CS-5340/6340, Solutions to Written Assignment #3

1. (28 pts) This question is about the Basilisk algorithm for semantic lexicon induction. Use the following seed words for the ANIMAL and HUMAN semantic classes:

   ANIMAL: bird, cat, dog, rat, snake
   HUMAN: boy, girl, person, man, student

   Consider the following contextual patterns paired with the words that they extract (i.e., co-occur with) in an imaginary text corpus:

Pattern	**Words**
SUBJ_climbed	*bear, boy, cat, cougar, monkey, squirrel*
SUBJ_ran	*bear, boy, cat, deer, dog, girl, man, mouse, squirrel, woman*
SUBJ_ate	*bird, boy, cat, dog, girl, man, owl, snake, woman*
SUBJ_flew	*bat, bird, canary, finch, hawk, owl, parrot, sparrow*
SUBJ_nested	*bird, finch, hawk, owl, parrot, sparrow, squirrel*
admired_DOBJ	*cougar, eagle, mother, father, hero*
caught_DOBJ	*bird, boy, cold, mouse, rat, snake, sparrow, squirrel*
hunted_DOBJ	*bear, deer, cougar*
treed_DOBJ	*bear, cougar*
invited_DOBJ	*boy, daughter, girl, lady, man, son, student, woman*
praised_DOBJ	*boy, daughter, dog, girl, son, student*
scared_by_NP	*bear, cougar, dog, person, rat, shark, snake, spider, thunder*
cage_for_NP	*bird, canary, finch, parrot, rat, snake*
wings_of_NP	*bat, bird, finch, hawk, owl, parrot, sparrow*

   For the RlogF score, assume that the logarithm always returns a value of at least 1. That is, RlogF = $\frac{F_i}{N_i} * log_2(F_i)$, UNLESS F=0 or F=1 in which case RlogF = $\frac{F_i}{N_i}$.

   (a) (8 pts) Using the seed words above, compute Basilisk's RlogF score for the ANIMAL class for each pattern below.

   - SUBJ_ran

     RlogF = 2/10 * log(2)

   - caught_DOBJ

     RlogF = 3/8 * log(3)

   - invited_DOBJ

     RlogF = 0/8 * log(0)

- cage_for_NP

  RlogF = 2/6 * log(2)

(b) (8 pts) Using the seed words above, compute Basilisk's RlogF score for the HUMAN class for each pattern below.

- SUBJ_ran

  RlogF = 3/10 * log(3)

- caught_DOBJ

  RlogF = 1/8 * log(1)

- invited_DOBJ

  RlogF = 4/8 * log(4)

- cage_for_NP

  RlogF = 0/6 * log(0)

(c) (6 pts) Assume that all patterns shown in the table are in Basilisk's pattern pool. Compute the AvgLog score for the following words for the ANIMAL class. Use the true log_2 value for this computation. **Please show your work!**

- *cougar*

  AvgLog = (log(1+1) + log(0+1) + log(3+1) + log(0+1) + log(0+1))/5 =
  (1+0+2+0+0)/5 = 3/5 = .60

- *squirrel*

  AvgLog = (log(1+1) + log(2+1) + log(1+1) + log(3+1))/4 =
  (1+1.585+1+2)/4 = 1.396

(d) (6 pts) Assume that all patterns shown in the table are in Basilisk's pattern pool. Compute the AvgLog score for the following words for the HUMAN class. Use the true log_2 value for this computation. **Please show your work!**

- *squirrel*

  AvgLog = (log(1+1) + log(3+1) + log(0+1) + log(1+1))/4 =
  (1+2+0+1)/4 = 1.0

- *woman*

  AvgLog = (log(3+1) + log(3+1) + log(4+1))/3 =
  (2+2+2.32)/3 = 2.11

2. (32 pts) This question relates to the Collins & Singer bootstrapping method for named entity recognition. The predicate Contains(w) is satisfied if an NP or Context includes the word w. Treat all words as being case-insensitive. For example, "city", "City", and "CITY" should all be considered to be the same word. TABLE 1 contains NP/Context pairs extracted from an imaginary text corpus.

TABLE 1

NP	CONTEXT	CLASS
Michael Jordan	Nike spokesman	PERSON
Jordan South	Nike client	PERSON
Jeff Jordan	Circuit City CEO	PERSON
Michael Jordan	Nike CEO	PERSON
Jeff West	CEO	PERSON
South Salt Lake	mall in	LOCATION
Jordan	country	LOCATION
South Jordan	city	LOCATION
Salt Lake	capital city	LOCATION
West Jordan	mall in	LOCATION

(a) (15 pts) Using the Contains(w) predicate, fill in the table below with all of the rules that would be generated from the NPs in TABLE 1 and compute the probability of each rule. **Leave the probabilities in fractional form!** You do not need to show rules that would have a probability of zero.

NP Rule	Probability
If Contains("Michael") → PERSON	2/2
If Contains("Jordan") → PERSON	4/7
If Contains("Jordan") → LOCATION	3/7
If Contains("Jeff") → PERSON	2/2
If Contains("West") → PERSON	1/2
If Contains("South") → PERSON	1/3
If Contains("South") → LOCATION	2/3
If Contains("Salt") → LOCATION	2/2
If Contains("Lake") → LOCATION	2/2
If Contains("West") → LOCATION	1/2

(b) (5 pts) Using only the NP rules in your previous table that have a probability > .50, apply those rules to the instances in TABLE 2. Fill in TABLE 2 with the class that would be assigned to each instance using those rules. If more than one rule applies, use the rule that has the highest probability. If no class would be assigned to an instance in TABLE 2, simply put *none*.

TABLE 2

NP	**CONTEXT**	**CLASS**
Ken Jordan	spokesman	**person**
Jeff Jones	City Mall CEO	**person**
Adam West	actor	**none**
Michael Williams Legal Consulting	firm	**person**
South Korea	country	**location**

(c) (12 pts) Using the Contains(w) predicate, fill in the table below with all of the **Context Rules** that would be generated from the instances in both TABLE 1 and TABLE 2 and compute the probability of each context rule based on the instances in both TABLE 1 and TABLE 2. **Leave the probabilities in fractional form!** You do not need to show rules that would have a probability of zero.

Context Rule	**Probability**
If Contains("spokesman") → PERSON	2/2
If Contains("CEO") → PERSON	4/4
If Contains("firm") → PERSON	1/1
If Contains("country") → LOCATION	2/2
If Contains("Nike") → PERSON	3/3
If Contains("client") → PERSON	1/1
If Contains("Circuit") → PERSON	1/1
If Contains("city") → PERSON	2/4
If Contains("city") → LOCATION	2/4
If Contains("capital") → LOCATION	1/1
If Contains("mall") → LOCATION	2/3
If Contains("mall") → PERSON	1/3
If Contains("in") → LOCATION	2/2

3. (22 pts) Consider the following quote from Dr. Seuss' *The Cat in the Hat*:

look at me
look at me
look at me now
it is fun to have fun
but you have
to know how

For this question, you will be using feature vectors that include a feature for every word w_i in the quote. Each feature vector should be of this form:

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
at	but	fun	have	how	is	it	know	look	me	now	to	you

(a) Create a co-occurrence feature vector for each word below based on the words that occur within 5 words to the left or 5 words to the right of the given word in the Dr. Seuss quote. For each word, you should count instances of the same word that occur within the 5-word window, but do not count the instance itself. For example, given "a b a", the 'b' has 0 b's in its context window and each 'a' has 1 'a' in its context window.

- Create a feature vector for the word "fun" using binary feature values.

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
at	but	fun	have	how	is	it	know	look	me	now	to	you
1	1	1	1	0	1	1	1	0	1	1	1	1

- Create a feature vector for the word "fun" using frequency feature values.

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
at	but	fun	have	how	is	it	know	look	me	now	to	you
1	2	2	3	0	2	2	1	0	1	1	3	2

- Create a feature vector for the word "look" using binary feature values.

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
at	but	fun	have	how	is	it	know	look	me	now	to	you
1	0	0	0	0	1	1	0	1	1	1	0	0

- Create a feature vector for the word "look" using frequency feature values.

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
at	but	fun	have	how	is	it	know	look	me	now	to	you
8	0	0	0	0	1	1	0	4	8	1	0	0

(b) Compute the similarity of the two feature vectors below using the specified similarity metric. *Please show all your work when computing the similarity score!*

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
	at	but	fun	have	how	is	it	know	look	me	now	to	you
X	2	8	9	3	4	0	0	6	1	9	1	2	5
Y	1	6	2	0	3	1	8	7	4	5	3	0	3

- Compute the similarity between vectors **X** and **Y** using Manhattan Distance.

  $1 + 2 + 7 + 3 + 1 + 1 + 8 + 1 + 3 + 4 + 2 + 2 + 2 = 37$

- Compute the similarity between vectors **X** and **Y** using the Jaccard metric.

  $\frac{1+6+2+0+3+0+0+6+1+5+1+0+3}{2+8+9+3+4+1+8+7+4+9+3+2+5} = \frac{28}{65} = .43$

- Compute the similarity between vectors **X** and **Y** using the Cosine metric.

  $\frac{2+48+18+0+12+0+0+42+4+45+3+0+15}{sqrt(4+64+81+9+16+0+0+36+1+81+1+4+25)*sqrt(1+36+4+0+9+1+64+49+16+25+9+0+9)} =$

  189/(sqrt(322)*sqrt(223)) = 189/(17.9*14.9) = .71

4. (18 pts) Consider the following short story:

> Samantha and Jake went to the grocery store. They used a basket and filled it with fruit, ice cream, and bread. On the way to the checkout counter, she knocked the shopping basket into a display case. The handle on the plastic container broke and their food rolled onto the floor. Sam picked up their groceries off the tile flooring, while Jake complained to the manager of the store about the broken handle. He offered an apology. But Jake thought his response was insincere, so he told him and vowed never to shop at this market again.

List all noun phrases (NPs) that are coreferent with each phrase below. If the same NP appears multiple times in the story, then you should list it multiple times (i.e., list every instance that is coreferent).

(a) Samantha and Jake

They, *their**, *their**

(b) Samantha

Sam, *she*

(c) the manager

He, *his**, *him*

(d) a basket

it, *the shopping basket*, *the plastic container*

(e) fruit, ice cream, and bread

their food, *their groceries*

(f) the grocery store

the store, *this market*

(g) an apology

his response

**NOTE: "their" and "his" are <u>not</u> noun phrases themselves, they are possessive pronouns that occur within a noun phrase in the story above. However, they are coreferent with earlier entities in the story. I did intend for you to include them in your answer, but unfortunately*

I forgot to include possessive pronouns in the instructions! Consequently, we did not deduct points if you omitted them since technically they are not NPs.