Justin Keeling
CSCI-491
WA-4
3 December 2019

1)

a.

bat is polysemous due to the fact it has related meanings **An example shows as follows:**

- "bat the ball" referring to a verb sense of striking a ball.
- "at bat" referring to a noun sense. In baseball, *at bat* means : "a turn trying to hit the ball".

b.

meat is homonymous because even in the same form, *meat* has unrelated meanings. **An example shows as follows:**

- "meat", as in; "the most important part of some idea or experience".

- "meat", as in; "the flesh of animals used as food".

c.

"big" and "large" are not synonymous in all contexts because "big" can be more of a reference "age" rather than large which is a reference to "size". **An example shows as follows:**

- "big brother" vs "large brother". *Big brother* is typically thought of in the context of "older". On the other hand, *large brother* would be more towards the reference of "size".

d.

Hyponymy explained by an example*: Country* would have a hyponym of *France*. The France is more specific as compared to country, which is a more broad sense.

Hypernym explained by an example: *Apple* would have a hypernym of *fruit.* Apple is more specific than its hypernym, fruit. Fruit is more encompassing than Apple.

2)

- pathlen(c1,c2) = 1 + # of edges in the shortest path in the hypernym graph.

- simpath(c1,2) = 1/ path(c1,c2)

a.

- pathlen(nickel, money) = 1+ 5 = 6

- simpath(nickel,money): 1/6

b.

- pathlen(money, Richter scale) = 1 + 4 = 5

- simpath(money, Richter scale) =1/5

c.

simpath(nickel, money) < simpath(money, Richter scale) == 1/6 < 1/5

- Therefore, since the simpath(money, Richter scale) is greater than simpath(nickel, money), then simpath(money, Richter scale) is the most similar pair.

d.

- Since basic path-based similarity assumes each link represents a uniform distance, it is not as accurate as is it were to have independent costs for each edge. The uniform weights also do not account for words that are connected only through abstract nodes.

- The strange conclusion in part *c (*money is more similar to Richter scale than it is to nickle), proves that only using a uniform weight does not truly give the best results regarding similarity.

3)

NOTE: Below are the formulas for the problems in 3)

$$\text{P(c)} = \frac{\sum_{w \in words(c)} \text{count}(w)}{\text{N}}$$

- where N is the total number of words in the corpus, words(c) is all words that are decendants of node c.

IC(c) = -log(P(c))

LCS(c_1 ,c_2) = most informative node in the hierarchy

$$\text{sim}_{resnik}(\ c_1,c_2\ ) = \text{-log(P(LCS(}c_1,c_2\text{)))}$$

$$\text{sim}_{Lin}(\ c_1,c_2\ ) = \frac{\text{2log(P(LCS(}c_1,c_2\text{)))}}{\text{log(P(}c_1\text{))+ log(P(}c_2\text{))}}$$

$$\text{sim}_{jiangconrath}(\ c_1,c_2) = \frac{1}{\text{log(P(}c_1\text{))+ log(P(}c_2\text{))- 2} \times \text{log(P(LCS(}c_1,c_2\text{)))}}$$

a.

The information content of “inanimate-object” is:

- P(inanimate object) = 0.167

- IC(inanimate object) = -log(0.167) = 1.7898

b.

- LCS(hill, geological-formation) = geological-formation

c.

sim_{resnik}(hill, shore) = -log(P(LCS(hill, shore)) = -log(P(geological-formation)) = -log(0.00176) = 6.342

d.

$$\text{sim}_{Lin}(\text{ hill,shore}) = \frac{\text{2log(P(LCS(hill, shore)))}}{\text{log(P(hill)+ log(P(shore)}} = \frac{\text{2log(P(geological-formation))}}{\text{log(0.0000189)+ log(0.0000836)}} = \frac{\text{2 log(0.00176)}}{-10.876 + -9.389}$$

$$= \frac{2 \times -6.342}{-20.266} = 0.626$$

e.

$$\text{sim}_{jiangconrath}(\text{ hill, shore}) = \frac{1}{\text{log(P(hill))+ log(P(shore))- 2} \times \text{log(P(LCS(hill, shore)))}}$$

$$= \frac{1}{\text{log(0.0000189) + log(0.0000836) - 2} \times \text{ log(P(geological-formation)}}$$

$$= \frac{1}{\text{-20.266 - 2} \times \text{-6.342}} = \frac{1}{\text{-7.582}} = -0.132$$

4)

a. 1,161,192 total number of words

b. see table

c. see table

d. The weakness being that there are many paths that can be taken to an entity based on of a word. Thus, the reason why entity has more counts than there are words in the table. Which also means that the others may be inaccurate as well.

Concept	**Count**	**Probability**
entity	2592000	2.23
inanimate-object	1	0.000000861
natural-object	16057	0.0138
geological-formation	6495	0.00559
natural-elevation	2989	0.00257
shore	312	0.000269
hill	642	0.000552
coast	134	0.000115

5)

a.

- “decal”: either a design that is fixed to some surface or a paper bearing the design which is to be transferred to the surface.

- “transfer paper”: a paper that is coated with a preparation for transferring a design to another surface.

- “silverpoint”: a drawing made on specially prepared paper with an instrument having a silver tip (15th and 16th centuries).

- “drawing paper”: paper that is specially prepared for use in drafting.

b.

- “decal” , “transfer paper”

    overlapped phrases: surface, a paper, that is, a design

    respectively, similarity_score += 1 + 4 + 4 + 4 = **13**

- “decal” , “silverpoint”

    overlapped words: paper

    respectively, similarity_score += 1= **1**

- “decal” , “drawing paper”

    overlapped words: that is, paper

    respectively, similarity_score += 4 +1 **5**

- “transfer paper” , “silverpoint”

    overlapped words: paper

    respectively, similarity_score += 1 = **1**

- “transfer paper” , “drawing paper”

    overlapped words: paper that is

    respectively, similarity_score += 9 = **9**

- “silverpoint”, “drawing paper”

    overlapped words: specially prepared, paper

    respectively, similarity_score += 4 + 1 = **5**

6)

a.

“glucose” synonyms:

Anhydrous Dextrose
D-Glucose
Dextrose
Glucose Monohydrate
Glucose, (DL)-Isomer
Glucose, (L)-Isomer
Glucose, (alpha-D)-Isomer
Glucose, (beta-D)-Isomer
L-Glucose

b. Hexoses

c. Blood Glucose

d.

Path: Glucose => Hexoses => Monosaccharides => Sugars

simpath = $1/1+1+1+1=\frac{1}{4}$

7)

a.

Context Words (Rows are terms)

	he	**is**	**not**	**lazy**	**intelligent**	**smart**
he	2	5	2	2	2	1
is	5	2	2	2	2	1
not	2	2	0	1	0	0
lazy	2	2	1	0	1	0
intelligent	2	2	0	1	0	1
smart	1	1	0	0	1	0

Below: Equations Used

$$p_{ij}=\frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

$$pmi_{ij}=\log_2\left(\frac{p_{ij}}{p_i * p *_j}\right)$$

$$p_i *=\frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

$$p *_j=\frac{\sum_{i=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

$$ppmi_{ij}=pmi_{ij} \text{ if } pmi_{ij}>0 \ \ else:0$$

b. p(w=he,c=intelligent) = $\frac{2}{48}=0.042$

c. p(w=he) = $14/48=0.292$

d. p(c=intelligent)$=\frac{6}{48}=0.125$

e. $PPMI(\text{he, intelligent})=\log_2(\frac{0.042}{\frac{14}{48}\times\frac{6}{48}})=\log_2(1.271)=0.1926$

f.

PPMI Matrix

Context Words (Rows are terms)

	he	**is**	**not**	**lazy**	**intelligent**	**smart**
he	0	0.292	0.456	0.193	0.193	0.193
is	0.292	0	0.456	0.193	0.193	0.193
not	0.456	0.456	0	0.168	0	0
lazy	0.193	0.193	0.678	0	0.415	0
intelligent	0.193	0.193	0	0.415	0	1.415
smart	0.193	0.193	0	0	1.415	0

8)

a. Rounded From 5 Decimals to 2

Term	**Dim1**	**Dim2**	**Dim3**	**Dim4**	**Dim5**
dog	0.31	0.31	0.53	-0.93	-0.74
cat	0.23	0.28	0.63	-0.59	-0.59
lion	0.20	0.44	0.34	-0.31	-0.52
tiger	-0.82	0.80	0.81	-0.10	-0.19
elephant	-0.07	0.82	0.61	-0.08	-0.19
cheetah	0.46	0.79	0.33	-0.22	-0.45
monkey	0.56	0.95	0.12	-0.87	-0.54
rabbit	0.27	0.04	0.59	-0.38	-0.47
mouse	-0.09	0.05	0.26	-0.53	-0.18
rat	-0.46	0.07	0.60	-1.37	-0.56

b. 18.12

c. 13.37

d. The dot product of of all the dimensions, show that mouse and rat produce a larger number as compared to mouse and elephant. Therefore, mouse is more similar to rat than elephant.