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]/SUBJ went hiking up [the mountain]/DOBJ .

   (b) [The software company]/SUBJ awarded [her]/IOBJ a [$10,000 prize]/DOBJ for [her excellent management]/OTHER .

   (c) [Dead squirrels]/DOBJ are occasionally found in [swimming pools]/OTHER .

   (d) Listen to that [loud thunder]/DOBJ !

   (e) [An old man]/SUBJ sold [his beloved car]/DOBJ to [several neighbors]/IOBJ .

   (f) [Natural language processing]/DOBJ is really fun .

   (g) [A family]/SUBJ from Idaho brought [the puppy]/IOBJ some [tasty treats]/DOBJ .

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

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

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

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

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

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

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

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

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

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

   (j) **Active Voice** 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.

Note, this is around 20 Minutes in the Lecture Video Sept1-ML+Eval.mp4

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

  i. R5 → smoky
  ii. R11 → smoke
  iii. Derived POS: ADJ

(b) inappropriateness

  i. R8 → inappropriate
  ii. R15 → appropriate
  iii. Derived POS: ADJ

(c) humanly

  i. R7 → human
  ii. But since the Root POS (ADJ) is not in the dictionary for human, there is no derivation

    iii. NO DERIVATIONS

(d) dehumanization

    i. R2 → dehumanize
    ii. R6 → dehuman
    iii. R13 → human
    iv. But since the Root POS (VERB) is not in the dictionary for human, there is no derivation.
    v. NO DERIVATIONS

(e) disinfects

    i. R9 → disinfect (noun)
    ii. R14 → infect (verb)
    iii. Now, rewind the recursion. R14 passes back a Verb. R9 however is expecting a Noun so in this case, it doesn't work.

Next case in the exhaustive Depth first search

    i. R10 → disinfect (Verb)
    ii. R14 → infect (verb)
    iii. Now, rewind the recursion. R14 passes back a Verb. R10 this time is expecting a Verb, so then the **derived POS is VERB**.

(f) disinfectants (note this is a case where the exhaustive depth-first search in the lecture around 39 minutes Sept1-ML+Eval.mp4 is really helpful)

    i. R9 → disinfectant
    ii. R1 → disinfect
    iii. R14 → infect
    iv. Now, rewind the recursion. R14 passes back a verb. R1 is expecting a verb and passes back a Noun. R9 is expected a Noun and thus the Derived POS: NOUN

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.

$$\kappa = \frac{P(agree) - P(expected)}{1 - P(expected)}$$

$$P(agree) = \frac{6}{10} = 0.6$$

$$P(expected) = P(P|Tom)*P(P|Jerry)+P(F|Tom)*P(F|Jerry)+P(A|Tom)*P(A|Jerry)$$

$$P(expected) = 5/10 * 3/10 + 2/10 * 3/10 + 3/10 * 4/10 = 0.33)$$

$$\kappa = \frac{0.6 - 0.33}{1 - 0.33} = 0.402$$

(b) Compute the Accuracy of Tom's labels when treating Jerry's labels as the gold standard. This is the same as $P(agree)$ from part (a)

$$\frac{6}{10} = 0.6$$

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

For recall, there are 4 newspaper articles Jerry labeled as Arts. Assuming Jerry is gold standard, recall would be what proportion of the 4 did Tom also categorize as Arts. Precision would be what proportion of the articles Tom labeled as Arts were actually Arts (according to Jerry).

$$Recall = 2/4 = 0.5$$

$$Precision = 2/3 = 0.666$$

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

$$Recall = 2/3 = 0.666$$

$$Precision = 2/2 = 1$$

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

$$Recall = 2/3$$

$$Precision = 2/5$$

(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.

If the system assigns every document to Arts, the Recall is always 1 (regardless of gold standard) since you're always capturing all the true positives

$$Recall = 1$$

$$Precision = 4/10$$

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?

   Each fold is $200/5 = 40$ documents. So for each experiment, 160 documents will be used (1 fold is left out)

   (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?

   Each fold is $500/25 = 20$ documents. So each experiment, 480 documents are used

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

   D this would be leave one out cross validation

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

   2 this would just be 2-fold 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”,+)

$$Likelihood = \frac{count(“good'', +)}{\sum_{w \in V} count(w, +)} = 6/8 = 0.75$$

$$loglikelihood = log_2(\frac{7}{9}) = -0.362$$

Remember add 1 for Laplace Smoothing

(b) Compute loglikelihood(“good”,-)

$$Likelihood = \frac{count(“good'', -)}{\sum_{w \in V} count(w, -)} = 6/12 = 0.5$$

$$loglikelihood = log_2(\frac{7}{13}) = -0.893$$

Remember add 1 for Laplace Smoothing

(c) Compute loglikelihood(“bad”,+)

$$Likelihood = \frac{count(“bad'', +)}{\sum_{w \in V} count(w, +)} = 1/8 = 0.125$$

$$loglikelihood = log_2(\frac{2}{9}) = -2.169$$

(d) Compute loglikelihood(“bad”,-)

$$Likelihood = \frac{count(“bad'', -)}{\sum_{w \in V} count(w, -)} = 4/12 = 0.333$$

$$loglikelihood = log_2(\frac{5}{13}) = -1.378$$

(e) Compute loglikelihood(“scary”,+)

$$Likelihood = \frac{count(\text{“}scary'', +)}{\sum_{w \in V} count(w, +)} = 1/12 = 0.0833$$

$$loglikelihood = log_2(\frac{2}{13}) = -2.7$$

(f) Compute loglikelihood(“scary”,-)

$$Likelihood = \frac{count(\text{“}scary'', -)}{\sum_{w \in V} count(w, -)} = 2/12 = 0.1666$$

$$loglikelihood = log_2(\frac{3}{13}) = -2.115$$

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

Note:

$$logprior(+) = log_2(1/3) = -1.584$$

$$logprior(-) = log_2(2/3) = -0.584$$

(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 .”*

i. +: 1 occurrence of the word “bad”‘ and 1 occurrence of the word “scary”

$$logprior(+) + 1 * loglikelihood(\text{“}bad'', +) + 1 * loglikelihood(\text{“}scary'', +) =$$

$$-1.583 + (-2.169) + (-2.7) = -6.452$$

ii. -: 1 occurrence of the word “bad”‘ and 1 occurrence of the word “scary”

$$logprior(-) + 1 * loglikelihood(\text{“}bad'', -) + 1 * loglikelihood(\text{“}scary'', -) =$$

$$-0.584 + (-1.378) + (-2.115) = -4.077$$

(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. +: 2 occurrences of the word “good” and 1 occurrence of the word “scary”

$$logprior(+) + 2 * loglikelihood(\text{“}good'', +)$$

$$+1 * loglikelihood(\text{“}scary'', +) =$$

$$-1.583 + (2 * -0.362) + (-2.7) = -5.007$$

ii. -: 2 occurrences of the word “good” and 1 occurrence of the word “scary”

$$logprior(-) + 2 * loglikelihood(\text{“}good'', -)$$

$$+1 * loglikelihood(\text{“}scary'', -) =$$

$$-0.584 + (2 * -0.893) + (-2.115) = -4.485$$

Since naive bayes would return the argmax, naive bayes would return "-" as the class

(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 ."*

   i. +: 1 occurrence of the word "good". 3 occurrences of the word "bad". 1 occurrence of the word "bad".

$$logprior(+) + 1 * loglikelihood(``good'', +) + 3 * loglikelihood(``bad'', +)$$

$$+1 * loglikelihood(``scary'', +) =$$

$$-1.583 + (1 * (-0.362)) + 3 * (-2.169) + (-2.7) = -11.152$$

   ii. -: 1 occurrence of the word "good". 3 occurrences of the word "bad". 1 occurrence of the word "bad".

$$logprior(-) + 1 * loglikelihood(``good'', -) + 3 * loglikelihood(``bad'', -)$$

$$+1 * loglikelihood(``scary'', -) =$$

$$-0.584 + (1 * (-0.893)) + 3 * (-1.378) + (-2.115) = -7.726$$

Since naive bayes would return the argmax, naive bayes would return "-" as the class