Command line lab 2

Creating a directory

- Create a directory for the PDFs we are going to experiment on
- `mkdir pdfs`
- `cd pdfs`

Downloading a file from the internet

- The curl command lets you fetch a given URL or file
- "client for URLs"
- To download the Mueller Report PDF from the DOJ website, type this on the command line and press return:

```
curl -o mueller.pdf https://www.justice.gov/archives/sco/file/1373816/dl
```

Installing software

Note: the exact commands for installing software differ somewhat across operating systems

1. Type on the command line and press return: `sudo apt-get update`
2. `sudo apt-get install poppler-utils`
3. Answer `Y` to the question:

```
After this operation, 17.2 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```

Understanding software installs: `sudo`

- Stands for "superuser do"
- Runs the commands that follow as superuser
- Superuser is a privileged administrative account
- Software installs typically require a privileged account

Understanding software installs: `apt-get`

- Software installer
- Accepts subcommands such as:
  - `update` update the package lists for available software packages
  - `install` install or upgrade packages

What is `poppler-utils`?

- Name of the software package we are installing
- A set of command-line programs for processing PDFs
- We will initially focus on two:
  - `pdfinfo`
  - `pdftotext`
- Description of all

`pdfinfo`

- Shows a PDF's metadata
- From your `pdfs` subdirectory, enter the following at the command line and press return:

```
pdfinfo mueller.pdf
```

`pdfinfo` output:

```
Title:          Report on the Investigation into Russian Interference in the 2016 Presidential Election
Subject:        Investigation into Russian Interference in the 2016 Presidential Election
Keywords:       2016 Presidential Election; Special Counsel; U.S. Department of Justice; Robert S. Mueller;
Author:         Special Counsel’s Office
Creator:        Adobe Acrobat Pro DC 19.12.20036
Producer:       Adobe Acrobat Pro DC 19.12.20036
CreationDate:   Thu Aug 29 17:13:28 2019 UTC
ModDate:        Tue Sep  3 17:22:42 2019 UTC
Tagged:         yes
UserProperties: no
Suspects:       no
Form:           AcroForm
JavaScript:     no
Pages:          448
Encrypted:      no
Page size:      612 x 792 pts (letter)
Page rot:       0
File size:      11446227 bytes
Optimized:      no
PDF version:    1.7
```

`pdftotext`

- Extracts text from a PDF
- From your pdfs subdirectory, enter the following at the command line and press return:

  ```
  pdftotext mueller.pdf
  ```
- Try `ls`

Exercises

1. Create a new directory at the same directory level as `pdfs` called `txt`. Move `mueller.txt` to your newly created `txt` subdirectory.
2. How can we ensure output generated from `pdftotext` goes to the `txt` directory in the future?
3. Test some flags on both `pdfinfo` and `pdftotext`. Find anything interesting?
4. Load another PDF into codespace - something that interests you. Run both `pdfinfo` and `pdftotext`. How's the text quality?