CS-5340/6340, Written Assignment #3
DUE: Thursday, November 5, 2015 by 11:00pm

1. (20 pts) Answer the questions below based on the Basilisk algorithm for semantic class induction, using the seed words for three semantic categories (ANIMAL, VEHICLE, and INSTRUMENT) and pattern data shown below. The table of pattern data includes four patterns and the nouns that each pattern extracted in an imaginary corpus. For logarithms, use log base 2.

   Animal Seeds: jaguar, shark, walrus, zebra
   Instrument Seeds: bass, flute, horn, violin
   Vehicle Seeds: altima, impala, mustang, prius

Pattern	Extracted Nouns
patternA	bass, bronco, dog, impala, jaguar, mustang, shark, tiger, zebra
patternB	beetle, bronco, horn, jaguar, mustang, prius, tire
patternC	bass, clarinet, flute, music, piano, sound, trumpet, violin
patternD	accord, altima, bronco, jaguar, legacy, prius, sound

   (a) Compute RlogF(patternA) for the ANIMAL category.
   Answer : RlogF(patternA) $= \frac{3}{9}.\log_2(3) = \frac{1}{3}.(1.585) = 0.5283$

   (b) Compute RlogF(patternA) for the VEHICLE category.
   Answer : RlogF(patternA) $= \frac{2}{9}.\log_2(2) = \frac{2}{9}.(1) = \frac{2}{9} = 0.22$

   (c) Compute RlogF(patternA) for the INSTRUMENT category.
   Answer : RlogF(patternA) $= \frac{1}{9}.\log_2(1) = \frac{1}{9}.(0) = 0$

   (d) Compute RlogF(patternB) for the ANIMAL category.
   Answer : RlogF(patternB) $= \frac{1}{7}.\log_2(1) = \frac{1}{7}.(0) = 0$

   (e) Compute RlogF(patternB) for the VEHICLE category.
   Answer : RlogF(patternB) $= \frac{2}{7}.\log_2(2) = \frac{2}{7}.(1) = 0.2857$

   (f) Compute RlogF(patternB) for the INSTRUMENT category.
   Answer : RlogF(patternB) $= \frac{1}{7}.\log_2(1) = \frac{1}{7}.(0) = 0$

   (g) Compute AvgLog("bronco") for the ANIMAL category.
   Answer : Avglog("bronco") $= \frac{\log_2(3+1)+\log_2(1+1)+\log_2(1+1)}{3}$
   $= \frac{\log_2(4)+\log_2(2)+\log_2(2)}{3} = \frac{2+1+1}{3} = \frac{4}{3} = 1.333$

   (h) Compute AvgLog("bronco") for the VEHICLE category.
   Answer : Avglog("bronco") $= \frac{\log_2(2+1)+\log_2(2+1)+\log_2(2+1)}{3}$
   $= \frac{\log_2(3)+\log_2(3)+\log_2(3)}{3} = \frac{1.585+1.585+1.585}{3} = \frac{4.755}{3} = 1.585$

   (i) Compute AvgLog("sound") for the INSTRUMENT category.
   Answer : Avglog("sound") $= \frac{\log_2(3+1)+\log_2(0+1)}{2}$
   $= \frac{\log_2(4)+\log_2(1)}{2} = \frac{2+0}{2} = 1$

   (j) Compute AvgLog("sound") for the VEHICLE category.
   Answer : Avglog("sound") $= \frac{\log_2(0+1)+\log_2(2+1)}{2}$
   $= \frac{\log_2(1)+\log_2(3)}{2} = \frac{0+1.585}{2} = \frac{1.585}{2} = 0.7925$

2. (16 pts) Consider the following context vectors:

$word1$: <5 3 4 0 7>
$word2$: <6 8 0 2 1>
$word3$: <2 7 1 5 4>

Compute the similarity scores below using the word vectors above. Please leave your answers in fractional form!

(a) Similarity($word1$, $word2$) using Manhattan Distance.
Answer: 18

(b) Similarity($word2$, $word3$) using Manhattan Distance.
Answer: 12

(c) Similarity($word1$, $word2$) using Jaccard Similarity.
Answer: 1/3

(d) Similarity($word2$, $word3$) using Jaccard Similarity.
Answer: 1/2

(e) Similarity($word1$, $word2$) using Cosine Similarity.
Answer: $\frac{61}{\sqrt{99}\sqrt{105}} = \frac{61}{(9.95).(10.25)} = \frac{61}{101.9875} = 0.598$

(f) Similarity($word2$, $word3$) using Cosine Similarity.
Answer: $\frac{82}{\sqrt{105}\sqrt{95}} = \frac{82}{(10.25).(9.75)} = \frac{82}{99.9375} = 0.821$

3. (32 pts) This question relates to the Collins & Singer bootstrapping method for named entity recognition. The predicate Contains(w) is satisfied if a sequence of words includes the word w. TABLE 1 shows contains NP/Context pairs extracted from an imaginary text corpus, with their labels for two classes: HUMAN (HUM) and LOCATION (LOC).

TABLE 1

NP	**CONTEXT**	**CLASS**
michael jordan	nike spokesman	HUM
jordan south	nike client	HUM
jeff jordan	circuit city ceo	HUM
michael jordan	nike ceo	HUM
jeff west	ceo	HUM
south salt lake	mall in	LOC
jordan	country	LOC
south jordan	city	LOC
salt lake	capital city	LOC
west jordan	mall in	LOC

(a) (14 pts) Using the Contains(w) predicate, list all of the **NP Rules** that would be generated from the NPs in TABLE 1 and compute the probabilities P(HUM) and P(LOC) for each rule. **Leave the probabilities in fractional form!**

NP Rule	**P(HUM)**	**P(LOC)**
Michael	2/2	0/2
Jordan	4/7	3/7
South	1/3	2/3
Jeff	2/2	0/2
West	1/2	1/2
Salt	0/2	2/2
Lake	0/2	2/2

(b) (6 pts) List the NP rules that would be produced by selecting rules from the table above that would have a probability > .60. Then apply these NP rules to the instances in TABLE 2 below (i.e., fill in TABLE 2 with the class label that would be assigned to each instance). If no class would be assigned, simply put *none*.
Answer: The NP rules that have a probability > .60 are:

if Contains(Michael)	$->$	HUM	1
if Contains(Jeff)	$->$	HUM	1
if Contains(Salt)	$->$	LOC	1
if Contains(Lake)	$->$	LOC	1
if Contains(South)	$->$	LOC	0.67

TABLE 2

NP	**CONTEXT**	**CLASS**
ken jordan	south lake corp	*none*
jeff jones	west corp ceo	HUM
adam west	salt lake	*none*
michael south	ceo	HUM
south salton sea	lake	LOC
mirror lake	west	LOC

(c) (12 pts) Using the Contains(w) predicate, list all of the **Context Rules** that would be generated from the CONTEXTS in TABLE 2 and compute the probabilities P(HUM) and P(LOC) for each rule (using the class labels that you assigned). **Leave the probabilities in fractional form!**

Context Rule	**P(HUM)**	**P(LOC)**
south	-	-
lake	0/1	1/1
corp	1/1	0/1
west	1/2	1/2
ceo	2/2	0/2
salt	-	-

4. (26 pts) For each sentence below, label the head noun of each noun phrase (NP) with the thematic role that is most appropriate based on its semantic relationship with the main verb.

   (a) Jenny sold a diamond necklace with a matching diamond bracelet to the actress.
   Jenny - Agent
   necklace - Theme
   bracelet - Co-Theme
   actress - Recipient

   (b) The man repaired the broken pipe with duct tape.
   man - Agent
   pipe - Theme
   tape - Instrument

   (c) Susan lent Thomas her car on Monday.
   Susan - Agent
   Thomas - Recipient
   car - Theme
   Monday - Time

   (d) The musician played his trumpet for President Obama.
   musician - Agent
   trumpet - Theme
   Obama - Beneficiary

   (e) The girl is hiking with her sister from Logan to Pocatello.
   girl - Agent
   sister - Co-Agent
   Logan - From-Loc
   Pocatello - To-Loc

   (f) The boat sank with its ten passengers.
   boat - Theme
   passengers - Co-Theme

   (g) The bird flew along the mountain trail with its powerful wings.
   bird - Agent
   trail - Path-Loc
   wings - Instrument

   (h) The Disney movie was watched by three parents with their children.
   movie - Theme
   parents - Agent
   children - Co-Agent

5. (6 pts) Imagine that you have 5 tiny documents that each contains just a few words, which are shown below.

DOC #1: *natural language processing rules*
DOC #2: *natural food book*
DOC #3: *natural gas*
DOC #4: *natural language book*
DOC #5: *language rules book*

Using these documents, compute the Pointwise Mutual Information (PMI) values below. Each probability P(x) should be the likelihood of x occurring in a document. For example, P(*food*) means the probability that a document will contain the word *food*. You must fill in the equation as well as show the final value.

(a) PMI(*language*,*rules*)
Answer : $\log_2\left(\frac{\mathcal{P}(f,w)}{\mathcal{P}(f).\mathcal{P}(w)}\right) = \log_2\left(\frac{\mathcal{P}(language,rules)}{\mathcal{P}(language).\mathcal{P}(rules)}\right) = \log_2\left(\frac{2/5}{(3/5).(2/5)}\right) = \log_2\left(\frac{2/5}{6/25}\right) = \log_2\left(\frac{5}{3}\right) = \log_2(5) - \log_2(3) = 2.322 - 1.585 = 0.737$

(b) PMI(*natural*,*book*)
Answer : $\log_2\left(\frac{\mathcal{P}(f,w)}{\mathcal{P}(f).\mathcal{P}(w)}\right) = \log_2\left(\frac{\mathcal{P}(natural,book)}{\mathcal{P}(natural).\mathcal{P}(book)}\right) = \log_2\left(\frac{2/5}{(4/5).(3/5)}\right) = \log_2\left(\frac{2/5}{12/25}\right) = \log_2\left(\frac{5}{6}\right) = \log_2(5) - \log_2(6) = 2.322 - 2.585 = -0.263$

ELECTRONIC SUBMISSION INSTRUCTIONS
(a.k.a. "What to turn in and how to do it")

Your written assignment must be in .pdf format. Please do not turn in .doc or .docx files ... convert them to .pdf format before submitting them!

To submit this assignment, the CADE provides a web-based facility for electronic handin, which can be found here:

https://webhandin.eng.utah.edu/

Or you can log in to any of the CADE machines and issue the command:

handin cs5340 written3 <filename>

Please name your file: YourName-written3.pdf (e.g., EllenRiloff-written3.pdf)

HELPFUL HINT: you can get a listing of the files that you've already turned in via electronic submission by using the 'handin' command without giving it a filename. For example:

handin cs5340 written3

will list all of the files that you've turned in thus far. If you submit a new file with the same name as a previous file, the new file will overwrite the old one.