Lecture 13: Bash Shell and Command Line, Grep

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- Finally, shell (bash, zsh)
  - Running things in command line
  - Interacting with text files in command line
  - Regex-based text search using grep

- Follow up of Lecture 11: Data formats, text file encoding & conversion (pages 7 – 11)
  - https://naraehan.github.io/Data-Science-for-Linguists-2025/lecture11.pdf

Bash/Zsh shell

- What is a "shell"?
  - https://en.wikipedia.org/wiki/Shell_(computing)
  - Usually refers to the command-line interface (CLI) as opposed to graphical user interface (GUI).
  - Bash is the most common flavor of shell in Unix-like OS.
- Mac:
  - Mac OS is a Unix-type OS.
  - Terminal is a built-in terminal. Zsh is the default shell, very similar to bash.
- Windows:
  - Not Unix-like OS, so does not come with native bash shell. But we installed "git bash": a bash environment for running command-line git.
  - As a bonus, it came with pretty much all of popular Unix command-line tools!

To find out which shell you're running:
`echo $SHELL`

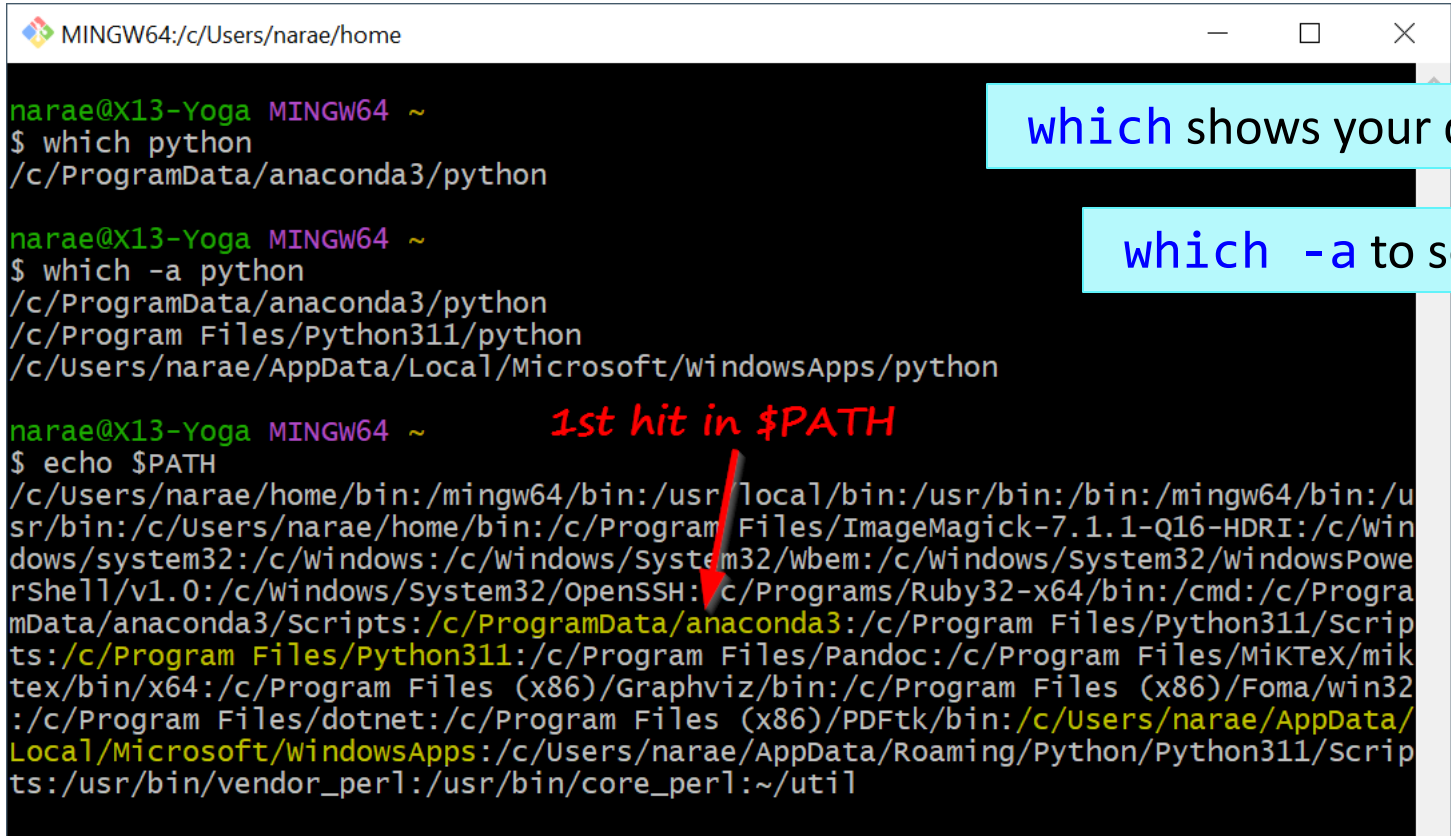
Shell introduction, navigating

- Introducing the shell
  - https://swcarpentry.github.io/shell-novice/01-intro.html
- Navigating & working with files and directories
  - https://swcarpentry.github.io/shell-novice/02-filedir.html
  - https://swcarpentry.github.io/shell-novice/03-create.html

- We've been doing some of these already, as part of our git routine. You should know:
  - `.  ..  ~  /`
  - `pwd`
  - `cd`
  - `ls`
  - Command-line history with ↑ and ↓
  - Using TAB for file name completion
  - Using Control+C to quit

PATH, which, where

- By now, we have multiple versions of Python on our system.
- But which Python is your default? Where is it exactly? 🡐 Depends on your PATH system variable.



`which` shows your default python

`which -a` to see all matches

`echo $PATH` lists all directories where my OS looks for executable commands. My Anaconda Python's location is listed first, so it is my default Python.

Running python script from command line

Ways to run python:

- In Ling1330, we ran .py scripts via IDLE.
- In this class, we are running Jupyter Notebooks.
- ← Surprise! Neither is *the primary way* to run Python.
- ← **Running .py scripts in command line** is.

1. `python hello.py`
   - Assuming python is in your $PATH, and hello.py is in your current working directory
2. `hello.py`
   - Assuming your current working directory is in your $PATH. If not, you should execute `./hello.py`
   - Assuming your script begins with a line (called 'shebang' line):

     `#!/systempath/to/python`
     - In my case, it's `#!/c/ProgramData/Anaconda3/python`
     - If your path contains a SPACE... tough luck! (kidding, there are ways around it.)

Method 1

MINGW64:/c/Users/narae/Documents/Data_Science

```
narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data_Science
$ ls
hello.py

narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data_Science
$ cat hello.py
print('hello, world!')

narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data_Science
$ which python
/c/ProgramData/anaconda3/python

narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data_Science
$ python hello.py
hello, world!

narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data_Science
$ which -a python
/c/ProgramData/anaconda3/python
/c/Program Files/Python311/python
/c/Users/narae/AppData/Local/Microsoft/WindowsApps/python

narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data_Science
$ /c/Program\ Files/Python311/python hello.py
hello, world!

narae@X13-Yoga
$
```

To run with different python program,
I must specify the full path
(note \ to escape space)

Method 2

MINGW64:/c/Users/narae/Documents/Data_Science

```
narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data
$ cat hello.py
#!/c/ProgramData/Anaconda3/python
print('hello, world!')

narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data
$ ./hello.py
hello, world!

narae@X13-Yoga MINGW64 /c/Users/narae/Documents/Data
$
```

Piping and I/O redirection

- **Piping** and **I/O redirection** make command-line ever so powerful.
- For people working mainly with text data (us!), piping enables us to manipulate data on the fly.
  - `hello.py > out.txt` redirect output to file
  - `hello.py | wc` pipe output to another application
  - `hello.py | wc > out.txt` daisy chain!

  Also:
  - `<` read in from a file input
  - `>>` *append* to existing file rather than overwriting

Download two files

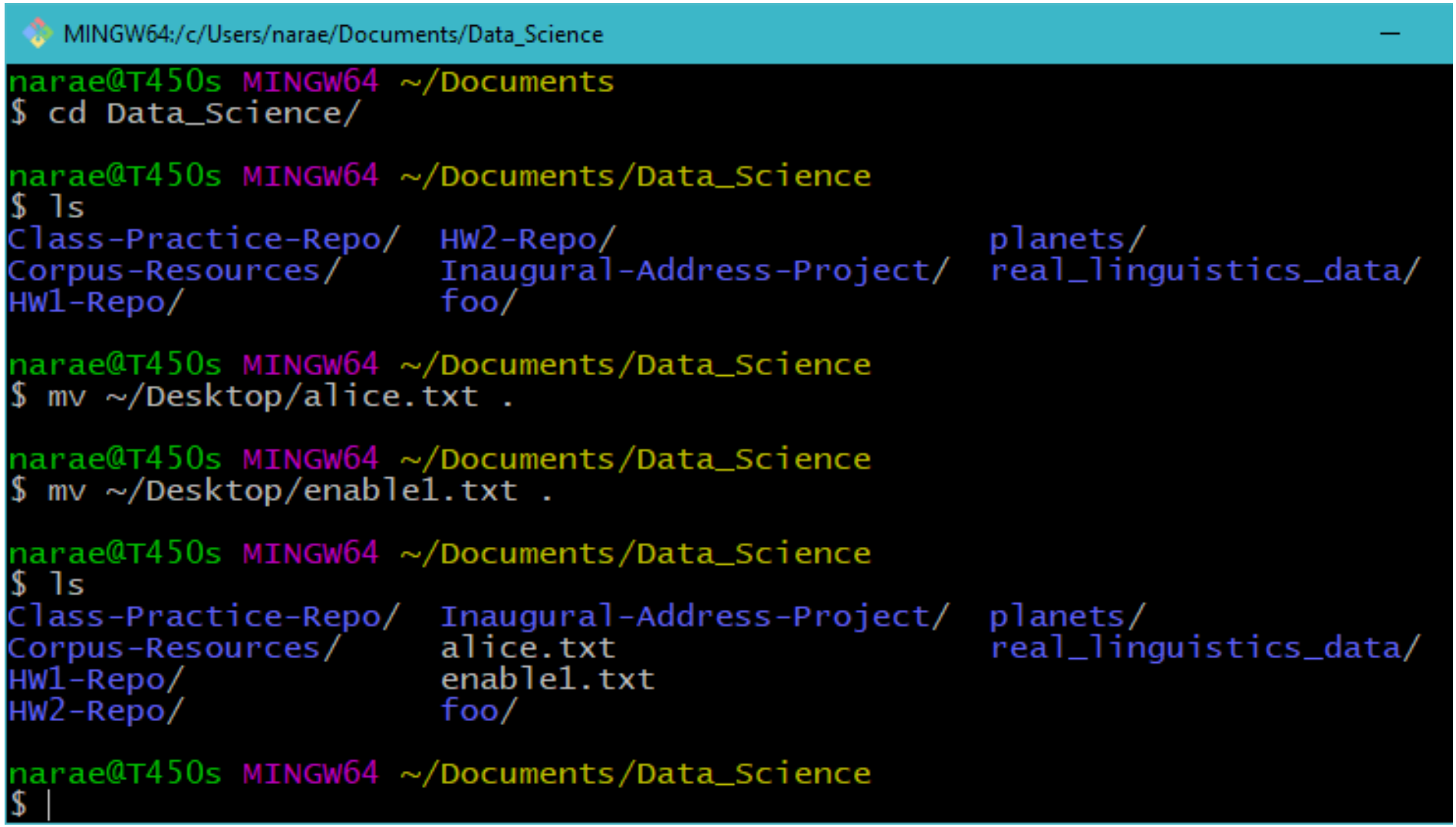
- Alice's Adventures in Wonderland
  - https://www.gutenberg.org/ebooks/11
  - Download the Plain Text UTF-8 version.
  - Rename the file to "alice.txt"
- ENABLE word list from Peter Norvig's site:
  - https://norvig.com/ngrams/
  - Download "enable1.txt"

🡐 Save them onto your Desktop.

🡐 Then, within bash shell, move the files into your Data_Science directory. (Wait if you are not sure how this is done.)

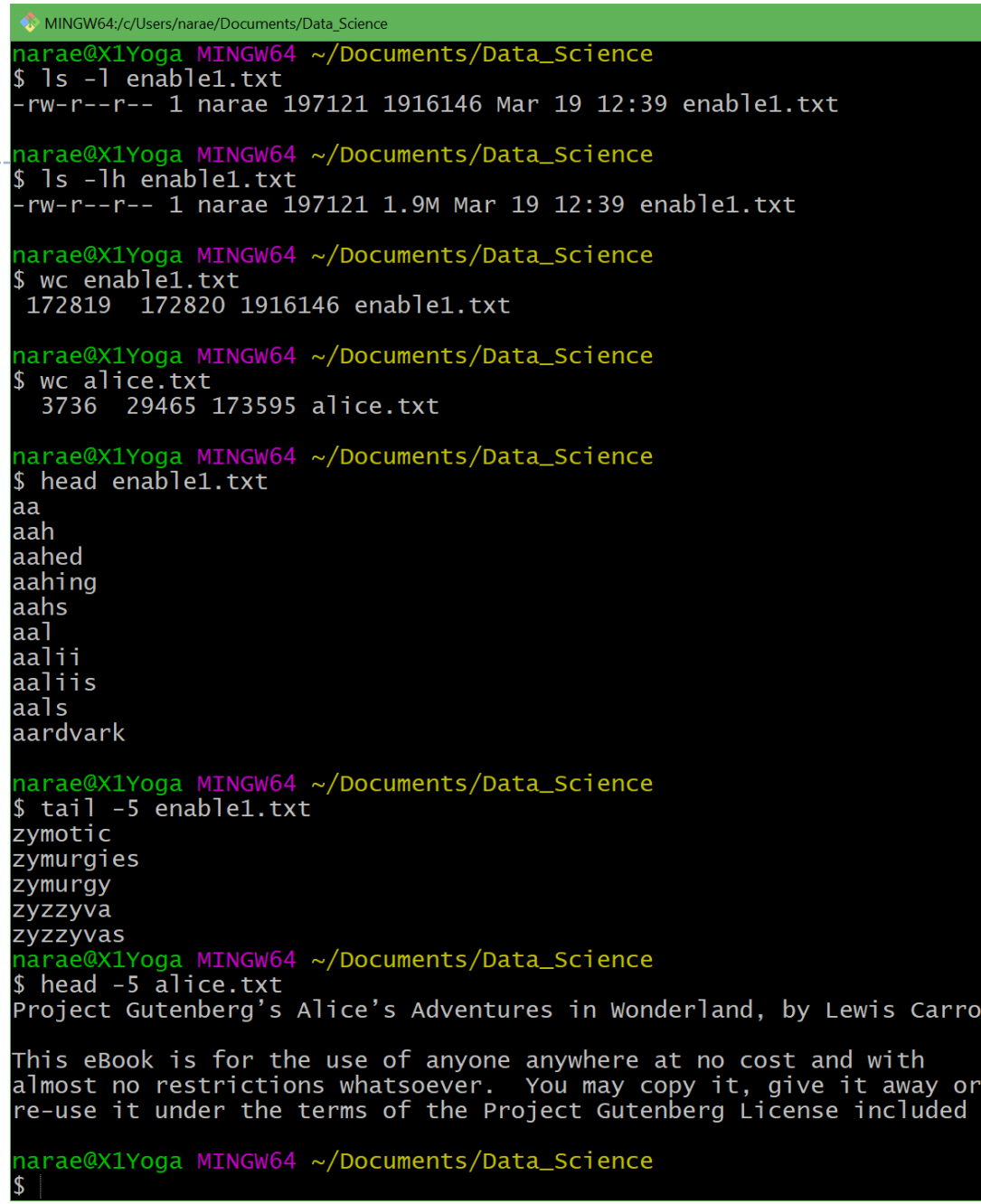
🡐 In command line, find out as much you can about these files.

Files in your Data_Science directory



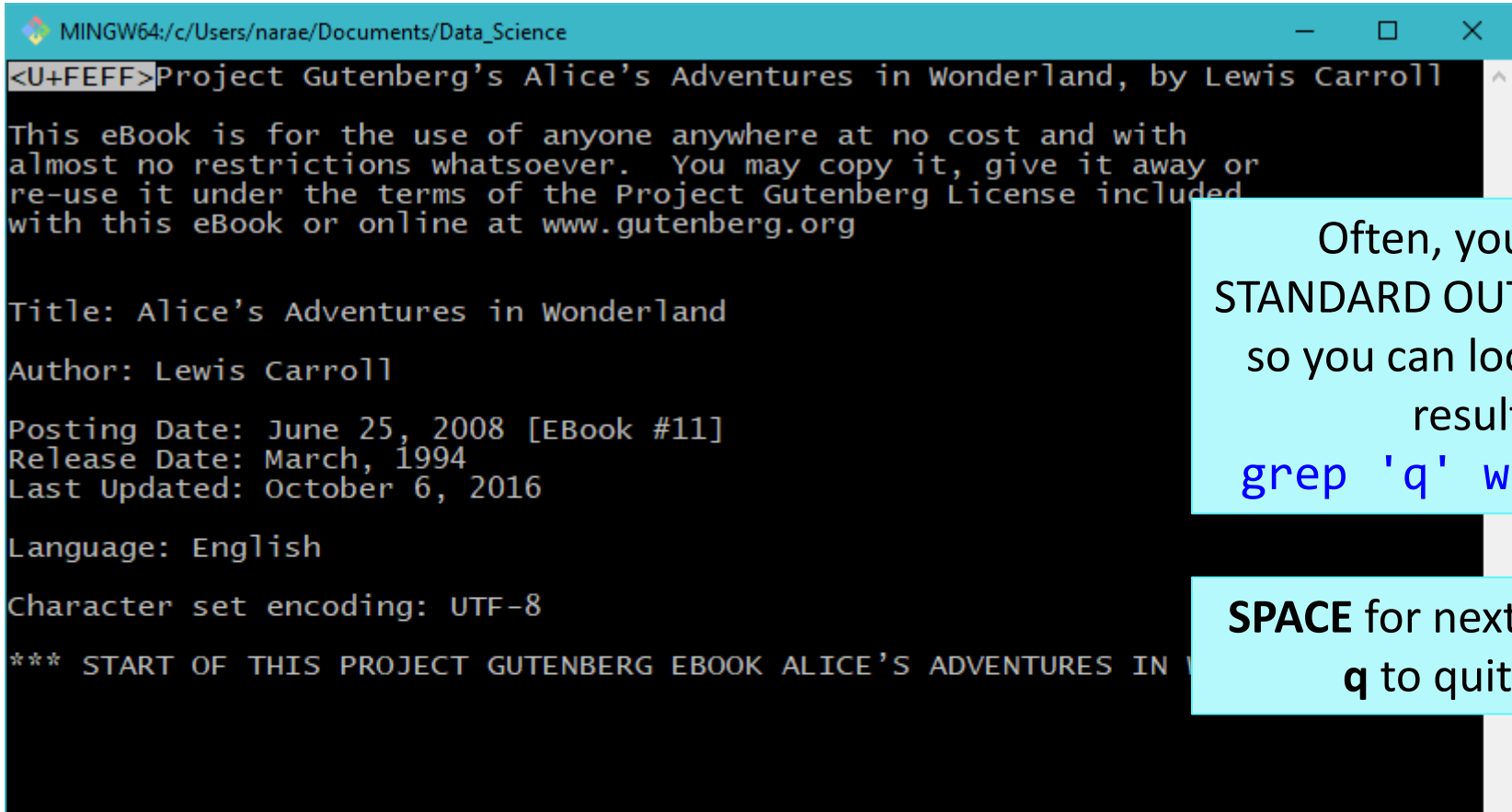
Examining a text file

- `ls (-lahF)`
  - Displays file info
  - Also: -G (Mac OS)
- `file (-i)`
  - Displays character encoding, line ending
- `wc`
  - Displays line count, word count, and character count
- `head -n`
  - Displays initial n lines
- `tail -n`
  - Displays last n lines



more or less

- more (and less) through a text file content, one screen-full at a time. Press **SPACE** for next page, **q** to quit.
  - Windows users: only less is available on git bash.



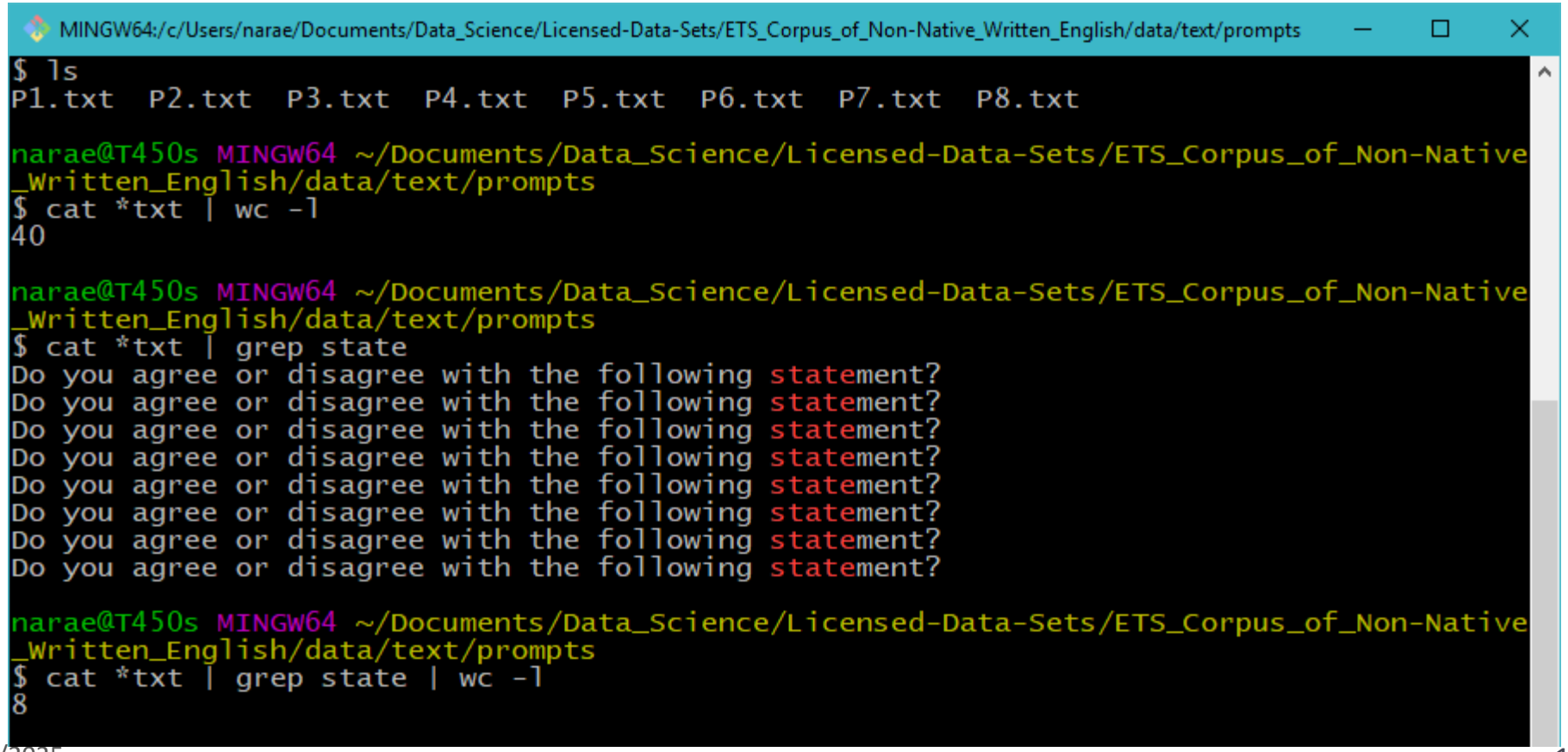
Often, you **pipe** your STANDARD OUTPUT into more, so you can look through the result, e.g.,

```
grep 'q' words | less
```

SPACE for next page
q to quit

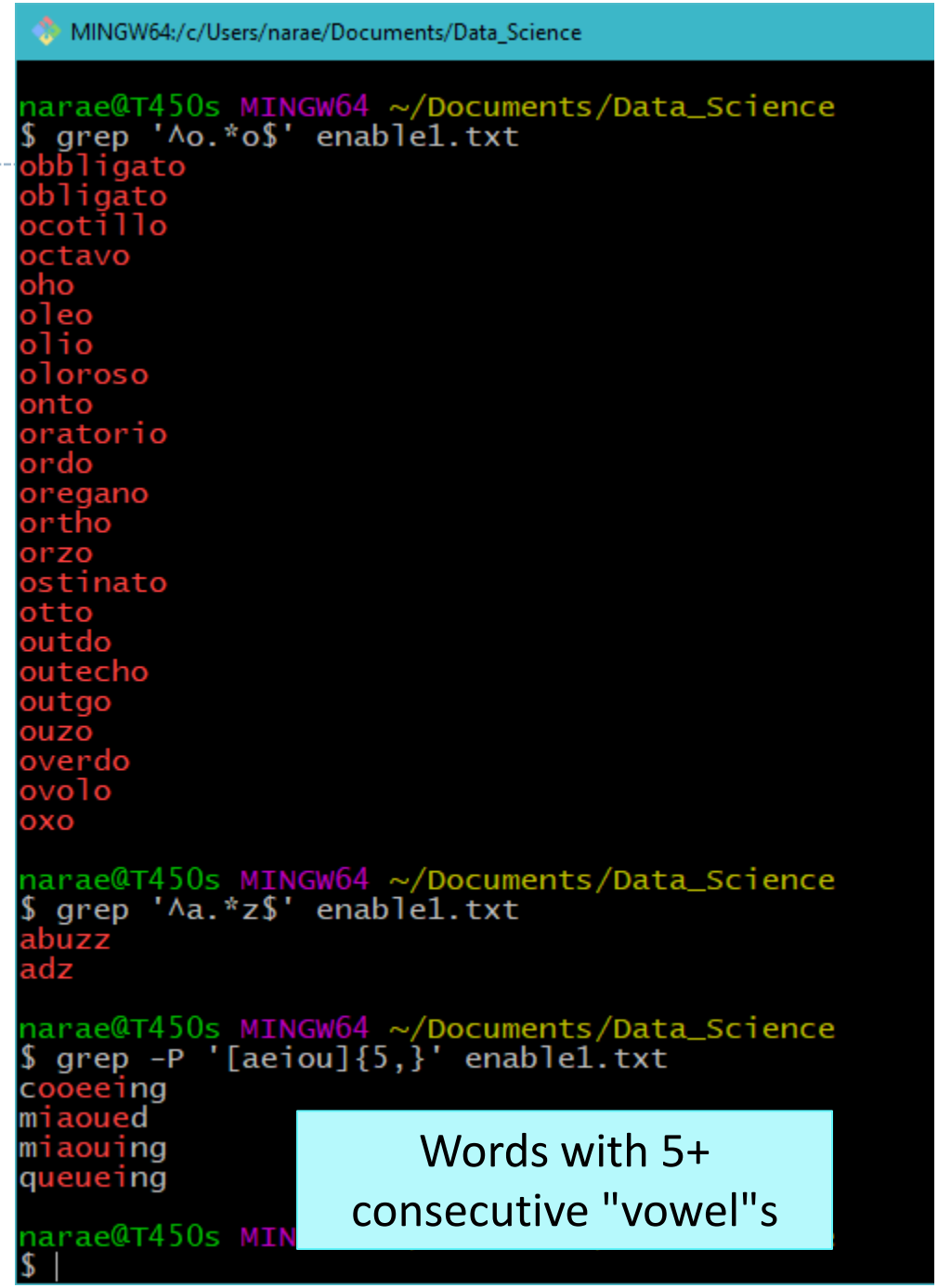
cat

- `cat` *concatenates* text file content and prints on the standard output.
  - Often used as the first step of piping.
  - Also useful in concatenating multiple file contents.



grep!!!

- `grep`
  - Searches each line in text for regular expression match
  - Excellent intro: https://softpanorama.org/Tools/grep.shtml
- `grep -P`
  - Already on git-Bash & Linux
    - **Mac users**: use `egrep` or `grep -E`
  - Accepts **perl-style** regular expressions
  - Perl-style = Python-style! Can use \s, \d etc.

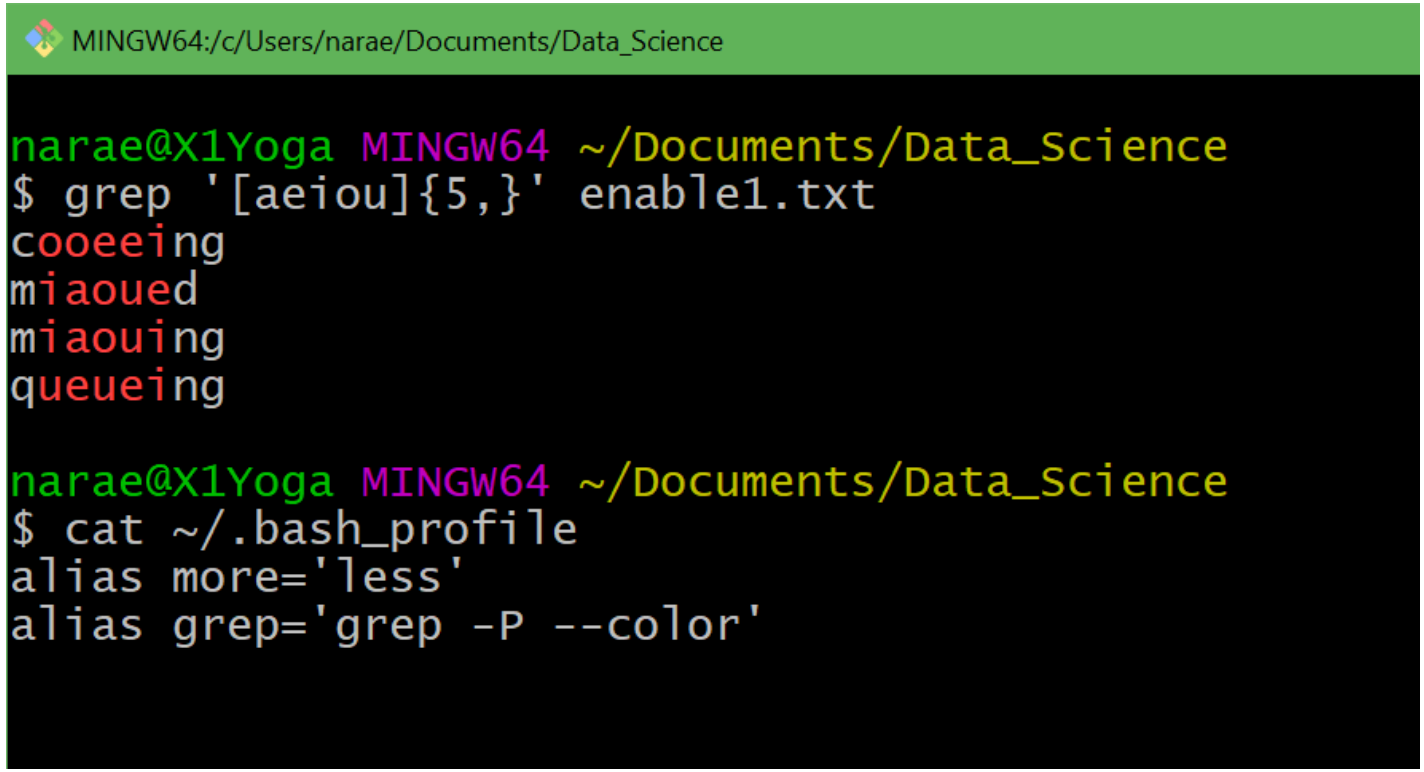


grep is better in color

- You might want to colorize your grep output: `grep --color`
- I have `grep` aliased to use color & perl-style regex in my `.bash_profile` configuration file:

Mac users: your config file is .zprofile

Mac users: you will want to alias `egrep` or `grep -E`



grep and piping, together

```
MINGW64:/c/Users/narae/Documents/Data_Science
unwarrantable
unwatchable
unwearable
unwinnable
unworkable

narae@T450s MINGW64 ~/Documents/Data_Science
$ grep '^un.*able$' enable1.txt  | wc -l
213

narae@T450s MINGW64 ~/Documents/Data_Science
$ grep '^un.*able$' enable1.txt  > able.txt

narae@T450s MINGW64 ~/Documents/Data_Science
$ tail -5 able.txt
unwarrantable
unwatchable
unwearable
unwinnable
unworkable

narae@T450s MINGW64 ~/Documents/Data_Science
$ grep '^in.*able$' enable1.txt  >> able.txt

narae@T450s MINGW64 ~/Documents/Data_Science
$ tail -5 able.txt
invariable
investable
inviable
inviolable
invulnerable

narae@T450s MINGW64 ~/Documents/Data_Science
$ wc -l able.txt
316 able.txt

narae@T450s MINGW64 ~/Documents/Data_Science
$
```

Pipe into `wc -l` to count

Write out to a file

Take a look at the last 5 lines of file

Append new search result to file

Take a look at the last 5 lines of file

File is now longer

grep -i, -v

- `grep -i`
  - ignores case
- `grep -v`
  - prints lines that DO NOT match

```
narae@T450s MINGW64 ~/Documents/Data_Science
$ grep -i 'q' enable1.txt | grep -v 'u'
faqir
faqirs
qaid
qaids
qanat
qanats
qat
qats
qindar
qindarka
qindars
qintar
qintars
qoph
qophs
qwerty
qwertys
sheqalim
sheqel
tranq
tranqs

narae@T450s MINGW
$
```

MINGW64:/c/Users/narae/Documents/Data_Science

```
narae@T450s MINGW64 ~/Documents/Data_Science
$ cat enable1.txt | grep -Pv '[aeiouy]'
brr
brrr
crwth
crwths
cwm
cwms
hm
hmm
mm
nth
pfft
phpht
pht
psst
sh
shh
tsk
tsks
tsktsk
tsktsks
```

grep -C n

- grep -C 2
  - prints context: 2 lines before and after

← capital C!

```
MINGW64:/c/Users/narae/Documents/Data_Science

narae@X1Yoga MINGW64 ~/Documents/Data_Science
$ grep -ic 2 "curious" alice.txt
  *    *    *    *    *    *    *

‘What a curious feeling!’ said Alice; ‘I must be shutting up like a
telescope.’

--
her eyes; and once she remembered trying to box her own ears for having
cheated herself in a game of croquet she was playing against herself,
for this curious child was very fond of pretending to be two people.
‘But it’s no use now,’ thought poor Alice, ‘to pretend to be two people!
why, there’s hardly enough of me left to make ONE respectable person!’
--
CHAPTER II. The Pool of Tears

‘Curiouser and curiouser!’ cried Alice (she was so much surprised, that
for the moment she quite forgot how to speak good English); ‘now I’m
opening out like the largest telescope that ever was! Good-bye, feet!’
--
It was high time to go, for the pool was getting quite crowded with the
birds and animals that had fallen into it: there were a Duck and a Dodo,
a Lory and an Eaglet, and several other curious creatures. Alice led the
way, and the whole party swam to the shore.

--
always growing larger and smaller, and being ordered about by mice and
rabbits. I almost wish I hadn’t gone down that rabbit-hole--and yet--and
yet--it’s rather curious, you know, this sort of life! I do wonder what
CAN have happened to me! When I used to read fairy-tales, I fancied that
kind of thing never happened, and now here I am in the middle of one!
--
by another footman in livery, with a round face, and large eyes like a
frog; and both footmen, Alice noticed, had powdered hair that curled all
over their heads. She felt very curious to know what it was all about,
and crept a little way out of the wood to listen.

--
```

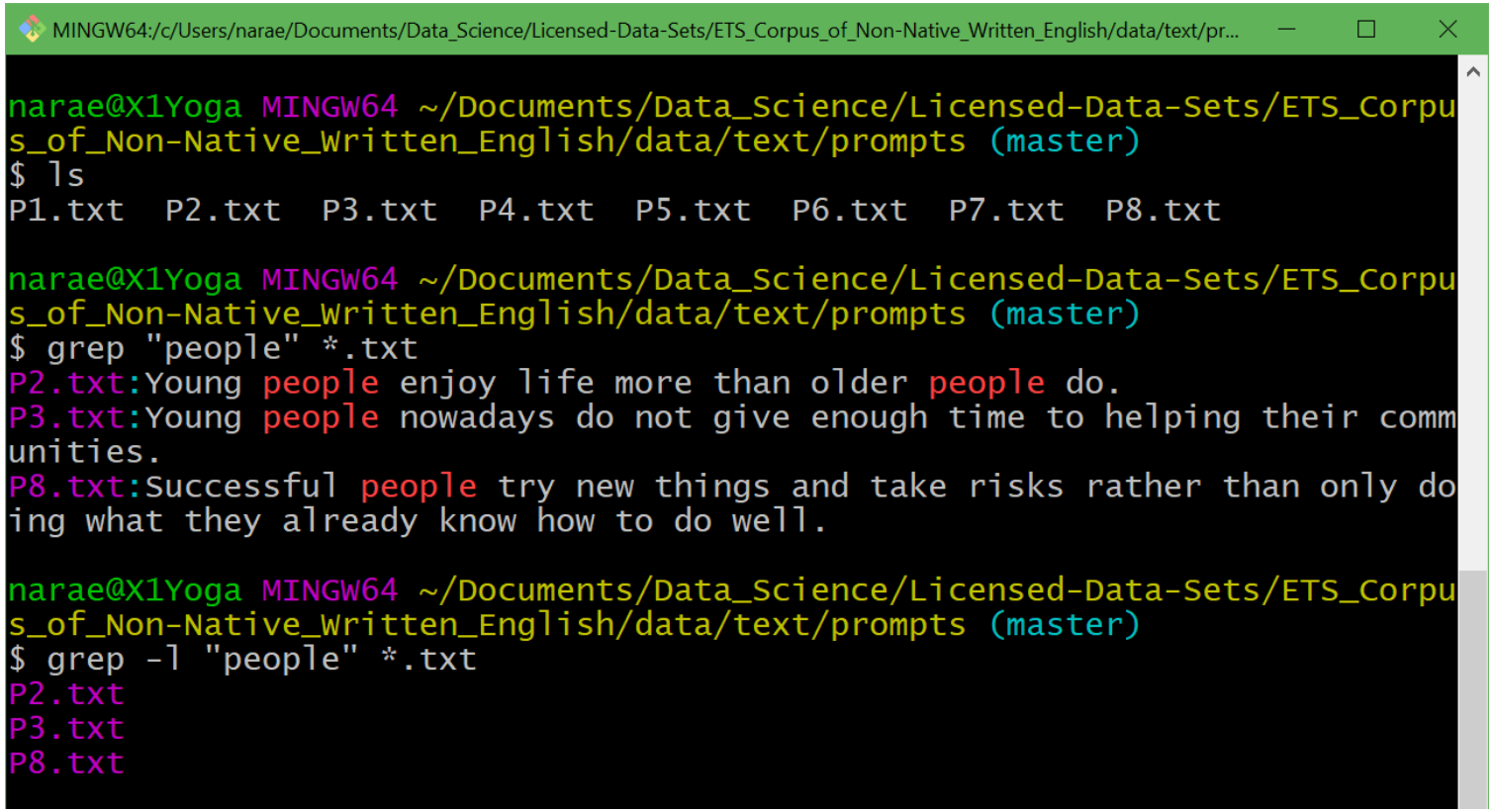
grep -n

- `grep -n`
  - prints out line number

```
MINGW64:/c/Users/narae/Documents/Data_Science

narae@T450s MINGW64 ~/Documents/Data_Science
$ head -25 alice.txt
Project Gutenberg's Alice's Adventures in Wonderland, by Lew

This eBook is for the use of anyone anywhere at no cost and
almost no restrictions whatsoever.  You may copy it, give it
re-use it under the terms of the Project Gutenberg License i
with this eBook or online at www.gutenberg.org


Title: Alice's Adventures in Wonderland

Author: Lewis Carroll

Posting Date: June 25, 2008 [EBook #11]
Release Date: March, 1994
Last Updated: October 6, 2016

Language: English

Character set encoding: UTF-8

*** START OF THIS PROJECT GUTENBERG EBOOK ALICE'S ADVENTURES




narae@T450s MINGW64 ~/Documents/Data_Science
$ grep "\*\*\*" -n alice.txt
21:*** START OF THIS PROJECT GUTENBERG EBOOK ALICE'S ADVENTU
*
3378:*** END OF THIS PROJECT GUTENBERG EBOOK ALICE'S ADVENTU
*
3380:***** This file should be named 11-0.txt or 11-0.zip **
3408:*** START: FULL LICENSE ***

narae@T450s MINGW64 ~/Documents/Data_Science
```

Searching multiple files

- `grep *.txt`
  - Searches through all files ending in .txt
- `grep -l`
  - prints file names *only if* a match is found



“informations”?

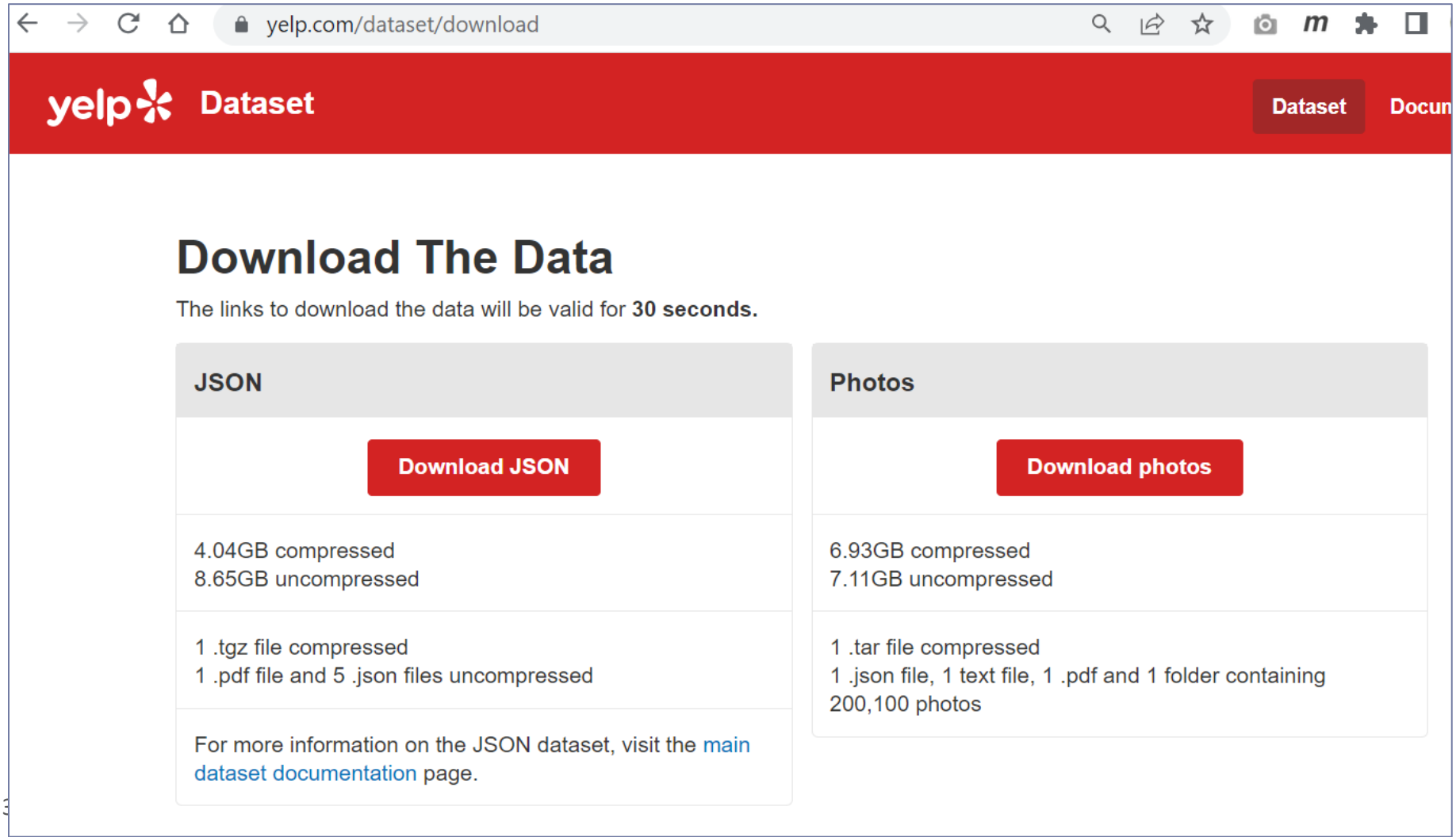
MINGW64:/d/Teaching/2022a.DS4Linguists/GitHub_repos/Licensed-Datasets/ETS_Corpus_of_Non-Native_Written_English/data/text/responses/original

```
e to be create a good product or a good service. So i divide in two grops. The relationship beetwin thee tw
o group is important to make dinamic the singolar group. The informations and the experience have to go in
two directions: from the administration and from the production, only with this continuous relationships is
 possible give a good response to the society. The broad knowledge of many academic subject is the point of
 start for the dinamism
98579.txt:2:    First, we can get new knowledge when we challenge new things. If people just do what they a
lready know, they will be bored and lazy to do because their work is same every day. However, if they focus
 on new things and try to do, they can get new informations; their view become wide, in addition, they can
make good and new relationship with other people and companys. It makes people success.
997714.txt:5:2.When you are looking for a job you can choose from a much larger variety of offers if you ha
ve an idea of more than one subject. Given the example you are a manager and you know many subjects very we
ll. Then you can draw conclusions, that could help you solving problems at work, with informations from a d
ifferent subject. Cross-subject thinking is an important qualification for higher management positions. Not
 only because of your intelligence but aswell because you can do smalltalk, which becomes more and more imp
ortant in business relations.
998126.txt:11:  to put in a nutshell, the student and te adult when they are not student  spend time in und
estand ideas and concepts and learn facts.On the whole, I am firmly convinced that all the time in your lif
e you learn and trie to understand concepts or informations.
998126.txt:12:Owing to  world change evryday at evrytime. you have to focus on all informations if you want
 to do not be drop out the society.

Jane Eyre@T480s MINGW64 /d/Teaching/2022a.DS4Linguists/GitHub_repos/Licensed-Datasets/ETS_Corpus_of_Non-Nat
ive_Written_English/data/text/responses/original (main)
$ grep -n 'informations' *.txt  | wc -l
252
```

Yep, a whole lot of them…

Bring on Big Data! The Yelp Open Dataset

- https://www.yelp.com/dataset



Working with big data files

MINGW64:/d/Corpora/Yelp_dataset_2023/archive

```
narae@T480s MINGW64 /d/Corpora/Yelp_dataset_2023/archive
$ ls -lah
total 8.7G
drwxr-xr-x 1 narae 197121    0 Mar 21 15:33 ./
drwxr-xr-x 1 narae 197121    0 Mar 21 15:37 ../
-rw-r--r-- 1 narae 197121  79K Mar 21 15:32 Dataset_User_Agreement.pdf
-rw-r--r-- 1 narae 197121 114M Mar 21 15:32 yelp_academic_dataset_business.json
-rw-r--r-- 1 narae 197121 274M Mar 21 15:32 yelp_academic_dataset_checkin.json
-rw-r--r-- 1 narae 197121 5.0G Mar 21 15:33 yelp_academic_dataset_review.json
-rw-r--r-- 1 narae 197121 173M Mar 21 15:33 yelp_academic_dataset_tip.json
-rw-r--r-- 1 narae 197121 3.2G Mar 21 15:34 yelp_academic_dataset_user.json

narae@T480s MINGW64 /d/Corpora/Yelp_dataset_2023/archive
$ wc -l yelp_academic_dataset_review.json
6990280 yelp_academic_dataset_review.json

narae@T480s MINGW64 /d/Corpora/Yelp_dataset_2023/archive
$ wc -l yelp_academic_dataset_user.json
1987897 yelp_academic_dataset_user.json
```

Each file is in JSON format, and they are huge:

- review.json is 5GB with 7 million records (=lines)
- user.json is 3.2GB with 2 million records (=lines)

- These are too big to open in most text editors (Notepad++ couldn't.)
- How to explore them? In command line. head/tail, grep and regular expression-based searching.

➔ To-do #14

Wrapping up

- To-do #14 out
  - Start early! Big data means big time (and space) commitment.
- Next class
  - More command line, grep, bash shell scripting
  - Supercomputing at CRC