COMS W4705: Natural Language Processing (sec:002) - Homework #4

Name: Geraldi Dzakwan (gd2551)

November 21, 2019

Problem 1

a. The most similar word to "animal" using Euclidean distance.

Suppose we denote the Euclidean distance between two words p and q as $dist_{euclid}(p, q)$. Since p and q are co-occurence vectors of size 6, then the Euclidean distance is defined by the equation below:

$$dist_{euclid}(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_6 - p_6)^2}$$

1. For $p = animal$ and $q = dog$:

$$dist_{euclid}(p, q) = \sqrt{(0-2)^2 + (4-3)^2 + (0-0)^2 + (4-3)^2 + (2-0)^2 + (2-3)^2}$$

$$dist_{euclid}(p, q) = \sqrt{(-2)^2 + (1)^2 + (0)^2 + (1)^2 + (2)^2 + (-1)^2}$$

$$dist_{euclid}(p, q) = \sqrt{4 + 1 + 0 + 1 + 4 + 1} = \sqrt{11} = 3.3166$$

2. For $p = animal$ and $q = cat$:

$$dist_{euclid}(p, q) = \sqrt{(4-2)^2 + (0-3)^2 + (0-0)^2 + (3-3)^2 + (3-0)^2 + (10-3)^2}$$

$$dist_{euclid}(p, q) = \sqrt{(2)^2 + (-3)^2 + (0)^2 + (0)^2 + (3)^2 + (7)^2}$$

$$dist_{euclid}(p, q) = \sqrt{4 + 9 + 0 + 0 + 9 + 49} = \sqrt{71} = 8.4261$$

3. For $p = animal$ and $q = computer$:

$$dist_{euclid}(p, q) = \sqrt{(0-2)^2 + (0-3)^2 + (0-0)^2 + (5-3)^2 + (0-0)^2 + (5-3)^2}$$

$$dist_{euclid}(p, q) = \sqrt{(2)^2 + (-3)^2 + (0)^2 + (2)^2 + (0)^2 + (2)^2}$$

$$dist_{euclid}(p, q) = \sqrt{4 + 9 + 0 + 4 + 0 + 4} = \sqrt{21} = 4.5826$$

4. For $p = animal$ and $q = run$:

$$dist_{euclid}(p,q) = \sqrt{(4-2)^2 + (3-3)^2 + (5-0)^2 + (0-3)^2 + (3-0)^2 + (4-3)^2}$$

$$dist_{euclid}(p,q) = \sqrt{(2)^2 + (0)^2 + (5)^2 + (-3)^2 + (3)^2 + (1)^2}$$

$$dist_{euclid}(p,q) = \sqrt{4+0+25+9+9+1} = \sqrt{48} = 6.9282$$

5. For $p = animal$ and $q = mouse$:

$$dist_{euclid}(p,q) = \sqrt{(2-2)^2 + (10-3)^2 + (5-0)^2 + (4-3)^2 + (3-0)^2 + (0-3)^2}$$

$$dist_{euclid}(p,q) = \sqrt{(0)^2 + (7)^2 + (5)^2 + (1)^2 + (3)^2 + (-3)^2}$$

$$dist_{euclid}(p,q) = \sqrt{0+49+25+1+9+9} = \sqrt{93} = 9.6436$$

Thus, the most similar word to "animal" based on Euclidean distance is "***dog***" because $dist_{euclid}(animal, dog)$ yields the minimum Euclidean distance: 3.3166.

b. The most similar word to "animal" using cosine similarity.

Suppose we denote the cosine similarity between two words p and q as $sim_{cos}(p,q)$. Since p and q are co-occurence vectors of size 6, then the cosine similarity is defined by the equation below:

$$sim_{cos}(p,q) = \frac{\sum_{i=1}^{6} p_i . q_i}{\sqrt{\sum_{i=1}^{6} p_i^2}\sqrt{\sum_{i=1}^{6} q_i^2}}$$

We first calculate the $L2$ norm of vector p, i.e. for word "animal":

$$\sqrt{\sum_{i=1}^{6} p_i^2} = \sqrt{2^2 + 3^2 + 0^2 + 3^2 + 0^2 + 3^2} = \sqrt{4+9+0+9+0+9} = \sqrt{31}$$

1. For $p = animal$ and $q = dog$:

$$sim_{cos}(p,q) = \frac{2*0 + 3*4 + 0*0 + 3*4 + 0*2 + 3*2}{\sqrt{31}\sqrt{0^2 + 4^2 + 0^2 + 4^2 + 2^2 + 2^2}}$$

$$sim_{cos}(p,q) = \frac{0+12+0+12+0+6}{\sqrt{31}\sqrt{0+16+0+16+4+4}}$$

$$sim_{cos}(p,q) = \frac{30}{\sqrt{31}\sqrt{40}} = 0.8519$$

2. For $p = animal$ and $q = cat$:

$$sim_{cos}(p,q) = \frac{2*4+3*0+0*0+3*3+0*3+3*10}{\sqrt{31}\sqrt{4^2+0^2+0^2+3^2+3^2+10^2}}$$

$$sim_{cos}(p,q) = \frac{8+0+0+9+0+30}{\sqrt{31}\sqrt{16+0+0+9+9+100}}$$

$$sim_{cos}(p,q) = \frac{47}{\sqrt{31}\sqrt{134}} = 0.7292$$

3. For $p = animal$ and $q = computer$:

$$sim_{cos}(p,q) = \frac{2*0+3*0+0*0+3*5+0*0+3*5}{\sqrt{31}\sqrt{0^2+0^2+0^2+5^2+0^2+5^2}}$$

$$sim_{cos}(p,q) = \frac{0+0+0+15+0+15}{\sqrt{31}\sqrt{0+0+0+25+0+25}}$$

$$sim_{cos}(p,q) = \frac{30}{\sqrt{31}\sqrt{50}} = 0.7620$$

4. For $p = animal$ and $q = run$:

$$sim_{cos}(p,q) = \frac{2*4+3*3+0*5+3*0+0*3+3*4}{\sqrt{31}\sqrt{4^2+3^2+5^2+0^2+3^2+4^2}}$$

$$sim_{cos}(p,q) = \frac{8+9+0+0+0+12}{\sqrt{31}\sqrt{16+9+25+0+9+16}}$$

$$sim_{cos}(p,q) = \frac{29}{\sqrt{31}\sqrt{75}} = 0.6014$$

5. For $p = animal$ and $q = mouse$:

$$sim_{cos}(p,q) = \frac{2*2+3*10+0*5+3*4+0*3+3*0}{\sqrt{31}\sqrt{2^2+10^2+5^2+4^2+3^2+0^2}}$$

$$sim_{cos}(p,q) = \frac{4+30+0+12+0+0}{\sqrt{31}\sqrt{4+100+25+16+9+0}}$$

$$sim_{cos}(p,q) = \frac{46}{\sqrt{31}\sqrt{154}} = 0.6658$$

Thus, the most similar word to "animal" based on cosine similarity is "***dog***" because $sim_{cos}(animal, dog)$ yields the maximum cosine similarity: 0.8519.

Problem 2

We can use the concept of hypernym for this problem. A hypernym of a sense is another sense that is in higher hierarchy in the WordNet hypernym tree. For example, the sense "animal" is a quite general sense that acts as the hypernym for more specific senses such as "mammal" and "insect".

The idea is divided into several steps:

1. First, locate a nearest common hypernym for our two input senses, i.e. the hypernym with the lowest distance to sense_1 and sense_2 (most specific hypernym).

2. Compute the distance from that nearest common hypernym to the most general sense in WordNet. For example, for noun senses, 'entity' is the most general sense.

3. Compute the similarity with below formula. Suppose depth(sense) denote the distance from the sense to the most general sense.

$$\text{similarity} = 2 * \frac{\text{depth(nearest common hypernym)}}{\text{depth(sense\_1)} + \text{depth(sense\_2)}}$$

   This is referred as the Wu Palmer similarity. This would result in a score ranging from 0 (completely unrelated) to 1 (completely related). Some threshold, for example 0.3, can be used to determine the relatedness. In other words, two senses are said to be related if the similarity score is above 0.3, else they are completely different.

   We can see that for noun senses, if the nearest common hypernym is close to "entity" we would likely to have lower similarity. Lower similarity can also be the result of the long distance between one or both senses and the nearest common hypernym.

   Reference: http://www.nltk.org/howto/wordnet.html (see wup_similarity function).

Another approach would be to use the Lesk algorithm and compute the overlap of words from both senses definition and examples. The more overlaps, the more related are the senses to each other. But, again, this would not yield a good result since often times there are no overlaps, even when we extend the overlap computation using both senses hypernyms. Thus, many related senses can be classified as unrelated in that case.