text mining for poets

DTL SU @ AU

kristoffer l nielbo
kln@cas.au.dk
github.com/digitaltxtlab

DAI|IMC|AARHUS UNIVERSITY

text-mining for poets

- text is available like never before
  - dictionaries, corpora, full-text DBs ...
  - information collection efforts [1]
  - `Information Super Highway Roadkill`, emails, blogs, websites ...
  - *billions & billions of words*
- what can we do with it all?
- it is better to do something simple, than nothing at all
- DIY is more satisfying than begging for “help” from a computer officer

[1] British National Corpus, MediaStream, Economic Data Series ...

Pannang Curry 1-2-3!

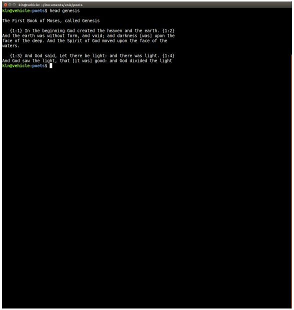
1. Open your right refrigerator door and remove ingredients from the following locations: Door shelf 2, Spot 1; Crisper drawer 1, Spot 3; Crisper drawer 1, Spot 5.
2. Open your third kitchen drawer and remove the utensils labeled “1”, “3”, “4”, “9”, and “12”.
3. Use your arms to apply utensil 1 to ingredients 1-3. Place ingredients inside utensil 3.

*Note: This recipe uses ShaKL the Shared Kitchen Layout. To use ShaKL, you’ll need to have installed ShaKL shelving, cabinetry, and utensils throughout your kitchen and pantry and have basic understanding of ShaKL managers. To learn more, read *Up and Running with ShaKL* (O’Billy Press, 2015). Want to improve ShaKL? Consider contributing to our team.

- a `shell` is a program whose primary purpose is to read commands and run other programs

- the `shell`'s main advantages are its high action-to-keystroke ratio, its support for automating repetitive tasks, and its capacity to access networked machines

- the `shell`'s main disadvantages are its primarily textual nature and how cryptic its commands and operation can be



```
kln@vehicle:poets$ head genesis

The First Book of Moses, called Genesis

   {1:1} In the beginning God created the heaven and the earth. {1:2}
And the earth was without form, and void; and darkness [was] upon the
face of the deep. And the Spirit of God moved upon the face of the
waters.

   {1:3} And God said, Let there be light: and there was light. {1:4}
And God saw the light, that [it was] good: and God divided the light
kln@vehicle:poets$ 
```

`PS1='$ '` sets prompt string in console to

```
$
```

`prompt` indicates that the `shell` is waiting for input

```
$ whoami
kln
$
```

user ID or who the shell thinks you are

`whoami`

1. finds a program called `whoami`
2. runs that program
3. displays that program's output
4. displays a new prompt to tell us that it's ready for more commands

`unknown command`

```
$ somecommand
somecommand: command not found
$
```

- the `shell` runs other programs, so it does not work if the program does not exist

`print working directory` - current default directory

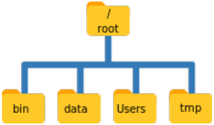
```
$ pwd
home/kln

$ a=$(pwd)
$ echo "current wd is: $a"
current wd is: /home/kln
```

the path to the `home` directory varies between operating systems:

- [linux] `/home/yourname`
- [mac] `/Users/yourname`
- [windows] `C:\Users\yourname`

organization of a filesystem (somewhat unix-centric view) #1

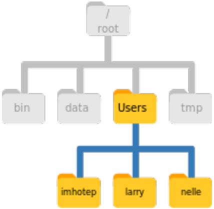


- at the top is the root directory that holds everything else (refer to by a / as the leading slash in /Users/kln

- the root directory contains multiple directories: bin for built-in programs, data for miscellaneous data files, Users for personal directories, tmp for temporary files &c

- so the current working directory /Users/kln is stored in Users which again is stored in root (because of /)

organization of a filesystem #2



- underneath `Users` each user with an account on the computer has a directory

`ls` - listing content of a directory

```
$ ls
Applications Documents    genesis      Movies        Pictures
Desktop      Downloads    Library      Music         Public

$ ls -F
Applications/ Documents/    genesis      Movies/        Pictures/
Desktop/      Downloads/    Library/     Music/         Public/

$ ls -F Desktop
creatures/          molecules/          notes.txt           solar.pdf
data/               north-pacific-gyre/ pizza.cfg           writing/

$ ls --help
```

navigating a filesystem

cd - change directory

```
$ cd Desktop
$ cd data
$ pwd
/Users/nelle/Desktop/data
$ cd ..
$ pwd
/Users/nelle/Desktop
```

shortcuts

```
$ pwd
/Users/nelle/Desktop
$ cd
$ pwd
/Users/nelle
$ cd -
/Users/nelle
$ pwd
/Users/nelle
$ cd ~/Documents/books
$ pwd
/Users/nelle/Documents/books
```

but the `shell` can do more than navigation, it can manipulate (text) data →

words|unstructured

The First Book of Moses, called Genesis

(1:1) In the beginning God created the heaven and the earth. (1:2) And the earth was without form, and void; and darkness [was] upon the face of the deep. And the Spirit of God moved upon the face of the waters.

(1:3) And God said, Let there be light: and there was light. (1:4) And God saw the light, that [it was] good: and God divided the light from the darkness. (1:5) And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day.

(1:6) And God said, Let there be a firmament in the midst of the waters, and let it divide the waters from the waters. (1:7) And God made the firmament, and divided the waters which [were] under the firmament from the waters which [were] above the firmament: and it was so. (1:8) And God called the firmament Heaven. And the evening and the morning were the second day.

(1:9) And God said, Let the waters under the heaven be gathered together unto one place, and let the dry [land] appear: and it was so. (1:10) And God called the dry [land] Earth; and the gathering together of the waters called he Seas: and God saw that [it was] good. (1:11) And God said, Let the earth bring forth grass, the herb yielding seed, [and] the fruit tree yielding fruit after his kind, whose seed [is] in itself, upon the earth: and it was so. (1:12) And the earth brought forth grass, [and] herb yielding seed after his kind, and the tree yielding fruit, whose seed [was] in itself, after his kind: and God saw that [it was] good. (1:13) And the evening and the morning were the third day.

(1:14) And God said, Let there be lights in the firmament of the heaven to divide the day from the night; and let them be for signs, and for seasons, and for days, and years: (1:15) And let them be for lights in the firmament of the heaven to give light upon the earth: and it was so. (1:16) And God made two great lights; the greater light to rule the day, and the lesser light to rule the night: [he made] the stars also. (1:17) And God set them in the firmament of the heaven to give light upon the earth, (1:18) And to rule over the day and over the night, and to divide the light from the darkness: and God saw that [it was] good. (1:19) And the evening and the morning were the fourth day. (1:20) And God said, Let the waters bring forth abundantly the moving creature that hath life, and fowl [that] may fly above the earth in the open firmament of heaven. (1:21) And God created great whales, and every living creature that moveth, which the waters brought forth abundantly, after their kind, and every winged fowl after his kind: and God saw that [it was] good. (1:22) And God blessed them, saying, Be fruitful, and multiply, and fill the waters in the seas, and let fowl multiply in the earth. (1:23) And the evening and the morning were the fifth day.

(1:24) And God said, Let the earth bring forth the living creature after his kind, cattle, and creeping thing, and beast of the earth after his kind: and it was so. (1:25) And God made the beast of the earth after his kind, and cattle after their kind, and every thing that creepeth upon the earth after his kind: and God saw that [it was] good.

(1:26) And God said, Let us make man in our image, after our likeness: and let them have dominion over the fish of the sea, and over the fowl of the air, and over the cattle, and over all the earth, and over every creeping thing that creepeth upon the earth. (1:27) So God created man in his [own] image, in the image of God created he him; male and female created he them. (1:28) And God blessed ...

case-folding

```
$ tr A-Z a-z
In the beginning God created the heaven and the earth
in the beginning god created the heaven and the earth
```

translate from and to file

```
$ tr "God" "Joe" < genesis > jenesis
$ head genesis | tr "God" "Joe"

$ tr "{}" " " < genesis > genesis_nobrace
```

remove (squeeze) repetitions

```
echo  "   In the beginning God created the heaven and the earth" | tr -s [:space:] " "
: In the beginning God created the heaven and the earth
```

remove characters

```
$ echo "{1:1} In the beginning God created the heaven and the earth" | tr -d "0-9"
{:} In the beginning God created the heaven and the earth
```

complement

```
$ echo "{1:1} In the beginning God created the heaven and the earth" | tr -cd "0-9:"
1:1
```

collapse lines

```
$ tr -s "\r\n" " " < genesis > genesis_line
$ tr -s "\n" " " < filename > new_filename
```

man(ual) page for more information

```
$ man tr
```

tokenization - unigrams

```
$ tr -sc "A-Za-z" "\n" < genesis
```

sort in alphabetic order

```
$ tr -sc "A-Za-z" "\n" < genesis | sort
```

uniq - lexicon of document

```
$ tr -sc "A-Za-z" "\n" < genesis | sort | uniq
$ tr -sc "A-Za-z" "\n" < genesis | sort | uniq -c
$ tr -sc "A-Za-z" "\n" < genesis | sort | uniq -c > lexicon.txt
```

file conversion

```
man pdftotext
man catdoc
```

script for char extraction from pdf files in directory

```
#!/bin/bash

for f in *.pdf
do
  echo "converting: - $f"
  pdftotext $f $f.txt
done
```

tired of cryptic commands and operations from the command line?

luckily we have:

python

```
if questions:
    try:
        answer()
    except RunTimeError:
        pass
    else:
        print "thank you"
```