

LING 331 - Text Processing for Linguists

Week 1

—

Intro,
Unix, Shell,
Environment, Files

Who are we?

Rob Voigt

`robvoigt@northwestern.edu`

Assistant Professor
of Linguistics
and Computer Science
(by courtesy)

Qingcheng Zeng

`qingchengzeng2027@u.northwestern.edu`

PhD Student
in Linguistics

Who is this class for?

- Linguists, social scientists, humanists
- Little-to-no programming experience
- Applications to research

Goals

- Lots of hands-on practice
- Teach you how to teach yourself



Who is this class *not* for?

- Folks with lots of programming experience
- CS Majors (probably - email me if this is you)
- COMP_SCI 110 is similar in focus (and uses one of the same textbooks) - what's different?
 - CS110 - broad, more CS-y (e.g. debugging and testing)
 - LING331 - narrow focus on applications to text, we will purposefully skip less-relevant stuff



What will we learn?

- Unix Command Line

basic usage, remote access, and tools for text

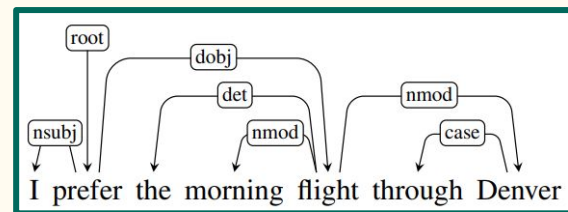
- Basic Python

programming concepts, syntax, useful libraries for text

- Applications (as much as we have time)

data munging, text analysis,
web scraping, APIs

```
[rfj5679@quser21 COHA]$ ls
db_lexicon_coha.zip sources.zip text urls.txt wc
[rfj5679@quser21 COHA]$ cd db
[rfj5679@quser21 db]$ ls
db_1810s_kwp.zip db_1860s_msl.zip db_1910s_aow.zip
db_1820s_lse.zip db_1870s_fhs.zip db_1920s_bsj.zip
db_1830s_sje.zip db_1880s_xjs.zip db_1930s_bkk.zip
db_1840s_ieo.zip db_1890s_lsp.zip db_1940s_jsk.zip
db_1850s_qoe.zip db_1900s_ahs.zip db_1950s_shy.zip
[rfj5679@quser21 db]$ for i in *zip; do unzip $i; done
Archive: db_1810s_kwp.zip
  inflating: 1810.txt
```



When and where will we see each other?

Here! Annenberg G30, MW 9:30am-11:00pm.

Office hours *Rob* Wednesdays 11am-noon / by appt
 Qingcheng Mondays 4-6pm / by appt

Ed discussion board for questions - help each other out!

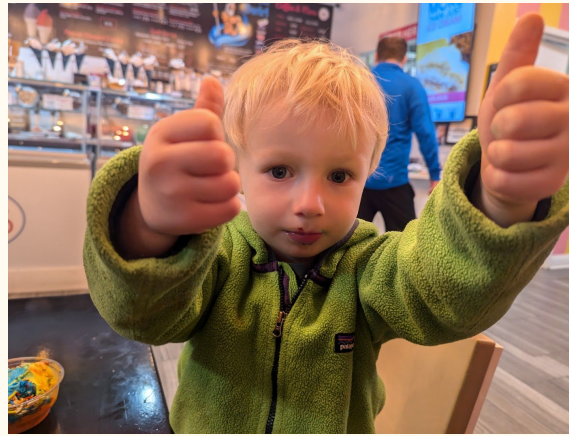
We will do a lot in person! Attendance is important - you could learn a lot of this material by doing online tutorials, the benefit here is us guiding you.

Kid considerations

I have little kids -
childcare loss can happen!

I will prioritize not missing class,
but be aware it's possible

I also try to minimize work time
outside of 9-5 business hours



Josie and Ollie
think you'll do
great in this class!



Why are we doing this?

1. Get computationally “free” -
GUIs only let you do things someone else decided on
2. Processing text data is useful for anyone’s research/work
3. This is the start of computational linguistics!
large language models, conversational/generative AI,
data science, web search, speech-to-text,
“big data” language analysis, etc etc

How will we do it?

Syllabus on course website:

https://faculty.wcas.northwestern.edu/robvoigt/courses/2025_winter/ling331/

Assignments, peer review, final project

Videos/readings before class;

Lectures and discussions

working on assignments during class in small groups

How will we do grading?

Heavy emphasis on qualitative feedback:

Qingcheng and/or I will read your work and comments and provide qualitative feedback inline.

No comment = “good job!”

Letter grades ultimately based on effortful completion,

Midterm and final self-evaluations

The point of this whole thing is for you to learn, period!

What constitutes **strong performance**?

There is a lower bound:

- Do basic reading/watching of course material
- Complete basic assignment (make it work)

There is no upper bound:

- Each week will have extra material listed for reference
- Assignments will often have a number of possible extensions
- You can start working early on your final assignment
- Plus whatever you can dream up

Agreements

I see this class as entering into a set of mutual agreements,
on top of the basic agreements of the university
(academic honesty etc)

We're building a community of learners interested in this topic!
(I'm a learner too.)

By registering, you agree to certain things -
By being the instructor, I agree to certain things.

You agree to:

Invest substantial time and effort in this course this quarter

Hold yourself accountable for your own progress

Be honest in assignments, self-evaluations

Stay on top of your work, and ask for help when needed

Be open to constructive feedback

Challenge yourself

Communicate with me when any of the above falls through

You also agree to our Generative AI Policy

100% banned in every form for work for this class

... until the final project.

You need to learn a new way of thinking!

GenAI will short-circuit this, leaving you unable to understand why things fail when they do.

Final project is more like the “real world” - anything goes!
(just document what you did)

I agree to:

Invest substantial time and effort in your process of learning

Prepare well for class, construct meaningful assignments

Make myself available to help

Be open to criticism and commentary

Provide structures for learning

Communicate with you when any of the above falls through

The Struggle!

Learning programming is like learning a new language

You have to soak in it and use it daily

It will feel unnatural at first, push through

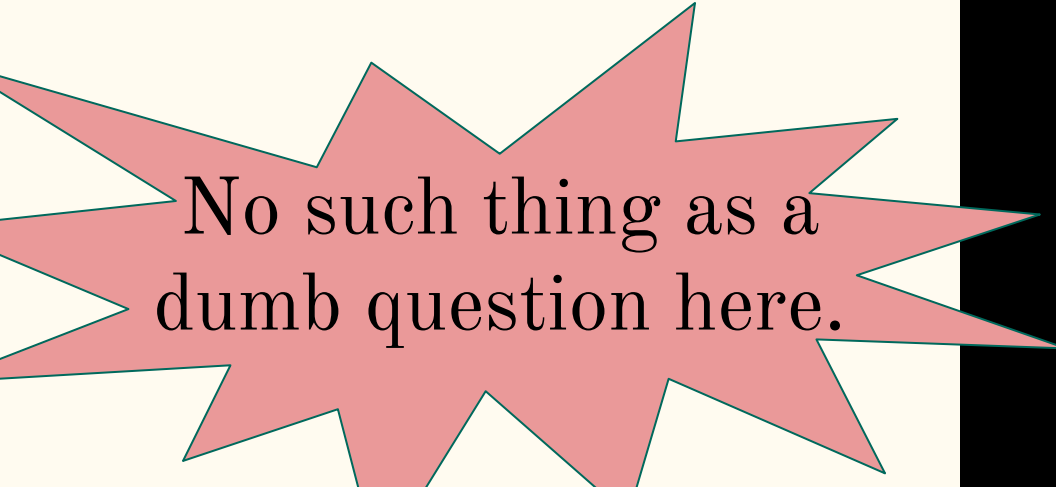
Don't be afraid to play around and break stuff

The Struggle Illustrated



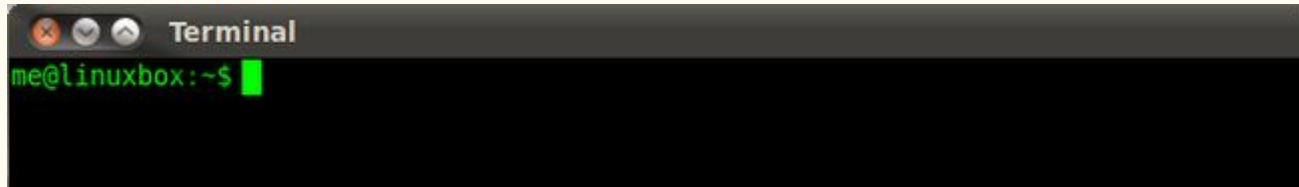
YOU
CAN
DO
IT

ERRORS
ARE
YOUR
NEW
FRIENDS



No such thing as a
dumb question here.

Our new home: the command line



Precision - the challenge of exactitude

One wrong letter, space, or punctuation mark
can easily derail you

These mistakes are at first *very hard to see*

Double-check, triple-check your code
and relevant documentation

(a beloved acronym by programmers is RTFM - read the flippin' manual!)

Take a break and come back to it

Benefits of command line interfaces

Automatable

easy to do
something 1000x

Fast

GUI interfaces are
computationally ‘heavy’

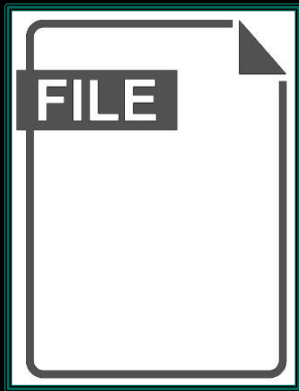
Consistent

same command always
does the same thing

Transparent

you’ll learn what your files
actually are

What is a file?



An abstraction!

... but ultimately,
an array of bytes

e.g., for ASCII text:

<i>Character</i>	L	I	N	G
<i>Bits</i>	100 1100	100 1001	100 1110	100 0111

Types of Files

Text

bytes representing characters
txt, code (like .py), html, logs

Executable

compiled code in binary format
to run as a program

Data

everything else: images, zip files,
doc/ppt/pdf, and so on

**file
extensions
are just a
helpful
suggestion!**

Quest!

Remote computing environment,
cluster of computers running Linux

Common for “big data” and
high-performance tasks

Can schedule complex stuff,
not waste your own machine

Ideal to use Quest
exclusively if you can

If it is slow for you at
home, you can do
everything locally, then
upload assignments

scp assignment1.txt [netid]@quest.it.northwestern.edu /projects/e31086/user/[netid]/assignment1/