Ling 473 Assignment 3
Due 4:30pm on Thursday August 23, 2018

1. (20 points) In Lecture 3, we looked at the outcomes of rolling two fair dice. For this problem, we will consider weighted dice—one white, and one red. For each die, `1` and `6` are twice as likely to show as the other four values.

   a. What is the probability that the total showing on the two dice will be 7?

   b. What is the probability that the total showing on the two dice will be 9 or higher?

   c. What is the probability that the red die will show a higher number than the white one?

2. (35 points) The following is the first paragraph of Ernest Hemmingway's *The Old Man and The Sea*. It has been POS-tagged using the online Brill tagger at the *Center for Sprogteknologi* at Københavns Universitet. A few minor changes have been applied.

PRP	VBD	DT	JJ	NN	WP	VBD		RB		IN	DT	NN		IN	DT	NNP	NNP		CC	PRP	VBD	VBN	CD			NNS	RB	IN		VBG		DT	NN	.
he	was	an	old	man	who	fished		alone		in	a	skiff		in	the	gulf	stream		and	he	had	gone	eighty-four			days	now	without		taking		a	fish	.

IN	DT	JJ		CD		NNS	DT	NN	VBD	VBN	IN		PRP	.	CC	IN		CD		NNS	IN		DT	NN		DT	NN	POS	NNS			VBD	VBN	PRP	IN		DT
in	the	first		forty		days	a	boy	had	been	with		him	.	but	after		forty		days	without		a	fish		the	boy	's	parents			had	told	him	that		the

JJ	NN	VBD	RB	RB			CC	RB		VBN	,	WDT	VBZ	DT	JJ		NN		IN	JJ		,	CC	DT	NN	VBD	VBN	IN	PRP$	NNS		IN
old	man	was	now	definitely			and	finally		salao	,	which	is	the	worst		form		of	unluck		,	and	the	boy	had	gone	at	their	orders		in

DT		NN	WDT	VBD		CD		JJ	NN	DT	JJ		NN		.	PRP	VBD	DT	NN	JJ	TO	VB	DT	JJ	NN	VB		IN	DT		NN	IN		PRP$
another		boat	which	caught		three		good	fish	the	first		week		.	it	made	the	boy	sad	to	see	the	old	man	come		in	each		day	with		his

NN		JJ		CC	PRP	RB		VBD	IN		TO	VB	PRP	VB		DT		DT	VBD		NNS	CC	DT	NN		CC	NN			CC	DT	NN	WDT	VBD
skiff		empty		and	he		always	went	down		to	help	him	carry		either		the	coiled		lines	or	the	gaff		and	harpoon			and	the	sail	that	was

VBD		IN		DT	NN		.	DT	NN		VBD	VBN		IN		NN		NNS	CC	,	VBD		,	PRP	VBD		IN	DT	NN		IN	JJ			NN		.
furled		around		the	mast		.	the	sail		was	patched		with		flour		sacks	and	,	furled		,	it	looked		like	the	flag		of	permanent			defeat		.

This assignment does not require programming, but if you wish to work with an electronic version of this information, you can refer to the following file:

/opt/dropbox/16-17/473/assignment3/old-man.txt

   a. How many bigrams does the sample contain?

   b. In a bigram model, we assume that a POS tag depends only on the POS tag of the preceding word. Calculate $P(. \mid$ `NN`$)$, assuming that the counts in the above sample are perfectly representative.

   c. We are interested in the probability of the bigram `DT JJ` in the sample text. What is the value of $P($`DT JJ`$)$?

   d. A trigram model predicates a POS tag on the POS tags of the preceding bigram. Calculate $P($`NN` $\mid$ `DT JJ`$)$ for the sample.

   e. Assume this sample characterizes a larger corpus. Assume that measured probabilities are <u>independent</u>. Estimate $P($`DT JJ` $\mid$ `NN`$)$ for the corpus. (Hint: this will use Bayes' Theorem.) Show your work.

3. (15 points) For phonetic elicitation with a group of American test subjects, we are using three word lists:

A = { *gnat*, *beet* }
B = { *loon*, *fee* }
C = { *peel*, *pool*, *he*, *sand* }

The test protocol is as follows: One of the lists is selected at random. Then, the subject is asked to pronounce a randomly selected word from that list. What is the probability that the word will have a high/close vowel (as opposed to low/open)? If you are not familiar with vowel phonetics, you can check the Lecture 5 recording, or listen to samples on http://en.wikipedia.org/wiki/Vowel.

4. (30 points) A classifier has portioned a set of eight biomedical documents into

C = { mentions the IL-2R α-promoter } (6 documents), and
$\bar{C}$ (the rest).

The gold standard indicates that only three documents actually mention the Interleukin-2 receptor alpha promoter, and we determine that exactly one of them is (incorrectly) in $\bar{C}$. In testing a post-processing heuristic, we select a document at random from C and move it in the class $\bar{C}$. Next, we randomly select a document from $\bar{C}$.

a. What is the probability that the document we selected from $\bar{C}$ mentions the IL-2R α-promoter (according to the gold standard)?

b. Next, we note that the document we selected from $\bar{C}$ *does*, in fact (according to the gold standard), mention the IL-2R α-promoter. Given this additional information, what is the probability that the document that we transferred from C to $\bar{C}$ mentioned (according to the gold standard) the IL-2R α-promoter (i.e., that we moved it to the wrong class)?