NLP Written Assignment 01

Sushmitha Sunkurdi Nataraj, u1265043

September 2020

(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 .

1. Three young boys went hiking up the mountain .
   [**Three young boys**]**/SUBJ** went [**hiking**]**/DOBJ** up [**the mountain**]**/OTHER**
2. The software company awarded her a $10,000 prize for her excellent management .
   [**The software company**]**/ SUBJ** awarded [**her**]**/IOBJ** [**a $10,000 prize**]**/DOBJ** for [**her excellent management**]**/OTHER** .
3. Dead squirrels are occasionally found in swimming pools .
   [**Dead squirrels**]**/SUBJ** are occasionally found in [**swimming pools**]**/OTHER**.
4. Listen to that loud thunder !
   Listen to [**that loud thunder**]**/DOBJ**!
5. An old man sold his beloved car to several neighbors.
   [**An old man**]**/SUBJ** sold [**his beloved car**]**/DOBJ** to [**several neighbors**]**/IOBJ** .
6. Natural language processing is really fun .
   [**Natural language processing**]**/SUBJ** is really fun .
7. A family from Idaho brought the puppy some tasty treats .
   [**A family**]**/SUBJ** from [**Idaho**]**/OTHER** brought [**the puppy**]**/IOBJ** [**some tasty treats**]**/DOBJ** .

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

1. The dog slept by the fire all night.
   Active Voice.

2. The boat in the harbor was sunk by a torpedo.
   Passive voice

3. The deer had been shot near the road.
   Passive voice

4. Susan will be awarded the grand prize at the science fair.
   Passive Voice

5. The new iPhone can not be purchased until 2021.
   Passive Voice.

6. Raccoons have been regularly digging in my garden.
   Active Voice.

7. The boy had been bullied at his previous school.
   Passive voice.

8. Tom has been preparing for the entrance exam for a month.
   Active Voice.

9. The kids were not smiling in the Christmas photo.
   Active Voice.

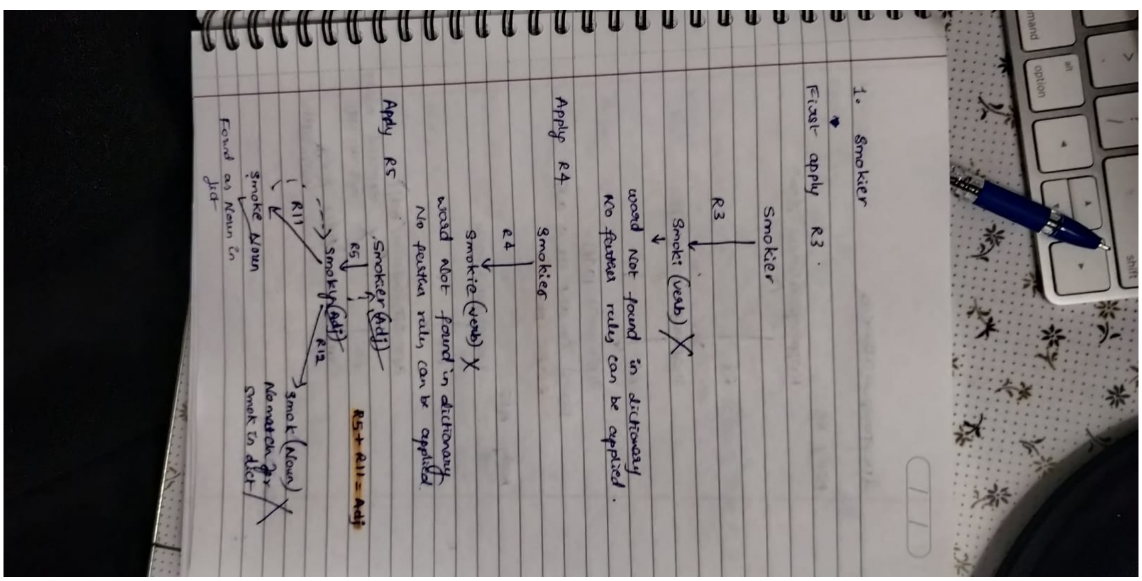
10. They should have seen the warning sign on the door.
    Active Voice.

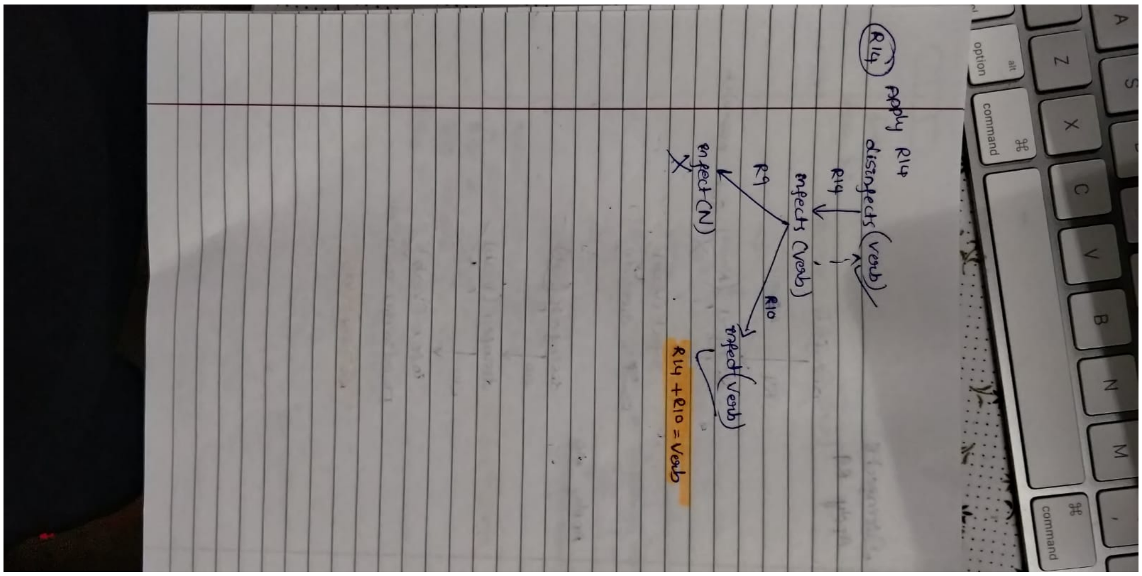
(20 pts) Use the Dictionary and Morphology Rules shown below to answer this question.

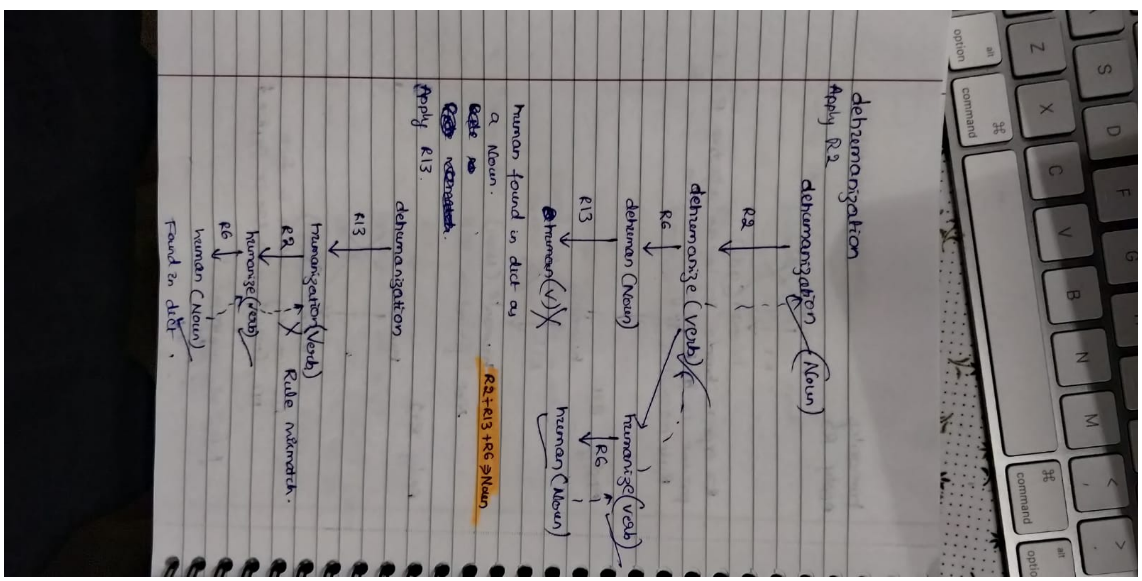
Dictionary	
appropriate	ADJ
infect	VERB
human	NOUN
humane	ADJ
smoke	NOUN, VERB

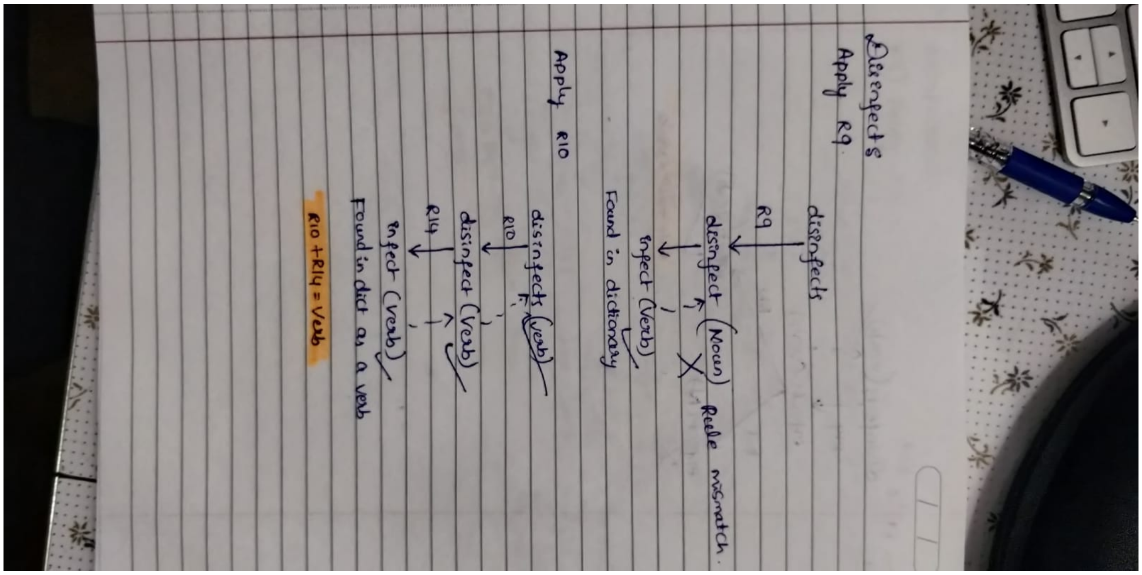
Rule ID	Prefix	Suffix	Replace Chars	Root POS	Derived POS
R1		ant		VERB	NOUN
R2		ation	e	VERB	NOUN
R3		er		VERB	NOUN
R4		er	e	VERB	NOUN
R5		ier	y	ADJ	ADJ
R6		ize		NOUN	VERB
R7		ly		ADJ	ADV
R8		ness		ADJ	NOUN
R9		s		NOUN	NOUN
R10		s		VERB	VERB
R11		y	e	NOUN	ADJ
R12		y		NOUN	ADJ
R13	de			VERB	VERB
R14	dis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

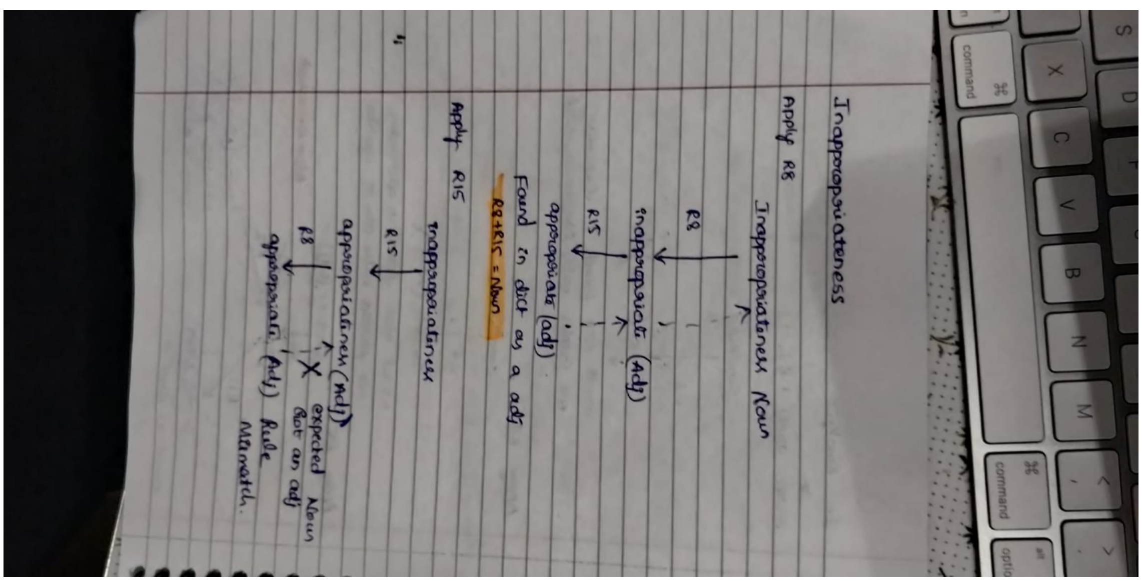
For each word given below, list all of the derivations that are possible using the Dictionary and Morphology Rules shown above. For each derivation, (1) list the rules that apply, *in the order that they would be applied, starting with the given word*, and (2) indicate the part-of-speech that would ultimately 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.




R14
Apply R14
disinfects (verb)
R14
infects (verb)
R9
R10
infect (N)
infect (verb)
R14 + R10 = Verb


dehumanization
Apply R2
dehumanization (Noun)
R2
dehumanize (verb)
R6
dehuman (Noun)
humanize (verb)
R6
R13
human (v)
human (Noun)
human found in dict as
a Noun.
R2+R13+R6 ⇒ Noun
Apply R13.
dehumanization
R13
humanization (Verb)
Rule mismatch.
R2
humanize (verb)
R6
human (Noun)
Found in dict.


Disinfects
Apply R9
disinfects
R9
disinfect (Noun) Rule mismatch.
infect (Verb)
Found in dictionary
Apply R10
disinfects (verb)
R10
disinfect (Verb)
R14
infect (verb)
Found in dict as a verb
R10 + R14 = Verb


Inappropriateness
Apply R8
Inappropriateness Noun
R8
inappropriate (Adj)
R15
appropriate (adj)
Found in dict as a adj
R8+R15 = Noun
Apply R15
inappropriateness
R15
appropriateness (Adj)
expected Noun
Got an adj
R8
appropriate (Adj) Rule Mismatch.

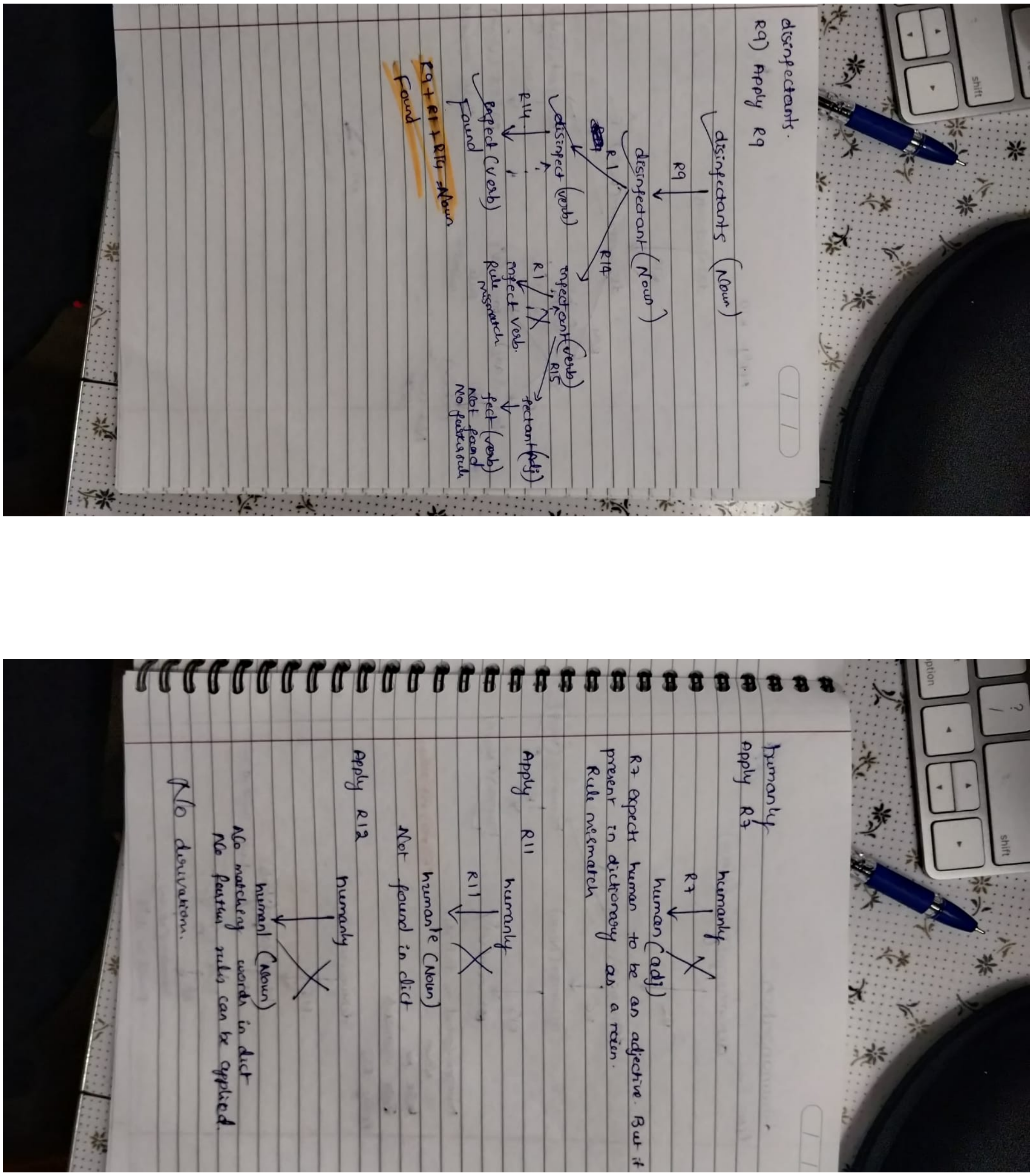


Figure 1: Depth First Tree with R3 applied

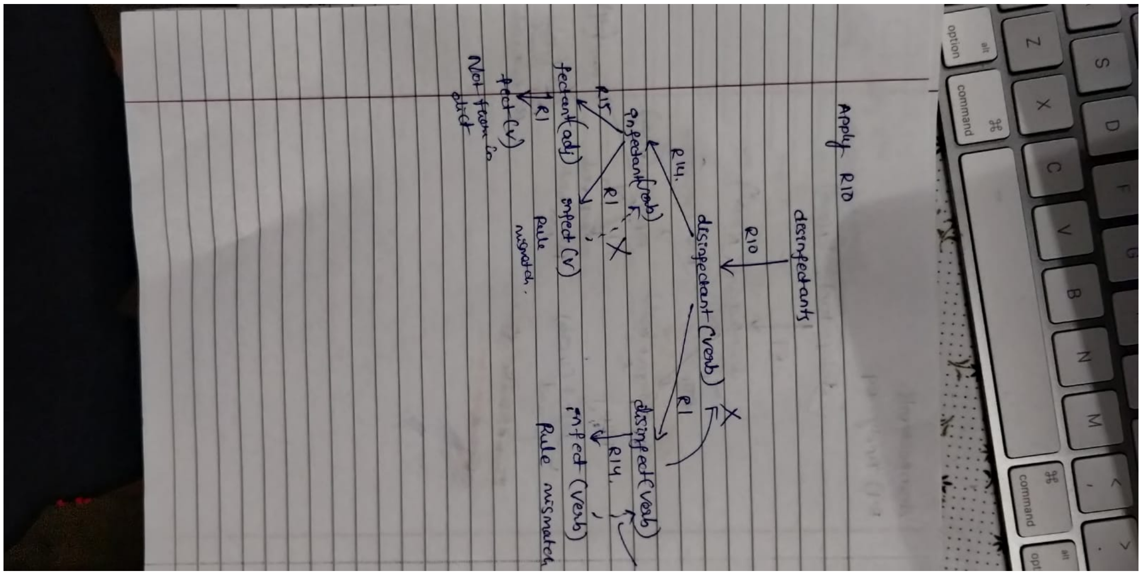


Figure 2: Depth First Tree with R3 applied

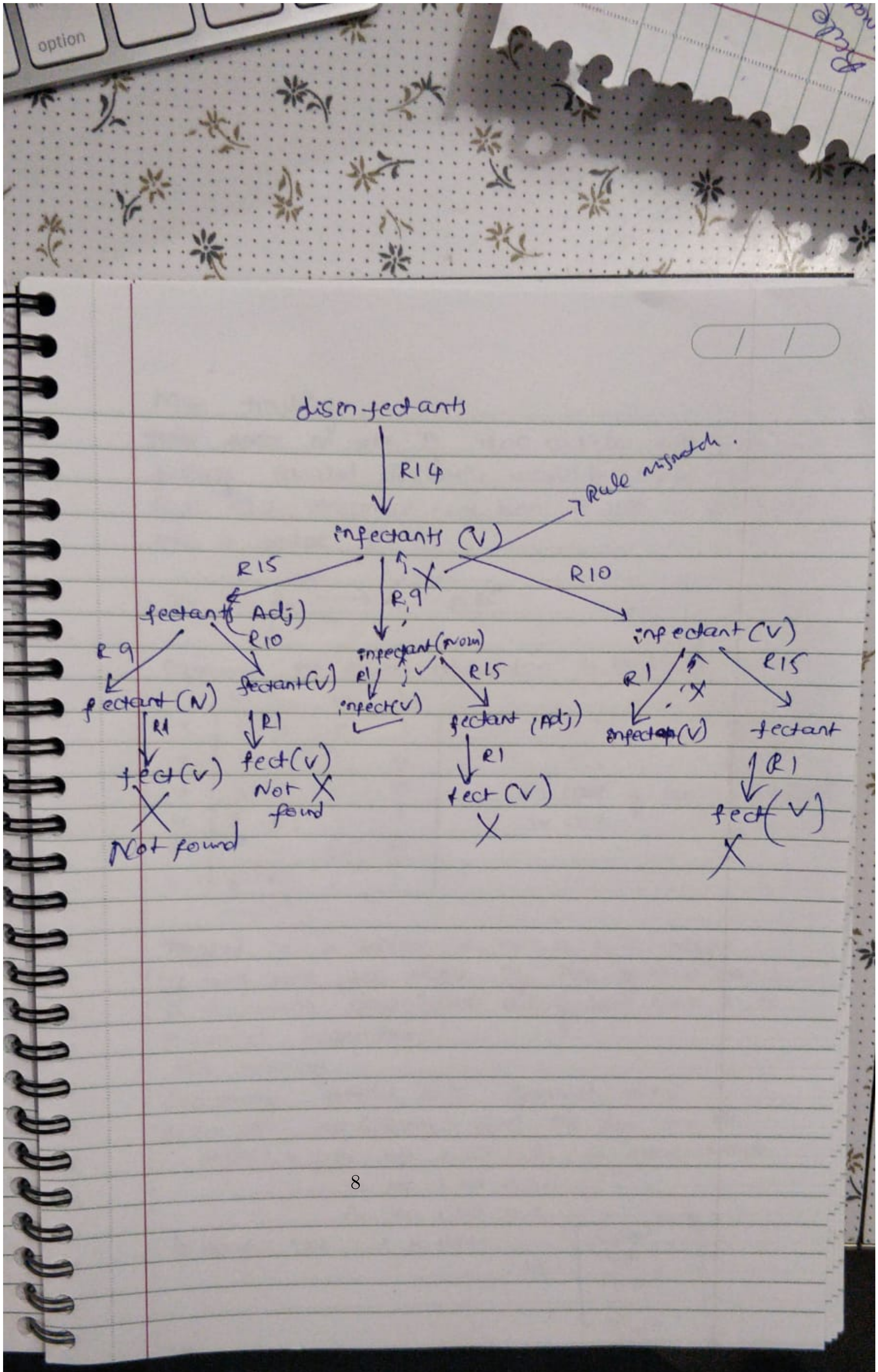
(18 pts) Tom and Jerry each labeled 10 newspaper articles (D1-D10) with respect to 3 categories: **Arts (A)**, **Finance (F)**, and **Politics (P)**. Their labels are shown below.

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.

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

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


disinfectants
R14
infectants (V)
Rule mismatch.
R15
R9
R10
fectant (Adj)
infectant (Noun)
infectant (V)
R9
R10
R1
R15
R1
R15
fectant (N)
fectant (V)
infect (V)
fectant (Adj)
infectant (V)
fectant
R1
R1
R1
R1
fect (V)
fect (V)
fect (V)
fect (V)
Not found
Not found

P(agree) = proportion of times the annotators agree
P(expected) = proportion of times the annotators are expected to agree by chance.

P(expected) = $\sum_{c \in C} P(c|A_1) * P(c|A_2)$ where: C = the set of possible classes (labels)
$A_1 = annotator\ 1's\ labels\ (Here\ Tom's\ label)$
$A_2 = annotator\ 2's\ labels\ (Here\ Jerry's\ label)$

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

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

$$P(expected) = \frac{3*4}{10*10} + \frac{2*3}{10*10} + \frac{5*3}{10*10}$$

$$P(expected) = \frac{12+6+15}{100}$$

$$P(expected) = \frac{33}{100} = 0.33$$

$$\kappa = \frac{P(agree) - P(expected)}{1 - P(expected)} = \frac{0.6 - 0.33}{1 - 0.33} = \frac{0.27}{0.67} = 0.40298$$

2. Compute the Accuracy of Tom's labels when treating Jerry's labels as the gold standard.

   Accuracy: the % of instances assigned a correct label

$$D_2, D_3, D_4, D_5, D_7, D_9\ are\ assigned\ correct\ labels.$$

$$Accuracy = \frac{6}{10} = 0.6 = 60\%$$

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

   $Recall:\ for\ a\ category\ C,\ the\ \%of\ true\ instances\ of\ C\ that\ are\ correctly\ labeled:$

$$Recall = \frac{number\ of\ records\ correctly\ labelled\ as\ C}{number\ of\ true\ instances\ of\ C}$$

$$\frac{2}{4} = 0.5 = 50\%$$

$$Precision:\ for\ a\ category\ C,\ the\ \%of\ instances\ assigned\ the\ label\ C\ that\ are\ correctly\ labeled$$

$$Precision = \frac{number\ of\ instances\ correctly\ labeled\ as\ C}{number\ of\ instances\ labelled\ as\ C}$$

$$Precision = \frac{2}{3} = 66.66\%$$

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

   I am using the same formulae as from the previous question for Recall and precision. So, I am not gonna write the formula again. I am reusing the formula from Q3

$$Recall = \frac{2}{3} = 66.66\%$$

$$Precision = \frac{2}{2} = 1 = 100\%$$

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

$$Recall = \frac{2}{3} = 66.66\%$$

$$Precision = \frac{2}{5} = 0.4 = 40\%$$

6. 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.

$$Recall = \frac{4}{4} = 1 = 100\%$$

$$Precision = \frac{4}{10} = 0.4 = 40\%$$

(8 pts) Cross-validation questions.

1. 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 will have $\frac{200}{5} = 40$ elements. There are 5 such folds. We will use $\frac{4}{5}$ folds to train the model. So that will be 40*4 = 160 documents.

2. 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?
   There are 25 folds. Each fold will have $\frac{500}{25} = 20$ elements. We will use 24 folds for training. So, that will be 24*20 = 480 documents.

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

   We can have a maximum of D folds, where each fold will have 1 element.

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

   2. One for training and one for testing.

Question #6 is for CS-6340 students ONLY!

(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	-

1. Compute loglikelihood("good",+)

$$loglikelihood(good,+) = log\frac{count(good,+)+1}{\sum_{w' \in V}(count(w',c)+1)}$$

   Where count(good, +) is number of occurences of good in all the documents with + label.

$$loglikelihood(good,+) = log_2\frac{6+1}{(6+1)+(1+1)+(1+1)} = log_2\frac{7}{11} = -0.652076$$

2. Compute loglikelihood("good",-)

$$loglikelihood(good,-) = log_2\frac{6+1}{(6+1)+(4+1)+(2+1)} = log_2\frac{7}{15} = -1.09952$$

3. Compute loglikelihood("bad",+)

$$loglikelihood(bad,+) = log_2\frac{1+1}{(1+1)+(6+1)+(1+1)} = log_2\frac{2}{11} = -2.4594$$

4. Compute loglikelihood("bad",-)

$$loglikelihood(bad,-) = log_2\frac{4+1}{(4+1)+(6+1)+(2+1)} = log_2\frac{5}{15} = -1.584963$$

5. Compute loglikelihood(“scary”,+)

$$loglikelihood(scary,+) = log_2 \frac{1+1}{(1+1)+(6+1)+(1+1)} = log_2 \frac{2}{11} = -2.45943$$

6. Compute loglikelihood(“scary”,-)

$$loglikelihood(scary,-) = log_2 \frac{2+1}{(2+1)+(4+1)+(6+1)} = log_2 \frac{3}{15} = -2.321928$$

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

7. 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 .”*

bad occurred 1 time.
scary occurred one time
There are 2 classes + and -.
For class +,
sum [+]= (logprior(+)) + logliklihood(bad, +)+ logloklihood(scary, +)

$$logprior(+) = log_2(number\ of\ documents\ with\ '+'class\ /total\ number\ of\ documents)$$

$$logprior(+) = log_2(2/6) = -1.58496$$

I am reusing the values of logliklihood from pervious question.
sum[+] = -1.58496 -2.4594 -2.45943 = -6.49482

For class -,
sum[-] = logprior[-] + logliklihood(bad, -) + logliklihood(scary, -)

$$logprior(-) = log_2(number\ of\ documents\ with\ '-'class\ /total\ number\ of\ documents)$$

I am reusing logliklihood values from previous question. sum[-] = -0.58496 -1.584963 - 2.321928 = -4.491853

Maximum value is class ’-’ = -4.491853

8. 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 !"*
   good has occurred 2 times and scary has occurred once.
   Numeric value of Naive Bayes algorithm for + and - class is calculated as follow as:
   sum[+] = logprior(+) + (logliklihood(good,+)*2) +logliklihood(scary, +)
   I am reusing the values of logprior and logliklihoods from previous questions
   sum[+] = -1.58496 - 0.652076 * 2 - 2.45943= -5.48542

   sum[-] = logprior(-) + (logliklihood(good,-)*2) +logliklihood(scary, -)
   sum[-] = -0.58496 -1.09952 * 2 -2.321928 = -5.1
   Maximum of above values is -5.1 with - class

9. 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 ."*

   bad has occurred 3 times,
   scary occurred once and good occurred once.

   Numeric value of Naive Bayes algorithm for + and - class is calculated as follow as:
   sum [+] = logprior(+) + logliklihood(bad, +) * 3 + logliklihood(scary, +) + logliklihood(good, +)

   $$logprior(+) = log_2(number\ of\ documents\ with\ '+'class\ /total\ number\ of\ documents)$$

   sum[+] = -1.58496 + 3*(-2.4594) - 0.652076-2.45943 = -12.0746

   sum [-] = logprior(-) + logliklihood(bad, -) * 3 + logliklihood(scary, -) + logliklihood(good, -)

   I am reusing logliklihhod and logprior values from previous questions. sum[-] = -0.58496 + 3*(-1.58496) - 1.09952 - 2.32198 = -8.76134

   ans is max(-8.76134,-12.0746) is -8.76134 with - class.