CS-5340/6340, Written Assignment #3
DUE: Tuesday, November 23, 2021 by 11:59pm

Submit your assignment on CANVAS in pdf format.

1. (20 pts) Show the instantiated patterns that would be generated by the AutoSlog pattern generator when given each sentence below as input, where the **bold-faced** noun phrase (NP) has been labeled with the event role indicated after the slash.

   For example, "**(John Smith)/VICTIM** was shot" indicates that "John Smith" was labeled as a VICTIM by a human.

   (a) **(The tourist)/VICTIM** was robbed in Spain by a young man.
   <VICTIM> was robbed

   (b) The tourist was robbed in Spain by **(a young man)/PERPETRATOR**.
   by <PERPETRATOR>

   (c) **(XYZ Co.)/SELLER** tried to sell ABC Inc. for 20 million dollars.
   <SELLER> to sell

   (d) XYZ Co. tried to sell **(ABC Inc.)/SELLEE** for 20 million dollars.
   to sell<SELLEE>

   (e) XYZ Co. tried to sell ABC Inc. for **(20 million dollars)/PRICE**.
   for <PRICE>

   (f) **(A tall woman)/PERPETRATOR** with a bomb was seen near the enormous explosion of a clothing factory.
   <PERPETRATOR> with a bomb

   (g) A tall woman with **(a bomb)/WEAPON** was seen near the enormous explosion of a clothing factory.
   with <WEAPON>

   (h) A tall woman with a bomb was seen near the enormous explosion of **(a clothing factory)/TARGET**.
   explosion of <TARGET>

2. (35 pts) For each sentence below, label each noun phrase (NP) with the appropriate thematic role. Put parentheses around the NP and indicate the thematic role with a slash, for example: (Mickey Mouse)/AGENT.

   (a) George flew to Las Vegas by helicopter.
   (George)/AGENT flew to (Las Vegas)/TO-LOC by (helicopter)/INSTRUMENT

   (b) Jim built his children a treehouse with his brother.
   (Jim)/AGENT built (his children)/RECIPIENT (a treehouse)THEME with (his brother)/CO-AGENT.

   (c) The young girl raced with her sister around the yard.
   (The young girl)/AGENT raced with (her sister)/CO-AGENT around (the yard)/THEME.

   (d) The waiter served Charlie a salad with Italian dressing.
   (The waiter)/AGENT served (Charlie)/RECIPIENT (a salad)/THEME with (Italian dressing)/INSTRUMENT.

   (e) The ballerina taught us dancing.
   (The ballerina)/AGENT taught (us)/BENEFICIARY (dancing)/THEME.

   (f) Two security cameras recorded the burglary.
   (Two security cameras)/AGENT recorded (the burglary)/THEME.

   (g) Three collared cougars were tracked by officials with GPS devices.
   (Three collared cougars)/THEME were tracked by (officials)/AGENT with (GPS devices)/INSTRUMENT.

   (h) Susan fixed the broken cabinet for her ailing grandmother.
   (Susan)/AGENT fixed (the broken cabinet)/THEME for (her ailing grandmother)/RECIPIENT.

   (i) Ted bought a kitchen table with chairs with his wife.
   (Ted)/AGENT bought (a kitchen table)/THEME with (chairs)/CO-THEME with (his wife)/CO-AGENT.

   (j) Jim's cancer was cured by chemotherapy.
   (Jim's cancer)/THEME was cured by (chemotherapy)/AGENT.

   (k) Cathy dislikes country music.
   (Cathy)/AGENT dislikes (country music)/THEME.

   (l) Skiing competitions will be held in northern Utah.
   (Skiing competitions)/AGENT will be held in (northern Utah)/AT-LOC.

3. (12 pts) Consider the (tiny!) text corpus below, which contains 5 documents (D1-D5). Each document consists of 2 sentences, so the entire corpus contains 10 sentences (S1-S10). For all computations below, treat the words as case-insensitive (e.g., "Fish" is the same as "fish"). Use log base 2 for the computations. **Show all your work!**

D1	S1: Nearly all birds can fly. S2: Penguins are water birds but can not fly.
D2	S3: Penguins eat fish and swim in the ocean. S4: Pelicans are birds that also eat fish and fly.
D3	S5: Pelicans and penguins eat ocean fish. S6: Fish swim in many rivers.
D4	S7: Fish can swim in both the ocean and rivers. S8: Most water birds eat fish in the ocean or rivers.
D5	S9: Some birds can swim, such as King penguins and Emperor penguins. S10: Penguins swim far into the ocean to catch deep sea squid.

(a) Compute TF-IDF(penguins, D5)

$TF - IDF = TF(penguins) * IDF(penguins, D5)$

$TF(penguins, D5) = 3$

$IDF(penguins) = log_2 \frac{5}{4} = 0.3219$

$TF - IDF = 3 * 0.3219 = 0.9657$

(b) Compute TF-IDF(fish, D2) $TF - IDF = TF(fish) * IDF(fish, D2)$

$TF(fish, D2) = 2$

$IDF(fish) = log_2 \frac{5}{3} = 0.7369$

$TF - IDF = 2 * 0.7369 = 1.4738$

Use the following definitions for the PMI computations below. Let x and y be terms in the vocabulary. Define P(x) as the probability that a sentence will contain at least one instance of x. Define P(x,y) as the probability that a sentence will contain at least one instance of x and at least one instance of y.

(c) Compute PMI(penguins,fish)

$PMI(penguins, fish) = log_2 \frac{P(penguins, fish)}{P(penguins) * P(fish)}$

$P(penguins, fish) = \frac{2}{10} = 0.2$

$P(penguins) = \frac{5}{10} = 0.5$

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

$PMI(penguins, fish) = log_2 \frac{0.2}{0.5 * 0.6} = -0.5849$

(d) Compute PMI(birds, fly)

$PMI(birds, fly) = log_2 \frac{P(birds, fly)}{P(birds) * P(fly)}$

$P(birds, fly) = \frac{3}{10} = 0.3$

$P(birds) = \frac{5}{10} = 0.5$

$P(fly) = \frac{3}{10} = 0.3$

$$PMI(birds, fly) = log_2 \frac{0.3}{0.5 * 0.3} = 1$$

4. (8 pts) Consider the following two vectors:

$\mathbf{X}$ = <7, 4, 11, 2, 9>
$\mathbf{Y}$ = <1, 8, 10, 3, 6>

Show all of your work for the computations below.

(a) Compute the similarity between $\mathbf{X}$ and $\mathbf{Y}$ using Manhattan Distance.
$|7-1|+|4-8|+|11-10|+|2-3|+|9-6| = 15$

(b) Compute the similarity between $\mathbf{X}$ and $\mathbf{Y}$ using Jaccard Similarity.
$(7+4+11+2+9)/(1+8+10+3+6) = \frac{33}{28}$

(c) Compute the similarity between $\mathbf{X}$ and $\mathbf{Y}$ using Cosine Similarity.
$\frac{(7*1)+(4*8)+(11*10)+(2*3)+(9*6)}{\sqrt{(7+4+11+2+9)}*\sqrt{(1+8+10+3+6)}} = 6.8755$

5. (12 pts) Calculate the following values for functions commonly used by neural networks. Show all your work!

   (a) Consider a neural network with no hidden units, where X is the input vector and Y is a single output node. Let X = <.4 .3 .7>, the weight vector W = <.8 .2 .1>, and the bias weight b = .6. Compute the output value for Y using the sigmoid activation function.
   $Y = \frac{1}{1+e^{-DOT(W,X)+b}}$
   $DOT(W, X) = 0.45$
   $Y = \frac{1}{1+e^{-.45+.6}} = 0.4625$

   (b) Consider a neural network with no hidden units, where X is the input vector and Y is a single output node. Let X = <.2 .6 .1>, the weight vector W = <.5 .4 .8>, and the bias weight b = .3. Compute the output value for Y using the tanh activation function.
   $Y = \frac{e^z - e^{-z}}{e^z + e^{-z}}$
   $z = DOT(W, X) + b = 0.42 + 0.3 = 0.72$
   $Y = \frac{e^{.72} - e^{-.72}}{e^{0.72} + e^{-0.72}} = 0.6169$

   (c) Consider the following vector Z = <4.2 0.3 2.1 .75 .99>. Compute softmax(Z).
   $softmax(Z_1) = e^{4.2}/(e^{0.3} + e^{2.1} + e^{.75} + e^{.99}) = 4.65548$
   $softmax(Z_2) = e^{0.3}/(e^{4.2} + e^{2.1} + e^{.75} + e^{.99}) = 0.01694$
   $softmax(Z_3) = e^{2.1}/(e^{4.2} + e^{0.3} + e^{.75} + e^{.99}) = 0.11210$
   $softmax(Z_4) = e^{.75}/(e^{4.2} + e^{0.3} + e^{2.1} + e^{.99}) = 0.02683$
   $softmax(Z_5) = e^{.99}/(e^{4.2} + e^{0.3} + e^{2.1} + e^{.75}) = 0.03436$

6. (13 pts) The questions below pertain to the LINKER event extraction system [Huang and Riloff, 2012]. Consider the following story with five sentences:

S1: A small boy set the local market on fire.
S2: Police quickly found and arrested the young boy.
S3: The cops charged the suspect with arson.
S4: The young suspect will appear in court tomorrow.
S5: A plea is expected by the suspect in court.

(a) Show the lexical bridge features that LINKER would create for the sentence pair S2+S3.
<Police, the cops> <found and arrested, charged> <the young boy, suspect>

(b) Show the discourse focus features that the LINKER event extraction system would produce for the sentence pair S1+S2.
<boy, object, subject>

(c) Show the discourse focus features that the LINKER event extraction system would produce for the sentence pair S4+S5.
<suspect, PP(by), subject>

Consider the following new story about an assassination that has six sentences:

S1: The mayor of Atlantis was assassinated by gunmen.
S2: Walter Roswell was shot yesterday.
S3: Atlantis has been plagued by violence and the deaths of many innocent people over the past decade.
S4: If more assassinations of politicians occur this year, lawmakers have threatened to end public access to city hall.
S5: The killing took place in front of several staff members.
S6: Three armed men entered city hall with rifles and shot Mayor Roswell.

(d) Assume that the underlined phrases in the story above were labeled as *candidate role fillers* for the Victim role by the role filler extractors. If the sentences S1, S2, S5, and S6 are considered to be "relevant" sentences, list the phrases that would ultimately be extracted by LINKER as Victims.

<Victim> was assassinated <Victim> was shot in front of <Victim>

Question #7 is for CS-6340 students ONLY!

7. (10 pts) Suppose a word sense disambiguation (WSD) system is trying to distinguish between 3 senses of the word "**bug**": $sense_A$ refers to an insect, $sense_B$ refers to a software error, and $sense_C$ refers to an obsession. Assume that a WSD system has disambiguated some instances of **bug** in the documents below, labeling them as sense A, B, or C. (Note that these labels may not be correct!)

   DOC #1
   The computer **bug/B** is probably in the print function.
   The **bug/B** is very subtle and may be hard to fix.
   But we will fix the **bug/?** by the end of the day!

   DOC #2
   The **bug/A** flew around the room and landed on a window sill.
   John hit the **bug/C** with a fly swatter.
   Miraculously, the **bug/?** survived!

   DOC #3
   A purple **bug/A** was sitting on a flower.
   The **bug/B** was very colorful and pretty!
   Mary suspected that the **bug/A** was an exotic type of beetle.
   Before she could check her book, the **bug/?** flew away!

   DOC #4
   George had been bitten by the writing **bug/C** and hoped to be a novelist!
   Last night when he was writing a book, a **bug/A** crawled across the page.
   He took the book outside and released the **bug/A**.
   The **bug/A** flew away and he resumed writing for 12 hours!
   George truly had a writing **bug/?** and hoped he would become famous.

In each document, one instance of "bug" is unlabeled (marked as **bug/?**). Using Yarowsky's "one sense per discourse" heuristic, indicate whether the unlabeled instance of **bug** in each document should be labeled as sense **A**, **B**, **C**, or answer **NONE** if it still cannot be labeled.

(a) What label should be assigned to **bug/?** in DOC #1?
B

(b) What label should be assigned to **bug/?** in DOC #2?
NONE

(c) What label should be assigned to **bug/?** in DOC #3?
A

(d) What label should be assigned to **bug/?** in DOC #4?
C

(e) Could Yarowsky's "one sense per discourse" heuristic be used to *change* any of the original sense labels in DOC #4? If YES, indicate which instances of **bug** should have their labels changed and briefly explain why. If not, just answer NO.
"released the **bug/A**" could be changed to /B as released a (software) bug could be a very common to have in the same sentance with the term (patch or build) release.