COMS W4705: Natural Language Processing
Written Homework 4

Yiran Wang (yw3201)

April 9, 2019

Problem 1

0.1 Euclidean Distance

$EuclideanDistance(X, Y) = ||X - Y||_2^2$

$$\begin{align} sim_{Euclidean}(animal, dog) &= ||[2,3,0,3,0,3] - [0,4,0,4,2,2]||_2^2 \simeq 3.3166 \tag{1}\\ sim_{Euclidean}(animal, cat) &= ||[2,3,0,3,0,3] - [4,0,0,3,3,10]||_2^2 \simeq 8.4261 \tag{2}\\ sim_{Euclidean}(animal, computer) &= ||[2,3,0,3,0,3] - [0,0,0,5,0,5]||_2^2 \simeq 4.5826 \tag{3}\\ sim_{Euclidean}(animal, run) &= ||[2,3,0,3,0,3] - [4,3,5,0,3,4]||_2^2 \simeq 6.9282 \tag{4}\\ sim_{Euclidean}(animal, mouse) &= ||[2,3,0,3,0,3] - [2,10,5,4,3,0]||_2^2 \simeq 9.6437 \tag{5} \end{align}$$

0.2 Cosine Distance

$CosineDistance = \frac{X \cdot Y}{|X|_2 \cdot |Y|_2}$

$$\begin{align} sim_{cos}(animal, dog) &= \frac{[2,3,0,3,0,3] \cdot [0,4,0,4,2,2]}{|[2,3,0,3,0,3]|_2 \cdot |[0,4,0,4,2,2]|_2} \simeq 0.8519 \tag{6}\\ sim_{cos}(animal, cat) &= \frac{[2,3,0,3,0,3] \cdot [4,0,0,3,3,10]}{|[2,3,0,3,0,3]|_2 \cdot |[4,0,0,3,3,10]|_2} \simeq 0.7292 \tag{7}\\ sim_{cos}(animal, computer) &= \frac{[2,3,0,3,0,3] \cdot [0,0,0,5,0,5]}{|[2,3,0,3,0,3]|_2 \cdot |[0,0,0,5,0,5]|_2} \simeq 0.7620 \tag{8}\\ sim_{cos}(animal, run) &= \frac{[2,3,0,3,0,3] \cdot [4,3,5,0,3,4]}{|[2,3,0,3,0,3]|_2 \cdot |[4,3,5,0,3,4]|_2} \simeq 0.6014 \tag{9}\\ sim_{cos}(animal, mouse) &= \frac{[2,3,0,3,0,3] \cdot [2,10,5,4,3,0]}{|[2,3,0,3,0,3]|_2 \cdot |[2,10,5,4,3,0]|_2} \simeq 0.6658 \tag{10}\\ & \tag{11} \end{align}$$

Problem 2

Homonymy defines multiple unrelated concepts correspond to the same word form, while Polysemy defines multiple semantically related concepts correspond to the same word form.
Given one word sense, we add all the synonyms, meronymy and holonymy (part-whole relation), hypernyms and hyponyms (IS-A relationship) of all synonyms in the same synset of the given sense

into a set. If the other sense doesn't appear in that set, they are completely different, otherwise, they are related.

Programming Component

Part 2 results

wn_frequency_predictor
Total = 298, attempted = 298
precision = 0.098, recall = 0.098
Total with mode 206 attempted 206
precision = 0.136, recall = 0.136

Part 3 results

wn_simple_lesk_predictor
Total = 298, attempted = 298
precision = 0.095, recall = 0.095
Total with mode 206 attempted 206
precision = 0.136, recall = 0.136

Part 4 results

predict_nearest
Total = 298, attempted = 298
precision = 0.115, recall = 0.115
Total with mode 206 attempted 206
precision = 0.170, recall = 0.170

Part 5 results

predict_nearest_with_context
Total = 298, attempted = 298
precision = 0.116, recall = 0.116
Total with mode 206 attempted 206
precision = 0.180, recall = 0.180

Part 6 results

I have used the linear combination of the cosine distance and the normalized frequency from part2. In addition, I have replaced the numeric context word with 'NUMBER' embedding.
predict_competition
Total = 298, attempted = 298
precision = 0.121, recall = 0.121
Total with mode 206 attempted 206

precision = 0.184, recall = 0.184