T-61.5020 Statistical Natural Language Processing

Exercises 2 — Entropy and perplexity
Version 1.0

1. Let us examine a small language with six words:

W		P(W)
'kissa'	(cat)	$\frac{3}{32}$
'tuuli'	(wind)	$\frac{3}{16}$
'kiipeilijä'	(climber)	$\frac{7}{32}$
'naukaisi'	(meowed)	$\frac{1}{8}$
'tuivertaa'	(blows)	$\frac{1}{8}$
'katosi'	(disappeared)	$\frac{1}{4}$

   a) Assume that a source generates values for a random variable X according to the table above. What is the entropy of the source $H(X)$?

   b) According to further examination, the language has a sentence structure 'SV' with categories S $\in$ {'kissa', 'tuuli', 'kiipeilijä'} and V $\in$ {'naukaisi', 'tuivertaa', 'katosi'}. The joint distribution P(S,V) of the variables is the following:

	'naukaisi'	'tuivertaa'	'katosi'	
'kissa'	$\frac{1}{8}$	0	$\frac{1}{16}$	$\frac{3}{16}$
'tuuli'	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$
'kiipeilijä'	$\frac{1}{16}$	0	$\frac{3}{8}$	$\frac{7}{16}$
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	

   What is the entropy of the source when we know that the previous symbol belongs to the set S, i.e. what is $H(X_i|X_{i-1} \in S)$?

   c) Assume that a language is generated by the model of part b. Now we model the two-word sentences of the language with the model given in part a. Calculate the average number of bits needed to encode the sentences with optimal code lengths given by the model.

2. Let us examine the random language presented in the first exercise session: We have 30 symbols including a word boundary, all with equal probabilities.

   a) The source generates symbols one by one. What is the entropy of the source?

   b) The source generates one word (one or more non-boundary symbols followed by a word boundary) at a time. Each word is treated as a whole. What is the entropy of the source?

3. We have three grammars presented in the following tables. (Translations: kissa = cat; koira = dog; valas = whale; kala = fish; istui = sat; menee = goes; on = is; puuhun = to tree; kuuhun = to moon; suuhun = to mouth) As a test set for the models, we have two sentences:

   a) Kissa menee puuhun. (Cat goes to tree.)

   b) Valas on kala paitsi ettei. (Whale is fish except not.)

   Calculate the perplexities of the three models for both sentences. Are the results comparable?

Model 1	Model 2
P(sana='kissa')=0.1	P(word=subject)=0.33
P(sana='koira')=0.1	P(word=verb)=0.33
P(sana='valas')=0.1	P(word=object)=0.33
P(sana='kala')=0.1	
P(sana='istui')=0.1	
P(sana='menee')=0.1	
P(sana='on')=0.1	
P(sana='puuhun')=0.1	
P(sana='kuuhun')=0.1	
P(sana='suuhun')=0.1	

Model 3	
P(sana='kissa' | word=first)	=0.25
P(sana='koira' | word=first)	=0.25
P(sana='valas' | word=first)	=0.25
P(sana='kala' | word=first)	=0.25
P(sana='istui' | previous_word ∈ {'kissa','koira','valas','kala'})	=0.33
P(sana='menee' | previous_word ∈ {'kissa','koira','valas','kala'})	=0.33
P(sana='on' | previous_word ∈ {'kissa','koira','valas','kala'})	=0.33
P(sana='puuhun' | previous_word ∈ {'istui','menee','on'})	=0.33
P(sana='kuuhun' | previous_word ∈ {'istui','menee','on'})	=0.33
P(sana='suuhun' | previous_word ∈ {'istui','menee','on'})	=0.33

Perplexity can be defined as the inverse of the geometric mean of the probabilities:

$$Perp(w_1, w_2, \ldots, w_n) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}}$$