CS-5340/6340, Written Assignment #1
DUE: Tuesday, September 3, 2019 by 11:59pm
Submit your assignment on CANVAS in pdf format.

1. (15 pts) Answer the questions below based on the part-of-speech tags that should be assigned to each word in the story below.

> *Tom was cycling in Moab when he hit a tree and broke his leg. He was rushed to a hospital where orthopedic doctors put his leg in a cast. He should recover within a month. Tom said that he could have avoided the tree if he had not been distracted by several young kids who were running down the steep trail. He yelled at the boys to move off the biking trail! The children were later scolded by a park ranger. Tom is lucky that he only broke a leg. Crashing into a tree can be fatal!*

   (a) List all of the verbs that appear in a passive voice construction.
   rushed, distracted, scolded

   (b) List all of the modal verbs.
   should, could, can

   (c) List all of the head nouns that are plural.
   doctors, kids, boys, children

   (d) List all of the gerunds.
   Crashing

   (e) List all of the adverbs.
   when, where, not, several, later, only

2. (20 pts) For each underlined word below, indicate whether it is functioning as a *particle* or as a *preposition* in the sentence.

   (a) Joe carried the injured girl down the hill and into an ambulance. **Preposition**

   (b) Tom had to catch up with his coursework after getting the flu. **Particle**

   (c) The 4-hour meeting finally wound up just before noon. **Particle**

   (d) The Jazz pulled off a win after a last-second slam dunk from Mitchell. **Particle**

   (e) The tiny boy was picked on by bullies in his class. **Particle**

   (f) Mary was fed up after the bus was late yet again.**Particle**

   (g) Susan swam across the river to rescue her dog. **Preposition**

   (h) The man dove off a tall cliff into the ocean. **Particle**

   (i) The armed robber gave up after being surrounded by police. **Particle**

   (j) Julie found out that she was nominated for an award. **Particle**

3. (20 pts) Fill in the table with morphology rules to derive all of the words below from the specified root form *in a linguistically sensible way.* Some derivations may require the application of multiple rules. In this case, put each rule in a separate row of the table. Also, some words may have multiple derivations. Be sure to include all derivations that make sense.

   For illustration, the table is already filled in with the derivation of "unfairly" from the root "fair".

   (a) abbreviation (root = "abbreviate")

   (b) eaters (root = "eat")

   (c) forgetfulness (root = "forget")

   (d) nonperishable (root = "perish")

   (e) neonationalism (root = "nation")

Derived Form	**Origin**	**Prefix**	**Suffix**	**Replacement Chars**	**POS of Origin**	**POS of Derived**
unfairly	unfair	-	ly	-	ADJ	ADV
unfair	fair	un	-	-	ADJ	ADJ
abbreviation	abbreviate	-	ion	e	Verb	Noun
eaters	eater	-	s	-	Noun	Noun
eater	eat	-	er	-	Verb	Noun
forgetfulness	forgetful	-	ness	-	Adjective	Noun
forgetful	forget	-	ful	-	Verb	Adjective
nonperishable	perishable	non	-	-	Adjective	Adjective/Noun
perishable	perish	-	able	-	Verb	Adjective/Noun
neonationalism	nationalism	neo	-	-	Noun	Noun
nationalism	national	-	ism	-	Adjective	Noun
National	Nation	-	al	-	Noun	Ajective

4. (25 pts) Assume that a part-of-speech tagger has been applied to the 4 sentences below with the following results:

   A/ART hungry/ADJ bear/NOUN eats/VERB a/ART pound/NOUN of/PREP fish/NOUN per/PREP day/NOUN

   A/ART hungry/ADJ bear/NOUN will/MOD often/ADV hunt/VERB for/PREP food/NOUN in/PREP a/ART garbage/NOUN can/NOUN

   The/ART brown/ADJ bear/NOUN hunt/NOUN starts/VERB tomorrow/ADV

   People/NOUN often/ADV fish/VERB for/PREP trout/NOUN and/CONJ hunt/VERB for/PREP deer/NOUN in/PREP the/ART forest/NOUN

   Fill in the table below with the probabilities that you would estimate based on the sentences above (i.e, treat these 4 sentences like a tiny text corpus). **Please leave your results in fractional form, even if the result is an integer like 0 or 1! For example, leave your answer as 0/5, 5/5, etc.**

   We define unigram, bigram , trigram, and emission 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}

   POS Trigram: $P(t_i \mid t_{i-2}\ t_{i-1})$ means probability of tag t_i following tags t_{i-2} t_{i-1}

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

Probability	**Value**
P(in)	2/40
P(bear)	3/40
P(PREP)	7/40
P(NOUN)	14/40
P(ART $\mid \phi$)	3/6
P(bear $\mid$ hungry)	2/2
P(hungry $\mid$ bear)	0/3
P(NOUN $\mid$ ADJ)	3/3
P(NOUN $\mid$ PREP)	4/7
P(NOUN $\mid$ PREP ART)	2/2
P(ADJ $\mid \phi$ ART)	3/3
P(food $\mid$ hunt for)	1/2
P(deer $\mid$ in the)	0/1
P(often $\mid$ ADV)	2/3
P(for $\mid$ PREP)	3/7

5. (20 pts) Use the following probabilities to answer the questions below. Assume that all probability values NOT listed in the table are zero!

$P(\text{I}) = .30$	$P(\text{I} \mid \phi) = .40$
$P(\text{am}) = .10$	$P(\text{is} \mid \phi) = .25$
$P(\text{Sam}) = .05$	$P(\text{Sam} \mid \phi) = .03$
$P(\text{is}) = .20$	$P(\text{Sam} \mid \text{am}) = .25$
$P(\text{am} \mid \text{I}) = .60$	$P(\text{Sam} \mid \text{is}) = .35$
$P(\text{is} \mid \text{I}) = .08$	$P(\text{I} \mid \text{is}) = .07$
$P(\text{Sam} \mid \text{I}) = .01$	$P(\text{I} \mid \text{Sam}) = .09$

Compute the perplexity of the following word sequences using a unigram language model. **Show all your work!**

(a) I am Sam

(b) Sam I am

(c) I is Sam

(d) is I Sam

I am Sam

$$\begin{aligned} PP(I\,am\,Sam) &= \sqrt[3]{\frac{1}{P(I)P(am)P(Sam)}} \\ &= \sqrt[3]{\frac{1}{(0.30)(0.10)(0.05)}} \\ &= 8.375 \end{aligned}$$

Sam I am

$$\begin{aligned} PP(Sam\,I\,am) &= \sqrt[3]{\frac{1}{P(Sam)P(I)P(am)}} \\ &= \sqrt[3]{\frac{1}{(0.05)(0.30)(0.10)}} \\ &= 8.375 \end{aligned}$$

I is Sam

$$\begin{aligned} PP(I\,is\,Sam) &= \sqrt[3]{\frac{1}{P(I)P(is)P(Sam)}} \\ &= \sqrt[3]{\frac{1}{(0.30)(0.10)(0.05)}} \\ &= 6.9336 \end{aligned}$$

is I Sam

$$\begin{aligned} PP(is\, I\, Sam) &= \sqrt[3]{\frac{1}{P(is)P(I)P(Sam)}} \\ &= \sqrt[3]{\frac{1}{(0.10)(0.30)(0.05)}} \\ &= 6.9336 \end{aligned}$$

Compute the perplexity of the following word sequences using a bigram language model. Show all your work!

(a) I am Sam

(b) Sam I am

(c) I is Sam

(d) is I Sam

I am Sam

$$\begin{aligned} PP(I\, am\, Sam) &= \sqrt[3]{\frac{1}{P(I/\varnothing)P(am/I)P(Sam/am)}} \\ &= \sqrt[3]{\frac{1}{(0.40)(0.60)(0.25)}} \\ &= 2.5543 \end{aligned}$$

Sam I am

$$\begin{aligned} PP(Sam\, I\, am) &= \sqrt[3]{\frac{1}{P(Sam/\varnothing)P(I/Sam)P(am/I)}} \\ &= \sqrt[3]{\frac{1}{(0.03)(0.090)(0.60)}} \\ &= 8.5145 \end{aligned}$$

I is Sam

$$\begin{aligned} PP(I\, is\, Sam) &= \sqrt[3]{\frac{1}{P(I/\varnothing)P(is/I)P(Sam/is)}} \\ &= \sqrt[3]{\frac{1}{(0.40)(0.08)(0.35)}} \\ &= 4.4695 \end{aligned}$$

is I Sam

$$\begin{aligned} PP(is\,I\,Sam) &= \sqrt[3]{\frac{1}{P(is/\varnothing)P(I/is)P(Sam/I)}} \\ &= \sqrt[3]{\frac{1}{(0.25)(0.07)(0.01)}} \\ &= 17.878 \end{aligned}$$

Question #6 is for CS-6340 students ONLY!

6. **(10 pts) The table below contains frequency values for two unigrams, two bigrams, and one trigram based on an imaginary text corpus. Fill in the table below with the smoothed probability of each n-gram using add-k smoothing with the specified value of *k*. You should assume that the vocabulary V consists of 100 distinct unigrams, and the total frequency count over all words (unigrams) in the corpus is 4,000. For the sake of simplicity, you should assume that none of the N-grams occur at the end of a sentence.**

   IMPORTANT: Leave your answer in fractional (numerator/denominator) form!

NOUN	FREQ	SMOOTHED PROB (k=1)	SMOOTHED PROB (k=7)
"natural"	200	8040/41	8280/41
"language"	500	20040/41	20280/41
"natural language"	80	81/300	87/300
"language processing"	60	61/600	67/600
"natural language processing"	20	21/180	27/180