RICE SOCIAL SCIENCES
Department of Linguistics

part-of-speech tagging

LING 430 Computational Linguistics Fall 2026
lecture 9a
Oct 20 2026

Siyu Liang
siyu.liang@rice.edu

logistics

- **HW 4 due tonight, 11:59 pm**
- HW 5 to be released Thursday

last time

- logistic regression
  - features & weights
  - the sigmoid
  - loss function
  - gradient descent

agenda

- part-of-speech
- tagsets
- part-of-speech tagging

today's notebook

class09a.ipynb

- course website → week 9 → activity
  - **File → Save a copy in Drive**

part-of-speech

part-of-speech

- which words are nouns in this line from a poem?
  - *'Twas brillig, and the slithy toves did gyre and gimble in the wabe*
    - Lewis Carroll, *Jabberwocky*, 1871
- how do you know?



Lewis Carroll

part-of-speech

- what makes a word a *noun*?
- not its meaning: you have no idea what a *tove* is
- its distribution: where it can go, and what it can take
  - after a determiner and an adjective: *the slithy toves*
  - a plural *-s*: *toves*
  - a possessive: *the tove's wabe*
- a **word class** is a set of words with similar grammatical behavior

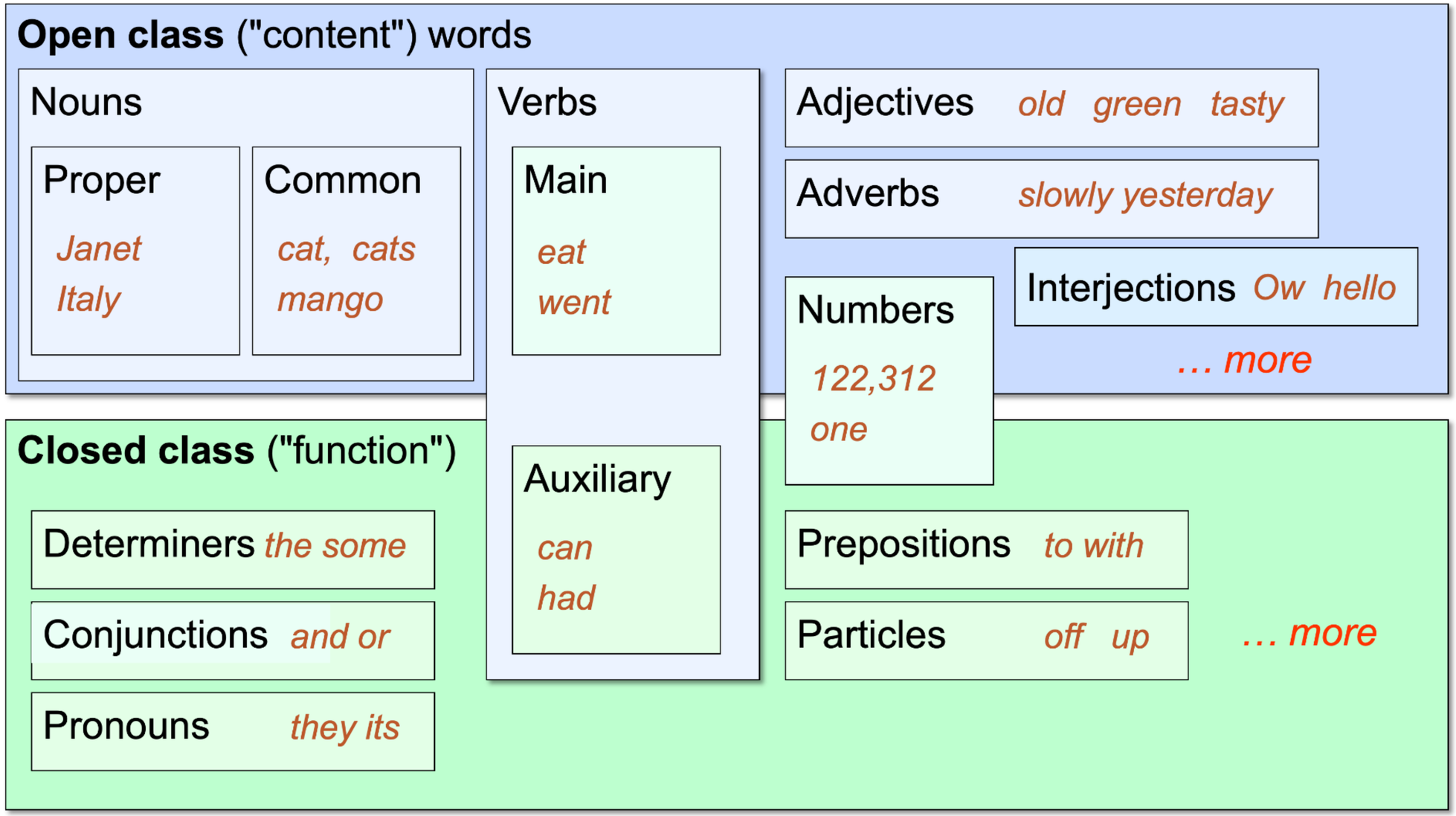
parts of speech

- Dionysius Thrax of Alexandria, around 100 B.C., listed eight
  - noun, verb, pronoun, preposition, adverb, conjunction, participle, article
- the idea is older than linguistics
  - Yāska in *Nirukta*, 5th century B.C.
  - Pāṇini in *Aṣṭādhyāyī*, 5th century B.C.
  - Aristotle, the 4th century B.C.
- still the basis of some linguistic concepts today
- **word classes** are also called **part of speech (POS)** or syntactic categories

open and closed classes

- **open classes** keep taking new members
  - *to google, doomscrolling, sus, delulu...*
  - every language has nouns and verbs
  - not every language has adjectives
    - Korean expresses *beautiful* with a verb
- **closed classes** have a fixed membership: nobody coins a new preposition
  - short and frequent
  - usually **function words**: *the, of, and, to...*

open and closed classes



tagsets

the universal tagset

open classes

- 17 tags shared by the Universal Dependencies treebanks of 100+ languages

tag	class	examples
ADJ	adjective	*red, young, awesome*
ADV	adverb	*very, slowly, home, yesterday*
NOUN	noun	*algorithm, cat, mango, beauty*
VERB	verb	*draw, provide, go*
PROPN	proper noun	*Regina, IBM, Houston*
INTJ	interjection	*oh, um, yes, hello*

the universal tagset

closed classes

tag	class	examples
ADP	adposition	*in, on, by, under*
AUX	auxiliary	*can, may, should, are*
CCONJ	coordinating conjunction	*and, or, but*
DET	determiner	*a, an, the, this*
NUM	numeral	*one, two, 2026*
PART	particle	*’s, not, to* (infinitive)
PRON	pronoun	*she, who, I, others*
SCONJ	subordinating conjunction	*that, which, because*

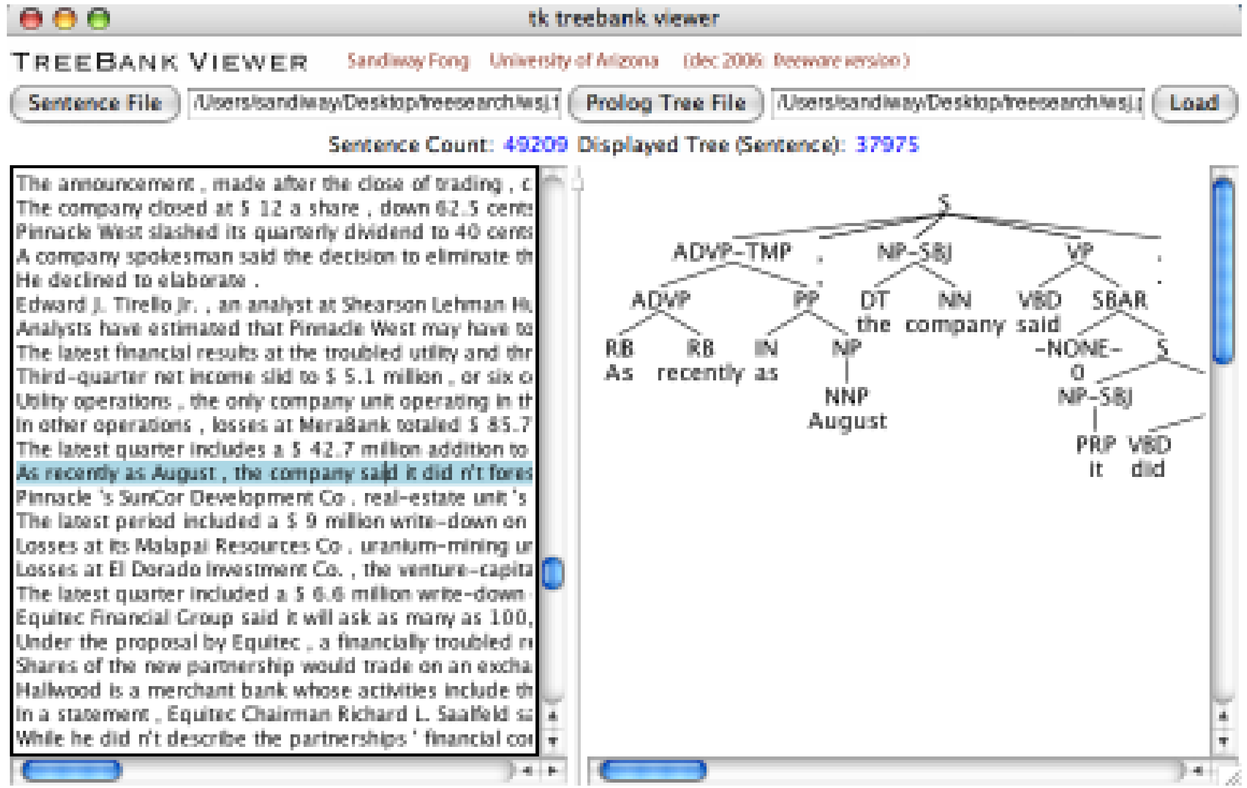
the universal tagset

everything else

tag	class	examples
PUNCT	punctuation	*, . ( )*
SYM	symbol	*$ % □*
X	other	*asdf, qwfg*

Treebank

- we briefly mentioned **Treebanks** in week 4
- POS tagging was considered a fundamental problem in natural language processing
- people have spent a lot of time building treebanks to train and test computers on it



the Penn Treebank

- The Penn Treebank (PTB) project annotated 2,499 stories on the Wall Street Journal (WSJ) from 1987 to 1989
- a finer distinct than the universal tagset
  - nouns: NN, NNS, NNP, NNPS
  - verbs: VB, VBD, VBG, VBN, VBP, VBZ
  - modals: MD; adjectives: JJ, JJR, JJS
  - ...

```
input:   I feel pretty good about it .
output:  I/PRP feel/VBP pretty/RB good/JJ about/IN it/PRP ./.
```

the Penn Treebank tagset

nouns and adjectives

tag	meaning	examples
NN	noun, singular or mass	*machine, air*
NNS	noun, plural	*machines*
NNP	proper noun, singular	*Houston, Janet*
NNPS	proper noun, plural	*Americas*
JJ	adjective	*nice, third*
JJR	adjective, comparative	*nicer*
JJS	adjective, superlative	*nicest*

the Penn Treebank tagset

verbs

tag	meaning	examples
VB	verb, base form	*eat* (to eat, will eat)
VBP	verb, present, not 3rd person singular	*eat* (I eat)
VBZ	verb, present, 3rd person singular	*eats*
VBD	verb, past tense	*ate*
VBG	verb, gerund or present participle	*eating*
VBN	verb, past participle	*eaten*
MD	modal	*can, should, will*

the Penn Treebank tagset

function words

tag	meaning	examples
DT	determiner	*the, a, this, some*
PDT	predeterminer	*all, both, half* (the)
PRP	personal pronoun	*I, him, we*
PRP$	possessive pronoun	*her, our*
POS	possessive ending	*’s*
CC	coordinating conjunction	*and, or, but*
IN	preposition or subordinating conjunction	*in, of, that, if*
TO	to, in any use	*to*
EX	existential there	*there* (is a cat)

the Penn Treebank tagset

adverbs, particles, wh-words

tag	meaning	examples
RB	adverb	*quickly, always*
RBR	adverb, comparative	*further, more*
RBS	adverb, superlative	*best, most*
RP	particle	*up, off* (buy back)
WDT	wh-determiner	*which, that*
WP	wh-pronoun	*who, what*
WP$	possessive wh-pronoun	*whose*
WRB	wh-adverb	*how, where, why*

the Penn Treebank tagset

the rest

tag	meaning	examples
CD	cardinal number	*371, one*
UH	interjection	*oh, um, yes*
FW	foreign word	*habeas, oui*
LS	list item marker	*1., a.*
SYM	symbol	*%, &*

the Penn Treebank tagset

- `nltk` ships with a portion of the Penn Treebank

```
from nltk.corpus import treebank
treebank.tagged_sents()[2170]
```

→ [('I', 'PRP'), ('feel', 'VBP'), ('pretty', 'RB'),
→ ('good', 'JJ'), ('about', 'IN'), ('it', 'PRP'), ('.', '.')]

```
that = [t for w, t in treebank.tagged_words() if w == "that"]
nltk.FreqDist(that).most_common()
```

→ [('IN', 513), ('WDT', 214), ('DT', 77), ('RB', 3)]

part-of-speech tagging

the Brown corpus

- the first million-word electronic corpus produced at Brown University in 1961
- 500 texts of 2,000 words each, published in the US in 1961, in 15 genres
- tagged by hand with 87 tags; nltk maps them to the universal tags

genre	texts	genre	texts
news	44	fiction	29
editorial	27	romance	29
learned	80	humor	9
religion	17	science fiction	6

tagsets compared

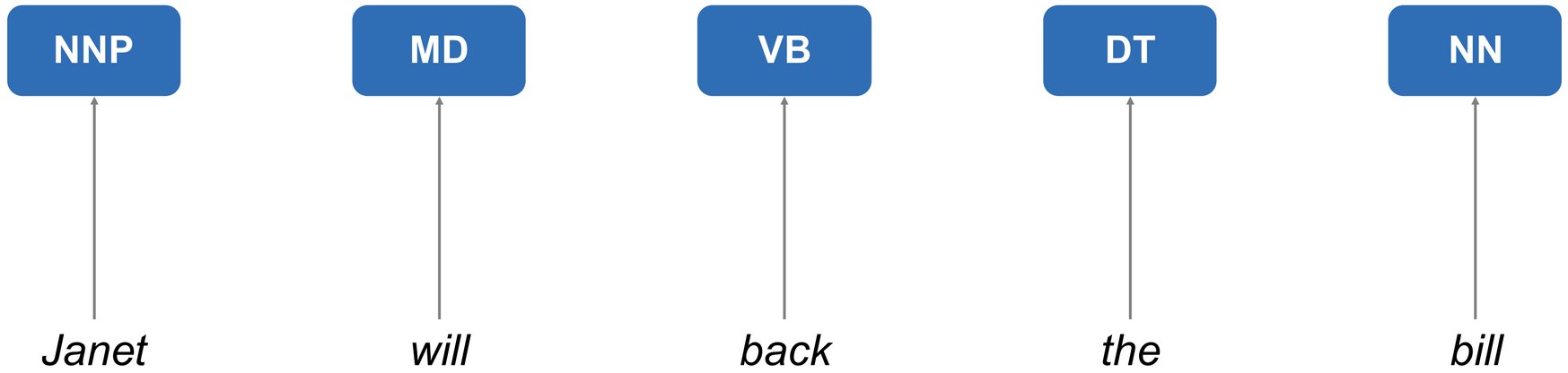
- *I want to go to Dallas*
  - Penn Treebank: *to/TO go/VB to/TO Dallas/NNP*
  - universal: *to/PART go/VERB to/ADP Dallas/PROPN*
- a tagset is backed by a theory of the grammar about which distinctions matter
- a tag only means something relative to its tagset

part-of-speech tagging

part-of-speech tagging

the task

- input: a sequence of words $x_1 \ldots x_n$
- output: one tag per word, $y_1 \ldots y_n$
- a **sequence labeling** task: the output is as long as the input



part-of-speech tagging

- try reading these aloud
  - *The content of the bag was content.*
  - *I object to that object.*
  - *They refuse to permit us to obtain the refuse permit.*

part-of-speech tagging

why tag?

- speech synthesis: *CONtent*/NN vs *conTENT*/JJ, *REfuse*/NN vs *reFUSE*/VB
- lemmatization: *saw*/NOUN → *saw*, *saw*/VERB → *see*
- parsing: the first step toward sentence structure
- corpus linguistics: how often is *like* a verb? 16% of the time in the Brown corpus

part-of-speech tagging

ambiguity

- what do these headlines mean?
  - *British Left Waffles on Falklands*
  - *Teacher Strikes Idle Kids*
  - *Juvenile Court to Try Shooting Defendant*

part-of-speech tagging

ambiguity

- part-of-speech tagging can explain the ambiguity

headline	reading 1	reading 2
British Left Waffles	Left/NN Waffles/VBZ	Left/VBD Waffles/NNS
Teacher Strikes Idle Kids	Strikes/VBZ Idle/JJ	Strikes/NNS Idle/VBP
Try Shooting Defendant	Shooting/VBG Defendant/NN	Shooting/NN Defendant/NN

part-of-speech tagging

ambiguity

- six tags for *back*

tag	example
JJ	*earnings growth took a back seat*
NN	*a small building in the back*
VBP	*a clear majority of senators back the bill*
VB	*Dave began to back toward the door*
RP	*enable the country to buy back debt*
RB	*I was twenty-one back then*

part-of-speech tagging

ambiguity

- guess
  - out of 100 word types (i.e. unique words) in a dictionary, how many have more than one tag?
  - out of 100 running words in a newspaper, how many are ambiguous?

part-of-speech tagging

ambiguity

- what concepts we've talked about this could explain this?

	WSJ	Brown
word types with one tag	86%	85%
word types with two or more tags	14%	15%
tokens that are ambiguous	55%	67%

part-of-speech tagging

ambiguity

- what does the ambiguity look like?

word	tags and counts
that	IN 514 · WDT 217 · DT 114 · RB 3
like	IN 49 · VB 8 · VBP 4 · JJ 1
back	RB 16 · RP 6 · NN 2 · VB 1 · JJ 1
set	VBN 12 · VBD 6 · VB 6 · NN 3 · VBP 2
up	RP 77 · RB 49 · IN 27

part-of-speech tagging

the dummy tagger

- so how do we tag a sentence automatically?
- we could assign the **most frequent tag**
  - give each word the tag it had most often in the training corpus
- a word never seen in training: NOUN
- no context at all
- what accuracy does this get?

part-of-speech tagging

the dummy tagger

- Janet will back the bill, with tag counts from the Brown corpus (universal tags)

word	counts	most frequent tag	correct?
Janet	never seen	NOUN	✓
will	VERB 2,137 · NOUN 108	VERB	✓
back	ADV 734 · NOUN 178 · ADJ 29 · VERB 25	ADV	✗ VERB
the	DET 69,968	DET	✓
bill	NOUN 143	NOUN	✓

part-of-speech tagging

the dummy tagger

- how good is it?
  - about 92% on the Wall Street Journal with Penn Treebank tags
  - 94.6% on the Brown corpus with universal tags
- why so high?
  - most tokens have one tag, and for the rest the most frequent tag is usually right
- this is a **baseline** system
  - any pos-tagger worth building should beat this

part-of-speech tagging

sources of information

- what follows a determiner?
- tag sequences in the Brown corpus, universal tags
- tag sequences are far from random

after	next tag
DET	NOUN .63 · ADJ .24 · VERB .06
ADJ	NOUN .65
PRON	VERB .71
ADP	DET .46 · NOUN .26

part-of-speech tagging

sources of information

- where did the baseline fail? *back*: the word alone says ADV
- what else could a tagger look at?
  - 1. the word itself: its tag counts
  - 2. the neighboring tags: after *will*/MD, a verb
  - 3. the shape of the word: *-ing*, *-ly*, *-ed*, a capital letter
- this will be covered on Thursday!

part-of-speech tagging

the ceiling

- how good can a tagger be?
- about 97% on the Wall Street Journal, for HMM, CRF and neural taggers alike
- two human annotators agree about 97% of the time too
  - part of the last 3% is disagreement about the gold standard, not error
- 97% still means one error every 33 words, about one per sentence

part-of-speech tagging

a tagger in nltk

```
import nltk
nltk.download("averaged_perceptron_tagger_eng")
nltk.pos_tag(["Janet", "will", "back", "the", "bill"])
```

→ [('Janet', 'NNP'), ('will', 'MD'), ('back', 'VB'),

→ ('the', 'DT'), ('bill', 'NN')]

activity 1

part 1: tag by hand

- in pairs, give every word a Penn Treebank tag
  - *Book that flight.*
  - *The old man the boat.*
  - *I will race you to the fence.*
- you can use this file as reference

activity 1

part 2: run the tagger

- in the notebook, run nltk.pos_tag on the same three sentences
- then try with your own sentence
- does the tagger disagree with you?

activity 1

debrief

- the tagger learned from newspaper text, where *book* and *race* are almost always nouns
- a sentence-initial capital looks like a name: *Book*/NNP

word	nltk	should be
Book that flight	NNP	VB
Book *that* flight	WDT	DT
The old *man* the boat	NN	VBP
I will *race* you	NN	VB

next time

- Thursday, Oct 22
  - HMM taggers and Viterbi
  - Jurafsky & Martin, Appendix A, §§A.1–A.4, and Ch. 18, §18.4
- **HW 4 due tonight, 11:59 pm**