COMS W4705: Natural Language Processing (Fall 2018)
Problem Set #4

Wenbo Gao - wg2313@columbia.edu

December 7, 2018

Problem 1

Using the raw co-occurence counts:

- Which word is the most similar to 'animal' using euclidean distance?
  'dog'.

  $\text{dis}(\text{dog}, \text{animal}) = \sqrt{(0-2)^2 + (4-3)^2 + (0-0)^2 + (4-3)^2 + (2-0)^2 + (2-3)^2} = \sqrt{11}$

  $\text{dis}(\text{cat}, \text{animal}) = \sqrt{(4-2)^2 + (0-3)^2 + (0-0)^2 + (3-3)^2 + (3-0)^2 + (10-3)^2} = \sqrt{71}$

  $\text{dis}(\text{computer}, \text{animal}) = \sqrt{(0-2)^2 + (0-3)^2 + (0-0)^2 + (5-3)^2 + (0-0)^2 + (5-3)^2} = \sqrt{21}$

  $\text{dis}(\text{run}, \text{animal}) = \sqrt{(4-2)^2 + (3-3)^2 + (5-0)^2 + (0-3)^2 + (3-0)^2 + (4-3)^2} = 4\sqrt{3}$

  $\text{dis}(\text{mouse}, \text{animal}) = \sqrt{(2-2)^2 + (10-3)^2 + (5-0)^2 + (4-3)^2 + (3-0)^2 + (0-3)^2} = \sqrt{93}$

- Which word is the most similar to 'animal' using cosine similarity?
  'run'.

$$\begin{aligned}\text{dis}(\text{dog}, \text{animal}) &= \frac{\vec{v}_{dog} \cdot \vec{v}_{animal}}{||\vec{v}_{dog}||\ ||\vec{v}_{animal}||} \\ &= \frac{0*2 + 4*3 + 0*0 + 4*3 + 2*0 + 2*3}{\sqrt{0^2+4^2+0^2+4^2+2^2+2^2}\sqrt{2^2+3^2+0^2+3^2+0^2+3^2}} \\ &= 0.851942751369\end{aligned}$$

$$\begin{aligned}\text{dis}(\text{cat}, \text{animal}) &= \frac{\vec{v}_{cat} \cdot \vec{v}_{animal}}{||\vec{v}_{cat}||\ ||\vec{v}_{animal}||} \\ &= \frac{4*2 + 0*3 + 0*0 + 3*3 + 3*0 + 10*3}{\sqrt{4^2+0^2+0^2+3^2+3^2+10^2}\sqrt{2^2+3^2+0^2+3^2+0^2+3^2}} \\ &= 0.729230142593\end{aligned}$$

$$\begin{aligned}\text{dis}(\text{computer}, \text{animal}) &= \frac{\vec{v}_{computer} \cdot \vec{v}_{animal}}{||\vec{v}_{computer}||\ ||\vec{v}_{animal}||} \\ &= \frac{0*2+0*3+0*0+5*3+0*0+5*3}{\sqrt{0^2+0^2+0^2+5^2+0^2+5^2}\sqrt{2^2+3^2+0^2+3^2+0^2+3^2}} \\ &= 0.762000762\end{aligned}$$

$$\begin{aligned}\text{dis}(\text{run}, \text{animal}) &= \frac{\vec{v}_{run} \cdot \vec{v}_{animal}}{||\vec{v}_{run}||\ ||\vec{v}_{animal}||} \\ &= \frac{4*2+3*3+5*0+0*3+3*0+4*3}{\sqrt{4^2+3^2+5^2+0^2+3^2+4^2}\sqrt{2^2+3^2+0^2+3^2+0^2+3^2}} \\ &= 0.601431982944\end{aligned}$$

$$\begin{aligned}\text{dis}(\text{mouse}, \text{animal}) &= \frac{\vec{v}_{mouse} \cdot \vec{v}_{animal}}{||\vec{v}_{mouse}||\ ||\vec{v}_{animal}||} \\ &= \frac{2*2+10*3+5*0+4*3+3*0+0*3}{\sqrt{2^2+10^2+5^2+4^2+3^2+0^2}\sqrt{2^2+3^2+0^2+3^2+0^2+3^2}} \\ &= 0.665758353436\end{aligned}$$

Problem 2

To determine if two senses of a lemma are homonyms or polysemy

- If there exists a subset of lemmas (with stop words removed) that coexists in both synsets or hyponyms of both synsets, then they are homonyms;
- Otherwise, polysemy.

Problem 3

a)

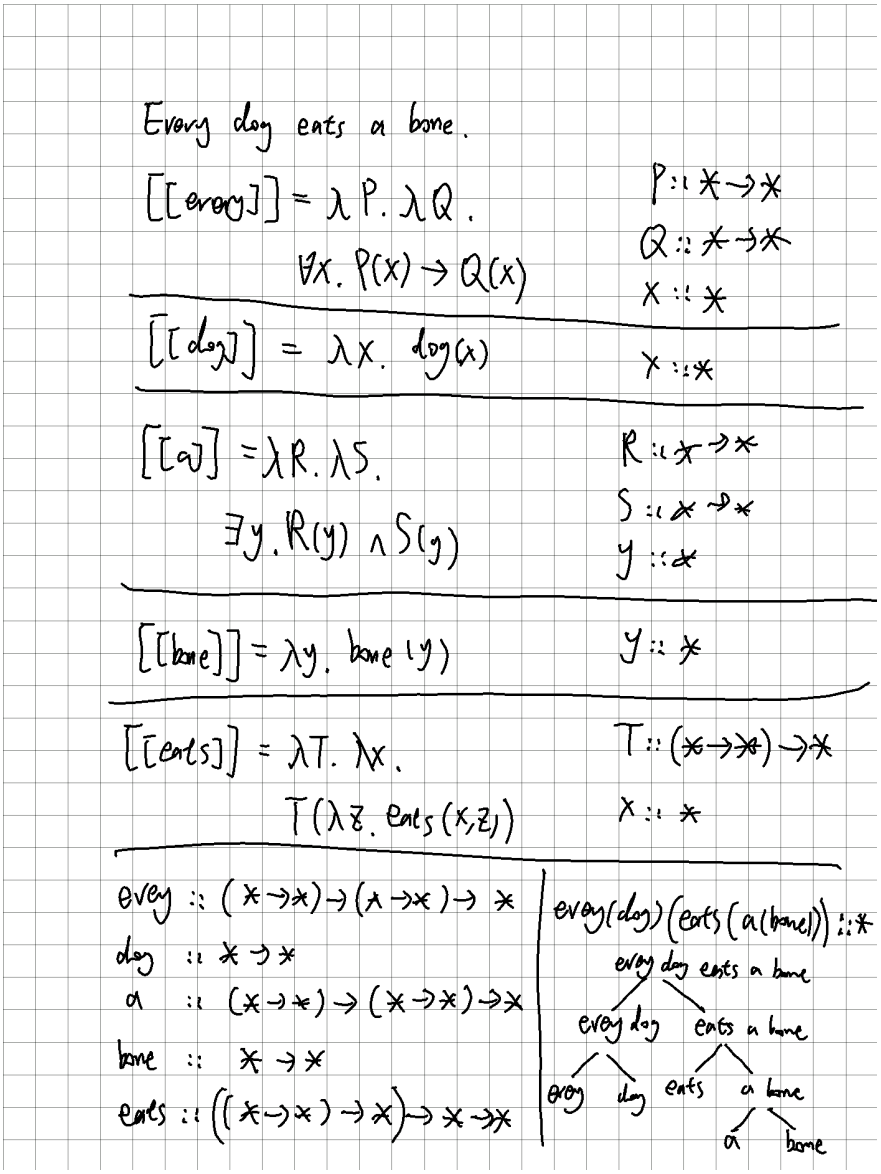

Figure 1: problem 3 (a)

b)

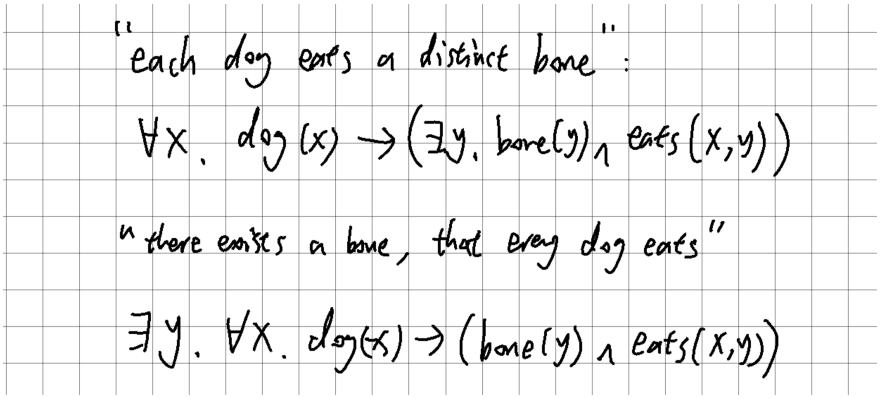

Figure 2: problem 3 (b)