

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

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.

$$3/9 * \log(3) = .53$$

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

$$2/9 * \log(2) = .22$$

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

$$1/9 * \log(1) = 0$$

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

$$1/7 * \log(1) = 0$$

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

$$2/7 * \log(2) = .29$$

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

$$1/7 * \log(1) = 0$$

(g) Compute $\text{AvgLog}(\text{"bronco"})$ for the ANIMAL category.

$$(\log(4) + \log(2) + \log(2))/3$$

(h) Compute $\text{AvgLog}(\text{"bronco"})$ for the VEHICLE category.

$$(\log(3) + \log(3) + \log(3))/3$$

(i) Compute $\text{AvgLog}(\text{"sound"})$ for the INSTRUMENT category.

$$(\log(4) + \log(1))/2$$

(j) Compute $\text{AvgLog}(\text{"sound"})$ for the VEHICLE category.

$$(\log(1) + \log(3))/2$$

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.

$$1 + 5 + 4 + 2 + 6 = 18$$

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

$$4 + 1 + 1 + 3 + 3 = 12$$

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

$$(5 + 3 + 0 + 0 + 1) / (6 + 8 + 4 + 2 + 7) = 9/27 = .33$$

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

$$(2 + 7 + 0 + 2 + 1) / (6 + 8 + 1 + 5 + 4) = 12/24 = .50$$

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

$$\begin{aligned} & \frac{(5*6)+(3*8)+(4*0)+(0*2)+(7*1)}{\sqrt{(5^2+3^2+4^2+0^2+7^2)}*\sqrt{(6^2+8^2+0^2+2^2+1^2)}} \\ &= \frac{30+24+0+0+7}{\sqrt{(25+9+16+0+49)}*\sqrt{(36+64+0+4+1)}} \\ &= \frac{61}{\sqrt{(99)}*\sqrt{(105)}} = .60 \end{aligned}$$

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

$$\begin{aligned} & \frac{(6*2)+(8*7)+(0*1)+(2*5)+(1*4)}{\sqrt{(6^2+8^2+0^2+2^2+1^2)}*\sqrt{(2^2+7^2+1^2+5^2+4^2)}} \\ &= \frac{12+56+0+10+4}{\sqrt{(36+64+0+4+1)}*\sqrt{(4+49+1+25+16)}} \\ &= \frac{82}{\sqrt{(105)}*\sqrt{(95)}} = .82 \end{aligned}$$

3. (32 pts) This question relates to the Collins & Singer bootstrapping method for named entity recognition. The predicate $\text{Contains}(w)$ 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 $\text{Contains}(w)$ predicate, list all of the **NP Rules** that would be generated from the NPs in TABLE 1 and compute the probabilities $P(\text{HUM})$ and $P(\text{LOC})$ for each rule. **Leave the probabilities in fractional form!**

NP Rule	$P(\text{HUM})$	$P(\text{LOC})$
Contains("michael")	2/2	0/2
Contains("jordan")	4/7	3/7
Contains("south")	1/3	2/3
Contains("jeff")	2/2	0/2
Contains("west")	1/2	1/2
Contains("salt")	0/2	2/2
Contains("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 for probability $> .60$:

If NP Contains("michael") \rightarrow HUM

If NP Contains("south") \rightarrow LOC

If NP Contains("jeff") \rightarrow HUM

If NP Contains("salt") \rightarrow LOC

If NP Contains("lake") \rightarrow LOC

TABLE 2

NP	CONTEXT	CLASS
ken jordan	south lake corp	NONE
jeff jones	west corp ceo	HUM
adam west	salt lake	NONE
michael smith	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 $P(\text{HUM})$ and $P(\text{LOC})$ for each rule (using the class labels that you assigned). **Leave the probabilities in fractional form!**

Context Rule	$P(\text{HUM})$	$P(\text{LOC})$
Contains("south")	—	—
Contains("lake")	0/1	1/1
Contains("corp")	1/1	0/1
Contains("west")	1/2	1/2
Contains("ceo")	2/2	0/2
Contains("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 matching diamond 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/AT-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/ORIGIN to Pocatello/DESTINATION.

- (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, $P(\text{food})$ 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) $\text{PMI}(\text{language}, \text{rules})$

$$\begin{aligned}\text{PMI} &= \log\left(\frac{P(\text{language}, \text{rules})}{P(\text{language}) * P(\text{rules})}\right) \\ &= \log\left(\frac{2/5}{(3/5) * (2/5)}\right) = \log(5/3) = .74\end{aligned}$$

- (b) $\text{PMI}(\text{natural}, \text{book})$

$$\begin{aligned}\text{PMI} &= \log\left(\frac{P(\text{natural}, \text{book})}{P(\text{natural}) * P(\text{book})}\right) \\ &= \log\left(\frac{2/5}{(4/5) * (3/5)}\right) = \log(5/6) = -.26\end{aligned}$$