**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.

   RlogF($pattern_i$)= $\frac{F_i}{N_i} * log_2(F_i)$
   =$\frac{3}{9} * log_2(3)$
   =$0.33 * 1.584$
   = 0.52303

   (b) Compute RlogF(patternA) for the VEHICLE category.

   RlogF($pattern_i$)= $\frac{F_i}{N_i} * log_2(F_i)$
   =$\frac{2}{9} * log_2(2)$
   =$0.2222 * 1$
   = 0.2222

   (c) Compute RlogF(patternA) for the INSTRUMENT category.

   RlogF($pattern_i$)= $\frac{F_i}{N_i} * log_2(F_i)$
   =$\frac{1}{9} * log_2(1)$
   =$0.11111 * 0$
   = 0

   (d) Compute RlogF(patternB) for the ANIMAL category.

   RlogF($pattern_i$)= $\frac{F_i}{N_i} * log_2(F_i)$
   =$\frac{1}{7} * log_2(1)$

$=0.14285 * 0$
$= 0$

(e) Compute RlogF(patternB) for the VEHICLE category.

RlogF($pattern_i$)= $\frac{F_i}{N_i} * log_2(F_i)$
$=\frac{2}{7} * log_2(2)$
$=0.2857 * 1$
$= 0.2857142857$

(f) Compute RlogF(patternB) for the INSTRUMENT category.

RlogF($pattern_i$)= $\frac{F_i}{N_i} * log_2(F_i)$
$=\frac{1}{7} * log_2(1)$
$=0.14285 * 0$
$= 0$

(g) Compute AvgLog("bronco") for the ANIMAL category.

AvgLog($word_i$)= $\frac{\sum_{j=1}^{N_i} log_2(F_j+1)}{N_i}$
$=\frac{log_2(3+1)+log_2(1+1)+log_2(1+1)}{3}$
$=\frac{2+1+1}{3}$
$=1.33333$

(h) Compute AvgLog("bronco") for the VEHICLE category.

AvgLog($word_i$)= $\frac{\sum_{j=1}^{N_i} log_2(F_j+1)}{N_i}$
$=\frac{log_2(2+1)+log_2(2+1)+log_2(2+1)}{3}$
$=\frac{3log_2(3)}{3}$
$=1.584963$

(i) Compute AvgLog("sound") for the INSTRUMENT category.

AvgLog($word_i$)= $\frac{\sum_{j=1}^{N_i} log_2(F_j+1)}{N_i}$
$=\frac{log_2(3+1)+log_2(0+1)}{2}$
$=\frac{log_2(4)}{2}$
$=1$

(j) Compute AvgLog("sound") for the VEHICLE category.

$\text{AvgLog}(word_i) = \frac{\sum_{j=1}^{N_i} log_2(F_j+1)}{N_i}$
$= \frac{log_2(0+1)+log_2(2+1)}{2}$
$= \frac{log_2(3)}{2}$
$= 0.7924815$

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.

$\text{ManhattanDistance(X,Y)} = \sum_{i=1}^{N} |x_i - y_i|$

$\text{Similarity}(word1, word2) = |(5-6) + (3-8) + (4-0) + (0-2) + (7-1)|$

$= 18$

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

$\text{ManhattanDistance(X,Y)} = \sum_{i=1}^{N} |x_i - y_i|$

$\text{Similarity}(word2, word3) = |(6-2) + (8-7) + (0-1) + (2-5) + (1-4)|$

$= 12$

(c) Similarity($word1$, $word2$) using Jaccard Similarity.

$\text{Jaccard(X,Y)} = \frac{\sum_{i=1}^{N} min(x_i, y_i)}{\sum_{i=1}^{N} max(x_i, y_i)}$

$\text{Similarity}(word1, word2) = \frac{5+3+0+0+1}{6+8+4+2+7}$

$= \frac{9}{27}$

$= \frac{1}{3}$

(d) Similarity($word2$, $word3$) using Jaccard Similarity.

$$\text{Jaccard(X,Y)}= \frac{\sum_{i=1}^{N} min(x_i,y_i)}{\sum_{i=1}^{N} max(x_i,y_i)}$$

Similarity($word2$, $word3$)$=\frac{2+7+0+2+1}{6+8+1+5+4}$

$=\frac{12}{24}$

$=\frac{1}{2}$

(e) Similarity($word1$, $word2$) using Cosine Similarity.

$$\text{Cosine(X,Y)}= \frac{\sum_{i=1}^{N} (x_i * y_i)}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}$$

Similarity($word1$, $word2$)$=\frac{5*6+3*8+4*0+0*2+7*1}{\sqrt{25+9+16+0+49}\sqrt{36+64+0+4+1}}$

$=\frac{61}{\sqrt{99}\sqrt{105}}$

$=0.5982980511087747$

(f) Similarity($word2$, $word3$) using Cosine Similarity.

$$\text{Cosine(X,Y)}= \frac{\sum_{i=1}^{N} (x_i * y_i)}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}$$

Similarity($word2$, $word3$)$=\frac{6*2+8*7+0*1+2*5+1*4}{\sqrt{36+64+0+4+1}\sqrt{4+49+1+25+16}}$

$=\frac{82}{\sqrt{105}\sqrt{95}}$

$=0.82102692$

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*.

NP rules that would be produced by selecting rules from the table above that would have a probability > 0.60

| **NP Rule** | **P(HUM)** | **P(LOC)** |
|---|---|---|
| michael | 2/2 | |
| south | | 2/3 |
| jeff | 2/2 | |
| salt | | 2/2 |
| lake | | 2/2 |

**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 |

We assign the class of the NP 'michael south' as HUM because the probability of micheal being human (1) is greater than the probability of south being a location (0.66)

(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 sold a diamond (necklace) / THEME with a matching (bracelet) / CO-THEME to the (actress) / RECIPIENT

   (b) The man repaired the broken pipe with duct tape.

   The (man) / AGENT repaired the broken (pipe) /THEME with duct (tape) / INSTRUMENT

   (c) Susan lent Thomas her car on Monday.

   (Susan) / AGENT lent (Thomas) / RECIPIENT her (car) / THEME on (Monday) / TIME

   (d) The musician played his trumpet for President Obama.

   The (musician) / AGENT played his (trumpet) / THEME for President (Obama) / BENEFICIARY

   (e) The girl is hiking with her sister from Logan to Pocatello.

   The (girl) / AGENT is hiking with her (sister) /CO-AGENT from (Logan) SOURCE/ ORIGIN/ FROM-LOC to (Pocatello) DESTINATION /TO -LOC

   (f) The boat sank with its ten passengers.

   The (boat) / THEME sank with its ten (passengers)/ CO-THEME

   (g) The bird flew along the mountain trail with its powerful wings.

   The (bird) / AGENT flew along the mountain (trail) / PATH-LOC with its powerful (wings) /INSTRUMENT

   (h) The Disney movie was watched by three parents with their children.

   The Disney (Movie) / THEME was watched by three (parents) / AGENT with their (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*)
PMI(f,w) = $log_2(P(f,w)/(P(w) * P(f)))$

$=log_2(2/5/(2/5 * 3/5))$

$=log_2(5/3)$

=0.7369078852103839

(b) PMI(*natural*,*book*)
PMI(f,w) = $log_2(P(f,w)/(P(w) * P(f)))$

$=log_2(2/5/(3/5 * 4/5))$

$=log_2(5/6)$

= -0.2630344058337938