Text Data

STAT 133

Gaston Sanchez

Department of Statistics, UC–Berkeley

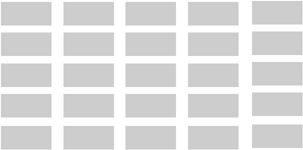
gastonsanchez.com
github.com/gastonstat/stat133
Course web: gastonsanchez.com/teaching/stat133

Datasets

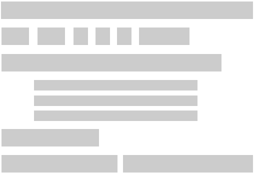
Datasets

You'll have some sort of (raw) data to work with

tabular



non-tabular



Data

- Much of the data we deal with are given to us as plain text
- The data are merely represented by their text form
- Sometimes the data are easily interpreted

Toy Data (tabular layout)

name	gender	height
Leia Skywalker	female	1.50
Luke Skywalker	male	1.72
Han Solo	male	1.80

Typically we get data formed of strings and numeric values

Comma Delimited (csv)

```
name,gender,height,weight,jedi,species,weapon
Luke Skywalker,male,1.72,77,jedi,human,lightsaber
Leia Skywalker,female,1.50,49,no_jedi,human,blaster
Obi-Wan Kenobi,male,1.82,77,jedi,human,lightsaber
Han Solo,male,1.80,80,no_jedi,human,blaster
R2-D2,male,0.96,32,no_jedi,droid,unarmed
C-3PO,male,1.67,75,no_jedi,droid,unarmed
Yoda,male,0.66,17,jedi,yoda,lightsaber
Chewbacca,male,2.28,112,no_jedi,wookiee,bowcaster
```

However ...

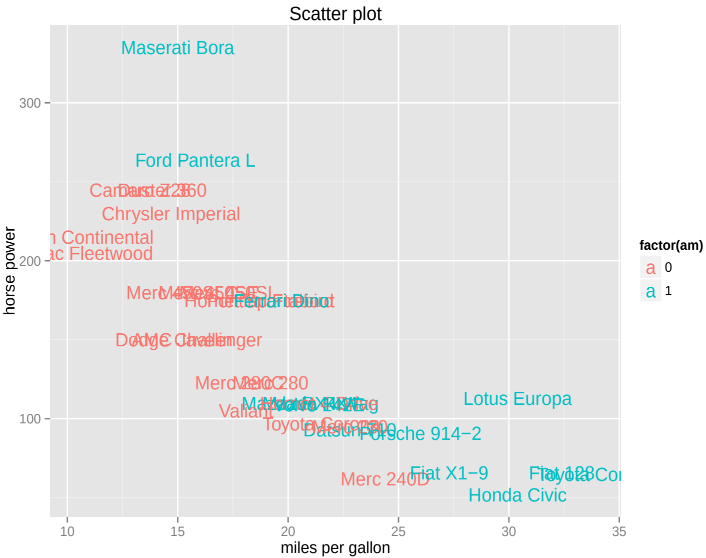
- There are many examples of more complex situations
- It is not uncommon to deal with data that are not as easily interpreted
- And thus the text must be processed to create values of interest

For instance ...

- e.g. when numeric values are embedded into text
- e.g. numeric values not in a regular or simple format
- e.g. numbers in an HTML table
- e.g. data in non-delimited-field formats

Text Everywhere

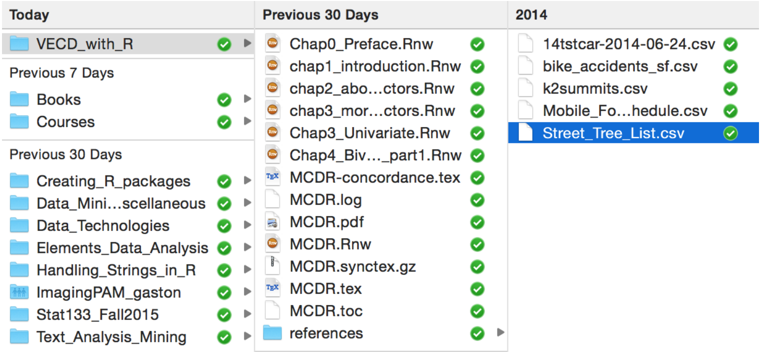
Text in plots



Text in scripts

```r
# =====================================================
# Stat133: Lab 2
# Description: Basics of data frames
# Data: Star Wars characters
# =====================================================

# load "readr
library("readr")

# read data using read_csv()
sw <- read_csv("~/stat133/datasets/starwarstoy.csv")

# use str() to get information about the data frame structure
str(sw)

# use summary() to get some descriptive statistics
summary(sw)

# convert column 'gender' as a factor
sw$gender <- factor(sw$gender)
```

Text: names of files and directories



Today

VECD_with_R

Previous 7 Days

Books
Courses

Previous 30 Days

Creating_R_packages
Data_Mini…scellaneous
Data_Technologies
Elements_Data_Analysis
Handling_Strings_in_R
ImagingPAM_gaston
Stat133_Fall2015
Text_Analysis_Mining

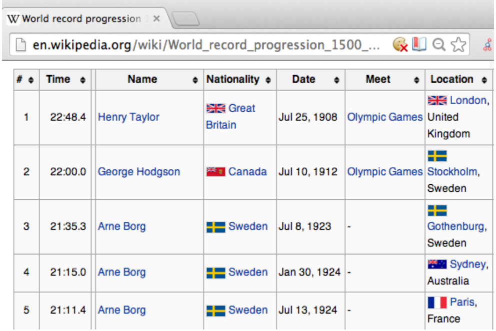
Previous 30 Days

Chap0_Preface.Rnw
chap1_introduction.Rnw
chap2_abo…ctors.Rnw
chap3_mor…ctors.Rnw
Chap3_Univariate.Rnw
Chap4_Biv…_part1.Rnw
MCDR-concordance.tex
MCDR.log
MCDR.pdf
MCDR.Rnw
MCDR.synctex.gz
MCDR.tex
MCDR.toc
references

2014

14tstcar-2014-06-24.csv
bike_accidents_sf.csv
k2summits.csv
Mobile_Fo…hedule.csv
Street_Tree_List.csv

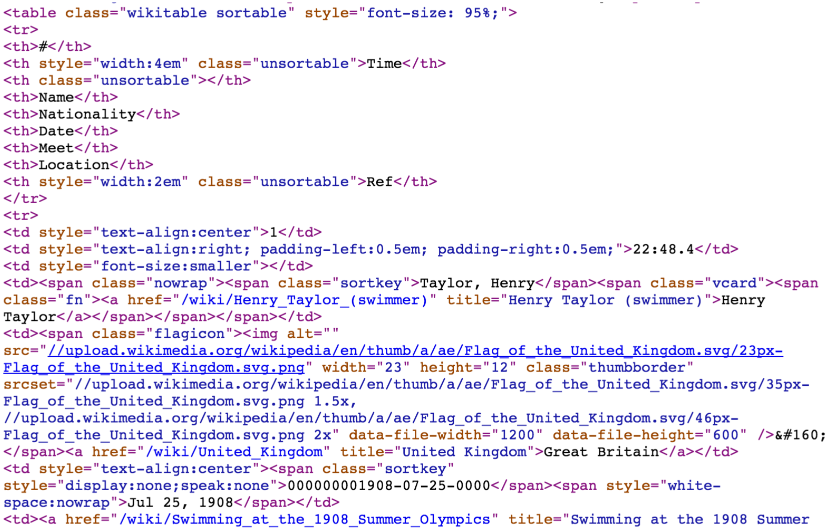
Wikipedia Table



#	Time	Name	Nationality	Date	Meet	Location
1	22:48.4	Henry Taylor	Great Britain	Jul 25, 1908	Olympic Games	London, United Kingdom
2	22:00.0	George Hodgson	Canada	Jul 10, 1912	Olympic Games	Stockholm, Sweden
3	21:35.3	Arne Borg	Sweden	Jul 8, 1923	-	Gothenburg, Sweden
4	21:15.0	Arne Borg	Sweden	Jan 30, 1924	-	Sydney, Australia
5	21:11.4	Arne Borg	Sweden	Jul 13, 1924	-	Paris, France

https://en.wikipedia.org/wiki/World_record_progression_1500_metres_freestyle

Wikipedia Table



```
<table class="wikitable sortable" style="font-size: 95%;">
<tr>
<th>#</th>
<th style="width:4em" class="unsortable">Time</th>
<th class="unsortable"></th>
<th>Name</th>
<th>Nationality</th>
<th>Date</th>
<th>Meet</th>
<th>Location</th>
<th style="width:2em" class="unsortable">Ref</th>
</tr>
<tr>
<td style="text-align:center">1</td>
<td style="text-align:right; padding-left:0.5em; padding-right:0.5em;">22:48.4</td>
<td style="font-size:smaller"></td>
<td><span class="nowrap"><span class="sortkey">Taylor, Henry</span><span class="vcard"><span
class="fn"><a href="/wiki/Henry_Taylor_(swimmer)" title="Henry Taylor (swimmer)">Henry
Taylor</a></span></span></span></td>
<td><span class="flagicon"><img alt=""
src="//upload.wikimedia.org/wikipedia/en/thumb/a/ae/Flag_of_the_United_Kingdom.svg/23px-
Flag_of_the_United_Kingdom.svg.png" width="23" height="12" class="thumbborder"
srcset="//upload.wikimedia.org/wikipedia/en/thumb/a/ae/Flag_of_the_United_Kingdom.svg/35px-
Flag_of_the_United_Kingdom.svg.png 1.5x,
//upload.wikimedia.org/wikipedia/en/thumb/a/ae/Flag_of_the_United_Kingdom.svg/46px-
Flag_of_the_United_Kingdom.svg.png 2x" data-file-width="1200" data-file-height="600" />&#160;
</span><a href="/wiki/United_Kingdom" title="United Kingdom">Great Britain</a></td>
<td style="text-align:center"><span class="sortkey"
style="display:none;speak:none">000000001908-07-25-0000</span><span style="white-
space:nowrap">Jul 25, 1908</span></td>
<td><a href="/wiki/Swimming_at_the_1908_Summer_Olympics" title="Swimming at the 1908 Summer
```

Example: XML Data

Toy Data (XML format)

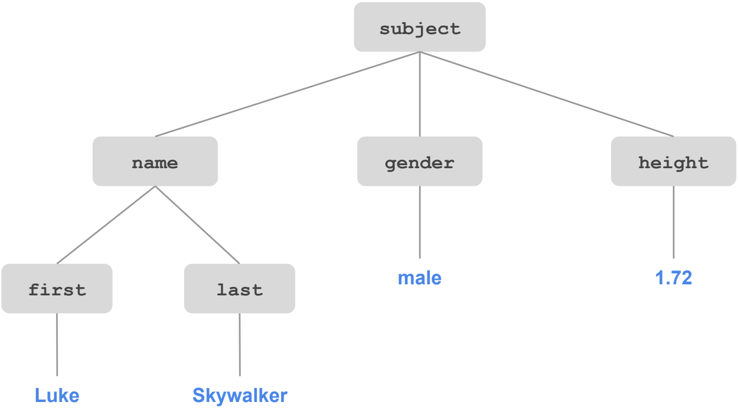
```
<subject>
  <name>
    <first>Luke</first>
    <last>Skywalker</last>
  </name>
  <gender>male</gender>
  <height>1.72</height>
</subject>
<subject>
  <name>
    <first>Leia</first>
    <last>Skywalker</last>
  </name>
  <gender>female</gender>
  <height>1.50</height>
</subject>
```

Toy Data (XML format)

Looking at one `<subject>` node:

```
<subject>
  <name>
    <first>Luke</first>
    <last>Skywalker</last>
  </name>
  <gender>male</gender>
  <height>1.72</height>
</subject>
```

XML hierarchical structure



Extracting Data

- Sometimes we must extract the elements of interest from the text content
- The extraction is done by identifying the patterns where the values occur

Extracting Data

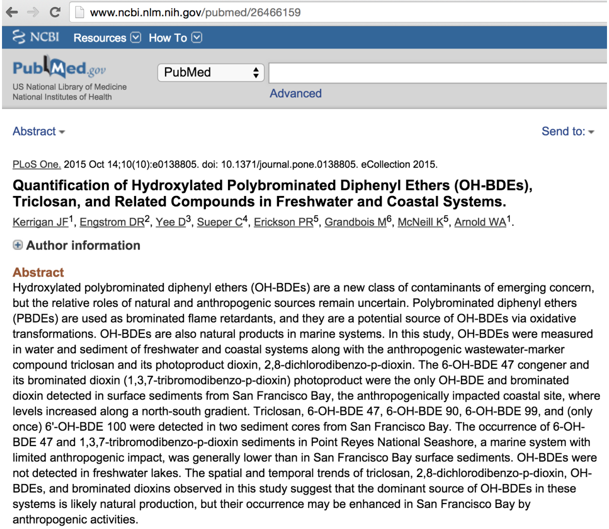
- A different example occurs when text itself makes up the data
- Speech
- Lyrics
- Email messages
- Abstract
- *etc*

Example: Speech

Text of President Barack Obama's State of the Union address, as provided by the White House:

> *Mr. Speaker, Mr. Vice President, members of Congress, distinguished guests and fellow Americans:*
>
> *Last month, I went to Andrews Air Force Base and welcomed home some of our last troops to serve in Iraq. Together, we offered a final, proud salute to the colors under which more than a million of our fellow citizens fought– and several thousand gave their lives.*

Example: Abstract


www.ncbi.nlm.nih.gov/pubmed/26466159
NCBI
Resources
How To
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Advanced
Abstract
Send to:

PLoS One. 2015 Oct 14;10(10):e0138805. doi: 10.1371/journal.pone.0138805. eCollection 2015.

Quantification of Hydroxylated Polybrominated Diphenyl Ethers (OH-BDEs), Triclosan, and Related Compounds in Freshwater and Coastal Systems.

Kerrigan JF[1], Engstrom DR[2], Yee D[3], Sueper C[4], Erickson PR[5], Grandbois M[6], McNeill K[5], Arnold WA[1].

⊕ **Author information**

Abstract

Hydroxylated polybrominated diphenyl ethers (OH-BDEs) are a new class of contaminants of emerging concern, but the relative roles of natural and anthropogenic sources remain uncertain. Polybrominated diphenyl ethers (PBDEs) are used as brominated flame retardants, and they are a potential source of OH-BDEs via oxidative transformations. OH-BDEs are also natural products in marine systems. In this study, OH-BDEs were measured in water and sediment of freshwater and coastal systems along with the anthropogenic wastewater-marker compound triclosan and its photoproduct dioxin, 2,8-dichlorodibenzo-p-dioxin. The 6-OH-BDE 47 congener and its brominated dioxin (1,3,7-tribromodibenzo-p-dioxin) photoproduct were the only OH-BDE and brominated dioxin detected in surface sediments from San Francisco Bay, the anthropogenically impacted coastal site, where levels increased along a north-south gradient. Triclosan, 6-OH-BDE 47, 6-OH-BDE 90, 6-OH-BDE 99, and (only once) 6'-OH-BDE 100 were detected in two sediment cores from San Francisco Bay. The occurrence of 6-OH-BDE 47 and 1,3,7-tribromodibenzo-p-dioxin sediments in Point Reyes National Seashore, a marine system with limited anthropogenic impact, was generally lower than in San Francisco Bay surface sediments. OH-BDEs were not detected in freshwater lakes. The spatial and temporal trends of triclosan, 2,8-dichlorodibenzo-p-dioxin, OH-BDEs, and brominated dioxins observed in this study suggest that the dominant source of OH-BDEs in these systems is likely natural production, but their occurrence may be enhanced in San Francisco Bay by anthropogenic activities.

Example: Web Log

Web log example

```
123.123.123.123 - - [26/Apr/2000:00:23:48 -0400]
"GET /pics/wpaper.gif HTTP/1.0" 200 6248
"http://www.jafsoft.com/asctortf/"
"Mozilla/4.05 (Macintosh; I; PPC)"
```

```
123.123.123.123 - - [26/Apr/2000:00:23:47 -0400]
"GET /asctortf/ HTTP/1.0" 200 8130
"http://search.netscape.com/Computers/Data_Formats"
"Mozilla/4.05 (Macintosh; I; PPC)"
```

Web log data

- The information in the log has a lot of structure
- e.g. the date always appears in square brackets
- However, the information is not consistently separated by the same characters
- Nor is it placed consistently in the same columns in the file

Web log example

Web log content structure:

```
ppp931.on.bellglobal.com
- -
[26/Apr/2000:00:16:12 -0400]
"GET /download/windows/asctab31.zip HTTP/1.0"
200
1540096
"http://www.htmlgoodies.com/downloads/freeware/15.html"
"Mozilla/4.7 [en]C-SYMPA  (Win95; U)"
```

Web log data

- IP address: `ppp931.on.bellglobal.com`
- Username etc: `"- -"`
- Timestamp: `"[26/Apr/2000:00:16:12 -0400]"`
- Access request: `"GET /download/windows/asctab31.zip HTTP/1.0"`
- Result status code: `"200"`
- Bytes transferred: `"1540096"`
- Referrer URL: `"http://www.htmlgoodies.com/downloads/freeware/15.html"`
- User Agent: `"Mozilla/4.7 [en]C-SYMPA (Win95; U)"`

Spam Filtering

Anatomy of an email message

- Three parts:
  - header
  - body
  - attachments (optional)
- Like regular mail, the header is the envelope and the body is the letter
- Plain text

Spam Filtering

Email header

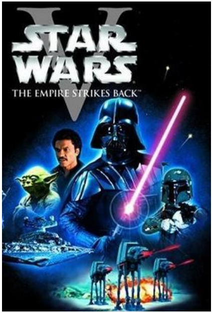
- date, sender, and subject
- message id
- who are the carbon-copy recipients
- return path

Example Email Header

```
Date: Mon, 29 Jun 2015 22:16:19 -0800 (PST)
From: doe@email.edu
X-X-Sender: smith@email.net
To: Txxxx Uxxx <txxxx@uclink.berkeley.edu>
Subject: Re: prof: did you receive my hw?
In-Reply-To: <web-569552@calmail-st.berkeley.edu>
MIME-Version: 1.0
Content-Type: TEXT/PLAIN; charset=US-ASCII
Status: 0
X-Status:
X-Keywords:
X-UID: 9079
```

Example: Movie Scripts

Episode IV



Episode V

Episode VI

STAR WARS

Episode V

THE EMPIRE STRIKES BACK

Script adaptation by
Lawrence Kasdan and Leigh Brackett
from a story by
George Lucas

LUCASFILM LTD.

Reading Text

```r
# read data as string vector
sw <- readLines("StarWars_EpisodeV_script.txt")

sw[1:13]
```

```
##  [1] ""
##  [2] "                              STAR WARS"
##  [3] ""
##  [4] "                              Episode V"
##  [5] "                                   "
##  [6] "                       THE EMPIRE STRIKES BACK"
##  [7] ""
##  [8] "                         Script adaptation by"
##  [9] "                  Lawrence Kasdan and Leigh Brackett"
## [10] "                           from a story by"
## [11] "                             George Lucas"
## [12] ""
## [13] "                            LUCASFILM LTD."
```

Star Wars Episode V script

```
A long time ago, in a galaxy far, far, away...


   It is a dark time for the Rebellion. Although the Death Star has
been destroyed, Imperial troops have driven the Rebel forces from
their hidden base and pursued them across the galaxy.
   Evading the dreaded Imperial Starfleet, a group of freedom fighters
led by Luke Skywalker has established a new secret base on the remote
ice world of Hoth.
   The evil lord Darth Vader, obsessed with finding young Skywalker,
has dispatched thousands of remote probes into the far reaches of
space...
```

Star Wars Episode V script

LUKE: (into comlink) Echo Three to Echo Seven. Han, old buddy, do you read me?

After a little static a familiar voice is heard.

HAN: (over comlink) Loud and clear, kid. What's up?

LUKE: (into comlink) Well, I finished my circle. I don't pick up any life readings.

HAN: (over comlink) There isn't enough life on this ice cube to fill a space cruiser. The sensors are placed. I'm going back.

Reading Text

```
sw[64:74]
```

```
##  [1] "LUKE: (into comlink) Echo Three to Echo Seven. Han, old buddy, do you"
##  [2] "read me?"
##  [3] "           After a little static a familiar voice is heard."
##  [4] ""
##  [5] "HAN: (over comlink) Loud and clear, kid. What's up?"
##  [6] ""
##  [7] "LUKE: (into comlink) Well, I finished my circle. I don't pick up any"
##  [8] "life readings."
##  [9] ""
## [10] "HAN: (over comlink) There isn't enough life on this ice cube to fill a"
## [11] "space cruiser. The sensors are placed. I'm going back."
```

Matching Text

```
grep('LUKE', sw[64:74])

## [1] 1 7

grep('LUKE', sw[64:74], value = TRUE)

## [1] "LUKE: (into comlink) Echo Three to Echo Seven. Han, old buddy, do you"
## [2] "LUKE: (into comlink) Well, I finished my circle. I don't pick up any"
```

Matching Text

```
force_lines <- grep('force', sw)

length(force_lines)
```

```
## [1] 6
```

```
sw[force_lines]
```

```
## [1] "been destroyed, Imperial troops have driven the Rebel forces from"
## [2] "        seconds, other Imperial reinforcements join the scuffle,"
## [3] "        Luke's feet forces the youth to jump back to protect himself."
## [4] "           Suddenly, Vader attacks so forcefully that Luke loses his"
## [5] "        exchange and Luke forces Vader back. Another exchange and"
## [6] "        finally forces him back, away from the edge. The wind soon"
```

Matching Text

```
(force_lines <- grep('Force', sw))
```

```
##  [1] 2550 2562 2877 2878 2890 2912 3126 3180 3184 3339 3340 3634 3637 3656
## [15] 4218 4421 4984
```

```
sw[force_lines]
```

```
##  [1] "EMPEROR: There is a great disturbance in the Force."
##  [2] "EMPEROR: The Force is strong with him. The son of Skywalker must not"
##  [3] "YODA: Run! Yes. A Jedi's strength flows from the Force. But beware of"
##  [4] "the dark side. Anger...fear...aggression. The dark side of the Force"
##  [5] "the Force for knowledge and defense, never for attack."
##  [6] "YODA: That place...is strong with the dark side of the Force. A domain"
##  [7] "YODA: Use the Force. Yes..."
##  [8] "YODA: And well you should not. For my ally in the Force. And a"
##  [9] "feel the Force around you. (gesturing) Here, between you...me...the"
## [10] "YODA: Concentrate...feel the Force flow. Yes. Good. Calm, yes. Through"
## [11] "the Force, things you will see. Other places. The future...the past."
## [12] "LUKE: But I can help them! I feel the Force!"
## [13] "you will be tempted by the dark side of the Force."
## [14] "Knight with the Force as his ally will conquer Vader and his Emperor."
## [15] "VADER: The Force is with you, young Skywalker. But you are not a Jedi"
## [16] "        hurtling at him. Using the Force, Luke manages to deflect it"
## [17] "LUKE: (into comlink) Take care, you two. May the Force be with you."
```

Matching Text

```
(dark_lines <- grep('dark side', sw))
```

```
## [1] 2878 2883 2912 3637 3677 4573
```

```
sw[dark_lines]
```

```
## [1] "the dark side. Anger...fear...aggression. The dark side of the Force"
## [2] "LUKE: Vader. Is the dark side stronger?"
## [3] "YODA: That place...is strong with the dark side of the Force. A domain"
## [4] "you will be tempted by the dark side of the Force."
## [5] "BEN: Luke, don't give in to hate -- that leads to the dark side."
## [6] "VADER: If you only knew the power of the dark side. Obi-Wan never told"
```

Example: Movie Script

What things would you analyze from a movie script?

Example: Movie Script

What things would you analyze from a movie script?

- How many characters?
- Most common words?
- Number of dialogues per character?
- Average number of words per dialogue?
- What's the longest word?

Extracting Text

```r
library(stringr)

# extract first word
str_extract(sw[64:74], "\\w+")
```

```
##  [1] "LUKE"  "read"  "After" NA      "HAN"   NA      "LUKE"  "life"  NA
## [10] "HAN"   "space"
```

Replacing Text

```r
# replace 'LUKE' by 'Luke'
str_replace(sw[64:74], "LUKE", "Luke")
```

```
##  [1] "Luke: (into comlink) Echo Three to Echo Seven. Han, old buddy, do you"
##  [2] "read me?"
##  [3] "           After a little static a familiar voice is heard."
##  [4] ""
##  [5] "HAN: (over comlink) Loud and clear, kid. What's up?"
##  [6] ""
##  [7] "Luke: (into comlink) Well, I finished my circle. I don't pick up any"
##  [8] "life readings."
##  [9] ""
## [10] "HAN: (over comlink) There isn't enough life on this ice cube to fill a"
## [11] "space cruiser. The sensors are placed. I'm going back."
```

Splitting Text

```r
# splitting a string into single characters
strsplit(sw[64], "")
```

```
## [[1]]
##  [1] "L" "U" "K" "E" ":" " " "(" "i" "n" "t" "o" " " "c" "o" "m" "l" "i" "n"
## [19] "k" ")" " " "E" "c" "h" "o" " " "T" "h" "r" "e" "e" " " "t" "o" " " "E"
## [37] "c" "h" "o" " " "S" "e" "v" "e" "n" "." " " "H" "a" "n" "," " " "o" "l"
## [55] "d" " " "b" "u" "d" "d" "y" "," " " "d" "o" " " "y" "o" "u"
```

Parsing Scripts

Dialogues

- Extracting the dialogues
- Identifying Star Wars characters (Luke, Han)
- Ignoring descriptions or non-dialogue remarks
  - e.g. `After a little static a familiar voice is heard`
- Ignoring annotations:
  - e.g. `(over comlink)`
  - e.g. `(into comlink)`

Star Wars Episode V script

```
HAN: Chewie!

           The Wookiee grumbles a reply.

HAN: All right, don't lose your temper. I'll come right back and give
you a hand.

           Chewbacca puts his mask back on and returns to his welding
        as Han leaves.
```

Star Wars Episode V script

```
BEN: If you choose to face Vader, you will do it alone. I cannot
interfere.

LUKE: I understand. (he moves to his X-wing) Artoo, fire up the
converters.

           Artoo whistles a happy reply.

BEN: Luke, don't give in to hate -- that leads to the dark side.

           Luke nods and climbs into his ship.

YODA: Strong is Vader. Mind what you have learned. Save you it can.

LUKE: I will. And I'll return. I promise.
```

Text Analysis

Dialogues

- Identifying words
- Counting frequencies of words
- Common words: prepositions, articles, conjunctions
- Exclamation symbols, numbers,

LUKE

LEIA

THREEPIO

HAN

VADER

YODA

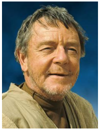
BEN

EMPEROR

LANDO

JABBA

ACKBAR


OWEN

PIETT

TARKIN

TROOPER

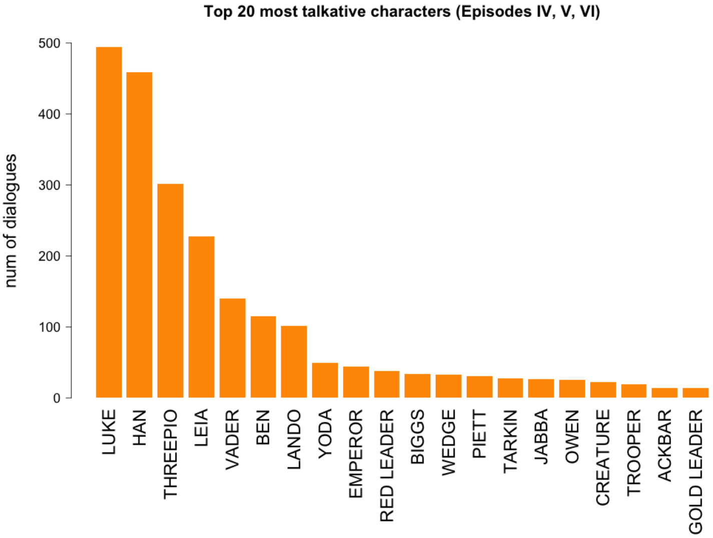
GOLD LEADER

RED LEADER

BIGGS

WEDGE

Top Characters



Excluded Characters