrTime  UniqueCarrier  Flig
## 1987  10  14  3  741  730  912  849  PS  1451  NA  91  79  NA  23  11  SAN  SFO  447  NA  NA  0
## 1987  10  15  4  729  730  903  849  PS  1451  NA  94  79  NA  14  0  SAN  SFO  447  NA  NA  0  N
```

Question

But what about
$SOME_ADDITIONAL_COMPLEXITY?

Example: The Airlines Dataset

```
grep "[0-9],TYS," ~/sw/data/airlines/csv/1987.csv | \
    sed -e 's/^.*TYS,//' -e 's/,.*//' | sort | uniq -c
```

```
##     717 ATL
##     389 BNA
##     348 CLT
##      90 CVG
##     340 MEM
##      60 ORD
##     360 PIT
##      92 TRI
```

Example: The Airlines Dataset

```
grep "[0-9],TYS," ~/sw/data/airlines/csv/2005.csv | \
    sed -e 's/^.*TYS,//' -e 's/,.*//' | sort | uniq -c
```

```
##    2670 ATL
##     389 CLE
##    2573 CVG
##    1425 DFW
##       4 DTW
##     621 EWR
##    1466 IAD
##    1036 IAH
##     359 LGA
##     484 MCO
##       4 MEM
##    1916 ORD
##     679 TPA
```

Wrapup

Wrapup

- `sed` is a very powerful find/replace tool
- It does significantly more as well
- Regular expressions aren't perfect!
- You now know what someone means when they say "Sorry s/a/b/".
- We've only worked on "lines"; how do we work on "columns"? 🤔

Questions?