CS-5340/6340, Written Assignment #1
DUE: Tuesday, September 8, 2020 by 11:59pm

Submit your assignment on CANVAS in pdf format.

1. (34 pts) In each sentence below, place brackets [ ] around each base noun phrase (NP). Then label each NP with one of the following syntactic roles: **SUBJ** (subject), **DOBJ** (direct object), **IOBJ** (indirect object), or **OTHER** if the NP is not a subject, direct object, or indirect object. Please format your answers as [NP]/ROLE, for example: [the clown]/SUBJ .

   (a) Three young boys went hiking up the mountain .

   (b) The software company awarded her a $10,000 prize for her excellent management .

   (c) Dead squirrels are occasionally found in swimming pools .

   (d) Listen to that loud thunder !

   (e) An old man sold his beloved car to several neighbors .

   (f) Natural language processing is really fun .

   (g) A family from Idaho brought the puppy some tasty treats .

2. (20 pts) For each sentence below, indicate whether the verb phrase is in an **active voice** or **passive voice** construction.

   (a) The dog slept by the fire all night.

   (b) The boat in the harbor was sunk by a torpedo.

   (c) The deer had been shot near the road.

   (d) Susan will be awarded the grand prize at the science fair.

   (e) The new iPhone can not be purchased until 2021.

   (f) Raccoons have been regularly digging in my garden.

   (g) The boy had been bullied at his previous school.

   (h) Tom has been preparing for the entrance exam for a month.

   (i) The kids were not smiling in the Christmas photo.

   (j) They should have seen the warning sign on the door.

3. (20 pts) Use the Dictionary and Morphology Rules shown below to answer this question.

Dictionary	
appropriate	ADJ
infect	VERB
human	NOUN
humane	ADJ
smoke	NOUN, VERB

Rule ID	**Prefix**	**Suffix**	**Replace Chars**	**Root POS**	**Derived POS**
R1		ant		VERB	NOUN
R2		ation	e	VERB	NOUN
R3		er		VERB	NOUN
R4		er	e	VERB	NOUN
R5		ier	y	ADJ	ADJ
R6		ize		NOUN	VERB
R7		ly		ADJ	ADV
R8		ness		ADJ	NOUN
R9		s		NOUN	NOUN
R10		s		VERB	VERB
R11		y	e	NOUN	ADJ
R12		y		NOUN	ADJ
R13	de			VERB	VERB
R14	dis			VERB	VERB
R15	in			ADJ	ADJ

For each word given below, list <u>all</u> of the derivations that are possible using the Dictionary and Morphology Rules shown above. For each derivation, (1) list the rules that apply, *in the order that they would be applied, starting with the given word*, and (2) indicate the part-of-speech that would ultimately be assigned to the given word. Be sure to list ALL legal derivations, even if some would result in the same part-of-speech assignment. If no derivations are possible for a word, then answer NO DERIVATIONS.

(a) smokier

(b) inappropriateness

(c) humanly

(d) dehumanization

(e) disinfects

(f) disinfectants

4. (18 pts) Tom and Jerry each labeled 10 newspaper articles (D1-D10) with respect to 3 categories: **Arts (A)**, **Finance (F)**, and **Politics (P)**. Their labels are shown below.

Document	**Tom**	**Jerry**
D1	P	A
D2	A	A
D3	A	A
D4	F	F
D5	P	P
D6	P	A
D7	F	F
D8	A	P
D9	P	P
D10	P	F

Show all your work, including the numerator and denominator of fractions! You will not get credit if you only show the final number as the answer to a question.

(a) Compute the inter-annotator agreement between Tom and Jerry's labels using the Kappa (κ) statistic.

(b) Compute the Accuracy of Tom's labels when treating Jerry's labels as the gold standard.

(c) Compute the Recall and Precision of Tom's labels for the **Arts** category when treating Jerry's labels as the gold standard.

(d) Compute the Recall and Precision of Tom's labels for the **Finance** category when treating Jerry's labels as the gold standard.

(e) Compute the Recall and Precision of Tom's labels for the **Politics** category when treating Jerry's labels as the gold standard.

(f) Imagine a trivial system that assigns every document to the **Arts** category. Compute the system's Recall and Precision for the **Arts** category when treating Jerry's labels as the gold standard.

5. (8 pts) Cross-validation questions.

   (a) Suppose you evaluate a machine learning (ML) system by performing 5-fold cross-validation using a collection of 200 annotated documents. For each experiment, how many documents will be used to train the ML model?

   (b) Suppose you evaluate a machine learning (ML) system by performing 25-fold cross-validation using a collection of 500 annotated documents. For each experiment, how many documents will be used to train the ML model?

   (c) Given a collection of D documents, what is the maximum number of folds that could be used to perform cross-validation?

   (d) Given a collection of D documents, what is the minimum number of folds that could be used to perform cross-validation?

Question #6 is for CS-6340 students ONLY!

6. (12 pts) The table below contains frequency counts for the words "good", "bad", and "scary" from a small (imaginary!) corpus of 6 movie review documents (D1-D6). Assume that these 3 words make up your entire vocabulary. Each document has been labeled as either a Positive (+) or Negative (-) review. Use the information in this table to answer the questions below. Use Log base 2 (log_2) in your equations. *Show all your work! You will not get credit if you only show the final number as an answer.*

	"good"	**"bad"**	**"scary"**	**Class**
D1	4	1	1	+
D2	2	0	0	+
D3	3	1	0	-
D4	0	2	1	-
D5	2	1	0	-
D6	1	0	1	-

(a) Compute loglikelihood("good",+)

(b) Compute loglikelihood("good",-)

(c) Compute loglikelihood("bad",+)

(d) Compute loglikelihood("bad",-)

(e) Compute loglikelihood("scary",+)

(f) Compute loglikelihood("scary",-)

For the questions below, assume that only "good", "bad" and "scary" are in your vocabulary (i.e., ignore all other words).

(g) For each Class, compute the numeric value that the Naive Bayes algorithm would produce for the review: *"This movie is so bad that it's scary ."*

(h) For each Class, compute the numeric value that the Naive Bayes algorithm would produce for the review: *"This movie is good ! So scary, really good !"*

(i) For each Class, compute the numeric value that the Naive Bayes algorithm would produce for the review: *"Bad bad movie . It is scary and the acting is good but the plot is bad ."*