CS-5340/6340, Written Assignment #2
DUE: Wednesday, September 21, 2016 by 11:00pm

1. (15 pts) The table below contains frequency values for a set of nouns referring to trees in an imaginary text corpus. Fill in the table below with the unsmoothed probability of each noun, as well as the smoothed frequency and smoothed probability of each noun using add-one smoothing. You should assume that the vocabulary consists only of the nouns listed below.

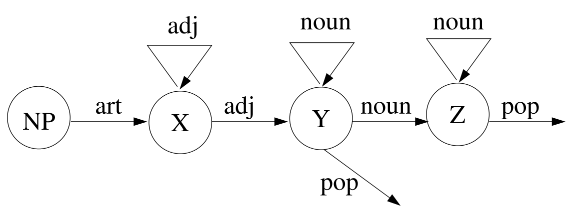
   IMPORTANT: Please show the fraction (numerator/denominator) used to compute each value as well as the final value (e.g., 2/4 = .50).

NOUN	FREQ	UNSMOOTHED PROB	SMOOTHED FREQ	SMOOTHED PROB
maple	600			
oak	400			
pine	180			
spruce	20			
aspen	0			

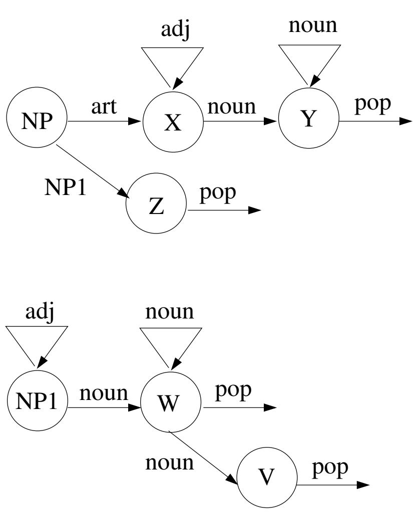
2. (16 pts) Consider the three Noun Phrase (NP) grammars and the three recursive transition networks (RTNs) below:

Grammar A	**Grammar B**	**Grammar C**
NP →art NP1	NP →NP1	NP →NP1
NP1 →adj NP1	NP1 →art NP2	NP1 →art NP2
NP1 →NP2	NP1 →NP2	NP2 →adj NP2
NP2 →noun	NP2 →adj NP2	NP2 →adj NP3
NP2 →noun NP2	NP2 →NP3	NP3 →noun
	NP3 →NP4	NP3 →noun noun
	NP4 →noun NP4	NP3 →noun NP3
	NP4 →noun	

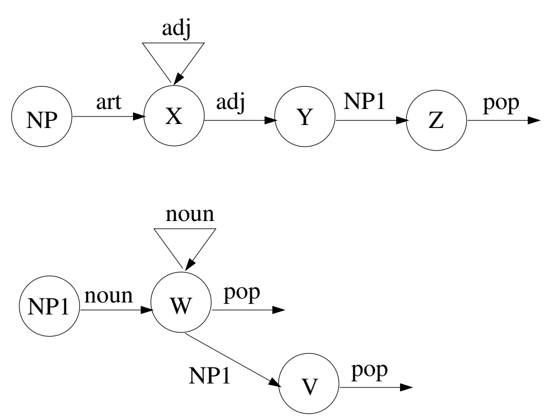
RTN−1



RTN−2



RTN−3



Each grammar and RTN accepts a noun phrase "language" consisting of sequences of part-of-speech (POS) tags that are considered to be legal noun phrases. For example, "adj art noun" might be a POS tag sequence in a noun phrase.

For each pair below, indicate whether they accept exactly the **SAME** NP language or **DIFFERENT** NP languages (i.e., do they accept exactly the same set of POS tag sequences or not). If you answer **DIFFERENT**, then briefly (1 sentence) explain how they are different and give an example of a POS tag sequence that is accepted by one of them but not the other (be sure to say *which* grammar or RTN would accept the example you give).

(a) Grammar A and Grammar B

(b) Grammar A and Grammar C

(c) Grammar A and RTN-2

(d) Grammar A and RTN-3

(e) Grammar B and RTN-2

(f) Grammar C and RTN-1

(g) Grammar C and RTN-3

(h) RTN-1 and RTN-3

3. (24 pts) Consider the following three sentences with assigned part-of-speech tags to be a (tiny!) text corpus. Treat the words as being case-insensitive (so "the" is the same as "The").

A/ART young/ADJ girl/NOUN helped/VERB an/ART old/ADJ woman/NOUN cross/VERB the/ART street/NOUN . The/ART old/ADJ woman/NOUN thanked/VERB the/ART young/ADJ girl/NOUN and/CONJ gave/VERB her/PRO five/NUM dollars/NOUN . The/ART girl/NOUN thanked/VERB the/ART old/ADJ woman/NOUN and/CONJ gave/VERB her/PRO a/ART big/ADJ hug/NOUN .

We define unigram, bigram, trigram, and lexical generation probabilities as:

Lexical Unigram: $P(w_i)$ means probability of word w_i

POS Unigram: $P(t_i)$ means probability of POS tag t_i

Lexical Bigram: $P(w_i \mid w_{i-1})$ means probability of word w_i following word w_{i-1}

POS Bigram: $P(t_i \mid t_{i-1})$ means probability of POS tag t_i following POS tag t_{i-1}

Lexical Trigram: $P(w_i \mid w_{i-2}\ w_{i-1})$ means probability of word w_i following words w_{i-2} w_{i-1}

Lexical Generation Probability: $P(w_i \mid t_i)$ means probability of word w_i given tag t_i.

Compute the probabilities listed below. Please show each probability as a fraction (numerator/denominator)!

(a) $P(the)$

(b) $P(VERB)$

(c) $P(young \mid girl)$

(d) $P(girl \mid young)$

(e) $P(and \mid woman)$

(f) $P(thanked \mid young\ girl)$

(g) $P(five \mid gave\ her)$

(h) $P(the \mid ART)$

(i) $P(cross \mid NOUN)$

(j) $P(thanked \mid VERB)$

(k) $P(NUM \mid PRO)$

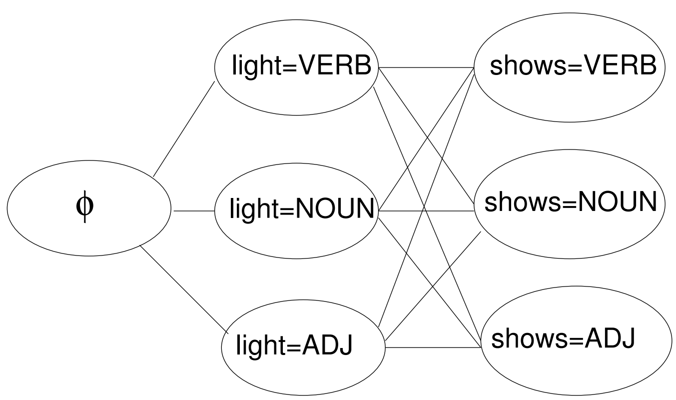
(l) $P(ART \mid VERB)$

4. (15 pts total) Use the following tables of probabilities to answer this question. Note that these numbers are completely fictional and not at all realistic! So don't worry that they don't make sense.

P(light | NOUN)	.70
P(light | VERB)	.50
P(light | ADJ)	.20
P(shows | NOUN)	.40
P(shows | VERB)	.30
P(shows | ADJ)	.10

P(NOUN | ϕ)	.60
P(VERB | ϕ)	.25
P(ADJ | ϕ)	.15
P(NOUN | NOUN)	.80
P(NOUN | VERB)	.30
P(NOUN | ADJ)	.60
P(VERB | NOUN)	.50
P(VERB | VERB)	.40
P(VERB | ADJ)	.10
P(ADJ | NOUN)	.20
P(ADJ | VERB)	.70
P(ADJ | ADJ)	.90

Assume that there are only 3 possible part-of-speech tags: NOUN, VERB, and ADJ. The following network would be used by the Viterbi algorithm to find the most likely sequence of POS tags for the sentence *"Light shows"*:



Using the Viterbi algorithm, compute the probability for each of the following nodes in the network. Show all your work!

(a) P(light=VERB)

(b) P(light=NOUN)

(c) P(light=ADJ)

(d) P(shows=VERB)

(e) P(shows=NOUN)

(f) P(shows=ADJ)

Question #5 is for CS-6340 students ONLY!

5. (15 pts) For this question, use the same Viterbi network and probability tables shown in Question #4. **Leave your answers in fractional form!**

   (a) Compute the lexical tag probability $P(light/VERB \mid light)$, which is the result of normalizing the forward probabilities in the Viterbi network.

   (b) Compute the lexical tag probability $P(light/NOUN \mid light)$, which is the result of normalizing forward probabilities in the Viterbi network.

   (c) Compute the lexical tag probability $P(light/ADJ \mid light)$, which is the result of normalizing forward probabilities in the Viterbi network.

   (d) Compute the lexical tag probability $P(shows/VERB \mid light\ shows)$, which is the result of normalizing the forward probabilities in the Viterbi network.

   (e) Compute the lexical tag probability $P(shows/NOUN \mid light\ shows)$, which is the result of normalizing forward probabilities in the Viterbi network.

   (f) Compute the lexical tag probability $P(shows/ADJ \mid light\ shows)$, which is the result of normalizing forward probabilities in the Viterbi network.

ELECTRONIC SUBMISSION INSTRUCTIONS
(a.k.a. "What to turn in and how to do it")

Your written assignment must be in .pdf format. Please do not turn in .doc or .docx files ... convert them to .pdf format before submitting them!

To submit this assignment, the CADE provides a web-based facility for electronic handin, which can be found here:

https://webhandin.eng.utah.edu/

Or you can log in to any of the CADE machines and issue the command:

handin cs5340 written2 <filename>

Please name your file: YourName-written2.pdf (e.g., EllenRiloff-written2.pdf)

HELPFUL HINT: you can get a listing of the files that you've already turned in via electronic submission by using the 'handin' command without giving it a filename. For example:

handin cs5340 written2

will list all of the files that you've turned in thus far. If you submit a new file with the same name as a previous file, the new file will overwrite the old one.