s1680791
s1781370

ANLP Assignment 3

Question for investigation: How is the outcome of different methods for measuring similarity affected by differences in word frequency?

I. Method comparison

We tried three different word similarity models: 1) cosine similarity of positive pointwise mutual information (PPMI) vectors (our baseline model), 2) cosine similarity of t-test vectors, and 3) Euclidean distance between PPMI vectors. Then we tested their outcome by looking at their respective Spearman correlations between word frequency and similarity. To construct the context vectors, we replaced PPMI with t-test values, which are a kind of significance-testing metric. We also tried cosine similarity and Euclidean distance as alternative methods for measuring similarity.

The t-test statistic computes the difference between observed and expected means, normalized by the variance, which can be used to measure how much more frequent the association is than chance. The higher the value of t, the greater the likelihood that we can reject the null hypothesis that the observed and expected means are the same. (Jurafsky and Martin 2017, pg.279).

The resulting t-test association measure is (Curran, 2003):

$$\text{t-test (a, b)} = \frac{P(a,\ b) - P(a)P(b)}{\sqrt{P(a)P(b)}}$$

Since $P(a,\ b) \geq P(a)P(b)$, the value of t-test is always positive.

Euclidean distance is an alternative similarity measure to cosine similarity which measures the absolute distance of each point in multidimensional space (word embeddings), which is directly related to the position coordinates of each point, while the cosine similarity measures the angle between vectors, which reflects the difference in direction rather than in position. The Euclidean distance can reflect the absolute difference of individual numerical features.

The calculation method of Euclidean distance is:

$$\text{Ed } (\mathbf{v}, \mathbf{w}) = \sqrt{\sum_i (v_i - w_i)^2}$$

II. Spearman correlation

Since a good similarity measure should measure only similarity, we usually assume that it should not be correlated with any form of word frequency. However, different similarity measures can be affected by their particular parameters. Hence, in order to see how the different methods are affected by word frequency, we explored the correlation between similarity and word frequency. The Spearman (rank) correlation (often called ρ) measures how close to a perfect monotonic relationship the inputs have. We expected a ρ value close to zero (indicating low correlation between similarity and word frequency) to reflect a good similarity measure method. We looked at the ρ of similarity and three word frequency measures: 1) the lower word frequency in a pair (minimum frequency), 2) the higher word frequency in a pair (maximum frequency), and 3) the difference in frequency in each pair.

III. Test word selection

We wanted a set of test words that would allow us to control for any extreme differences in context vector values or magnitudes. The main variable we wanted to investigate was frequency. For this reason, we decided to select words that fall neatly into one semantic category and which typically have the same syntactic role. We decided on adjectival colour words. Because colour adjectives are used in very similar contexts and usually the context is only distinguished by the object bearing a

certain colour quality (e.g. a green lime, a red brick, a blue sky, etc.), there was a high likelihood that a context vector based on co-occurrence with these object words would be a reasonable measure of relative meaning. Since we wanted to see the effects of word frequency on different models, we picked a mix of presumably high frequency words (red, yellow, blue, orange, green, purple, pink, black, white, gray, grey) and low frequency words (fuchsia, beige, amber, crimson, indigo, magenta).

IV. Results

A. PPMI and cosine similarity

Figure 2 illustrates that this model performed fairly well in that the correlation between similarity and max frequency, as well as the correlation between similarity and difference in frequency, were very close to zero. However, there was a fairly high correlation between the minimum word frequency and similarity. This was not surprising, since PPMI word similarity models have a bias toward assigning high similarities to infrequent words. (Jurafsky and Martin 2017, pg.276) So it makes sense that in word pairs where the less frequent word was particularly infrequent, there was a tendency to assign a higher PPMI to the word pair, thereby yielding a correlation value of 0.69 between similarity and minimum word frequency.

B. T-test and cosine similarity

As we can see in Figure 2, the Spearman correlation values from this method are relatively small which shows its similarity is less affected by word frequency. For both minimum and maximum frequency in each word pair, the correlation with similarity fell between the respective correlation values for the other two methods, and the similarity correlation with difference in frequency was the lowest of the three. Furthermore, compared to the baseline method, only the correlation between maximum frequency and similarity increased, though it was still fairly small, and both other correlations decreased, indicating an overall improvement from the baseline model. Hence, we can conclude that the t-test similarity measure is generally less affected by word frequency than PPMI, especially regarding difference in frequency.
However, while the results of this method are reasonable, t-test values should be more accurate with the population (N) is under 30. Here, N is the total number of documents in the dataset, which is a very large value.

C. PPMI and Euclidean distance

This model was clearly the most affected by word frequency, as expected (see section 1). A difference in magnitude of two vectors (in this case determined by word frequencies) would yield a greater Euclidean distance. This explains the very high correlation (0.68) between similarity and difference in word frequency. It is also not surprising that the minimum and, in particular, maximum word frequencies had very high correlations with the similarity measure, 0.38 and 0.77, respectively. Again, as expected, this model dealt especially poorly with high frequency words (see Figure 3) because high magnitude vectors are likely to be a high Euclidean distance away from each other and especially from low magnitude vectors. It seems the Euclidean distance is an appropriate similarity measure only if the vectors in question have fairly uniform and very low magnitudes, i.e. words with very similar and very low frequencies.

s1680791
s1781370

ANLP Assignment 3

1. Preliminary task results:

Cosine similarity	Word pair	Word1 frequency	Word2 frequency	Difference in frequency
0.36133	('cat', 'dog')	169733	287114	11528
0.16684	('comput', 'mous')	160828	22265	619
0.12069	('cat', 'mous')	169733	22265	975
0.09086	('mous', 'dog')	22265	287114	196
0.07188	('cat', 'comput')	169733	160828	215
0.06085	('comput', 'dog')	160828	287114	179
0.01551	('@justinbieber', 'dog')	703307	287114	300
0.01485	('cat', '@justinbieber')	169733	703307	240
0.01423	('@justinbieber', 'comput')	703307	160828	405
0.00873	('@justinbieber', 'mous')	703307	22265	66

Figure 1

2. Freq vs Similarity Spearman correlation

	Min frequency	Max frequency	Difference in frequency
PPMI + cos sim.	0.69	0.08	-0.18
T-test + cos sim.	0.46	0.24	0.05
PPMI + Euclidean dist.	0.38	0.77	0.68

Figure 2

3. The top and bottom five results of 3 methods

The top and bottom five results of the cosine similarity computations (using PPMI) are:

0.45635('black', 'white') 509536 383490 48705
0.45023('pink', 'blue') 98870 199619 3632
0.41944('blue', 'red') 199619 307667 18578
0.40528('purpl', 'pink') 54737 98870 3012
...
0.06703('blue', 'indigo') 199619 1984 111
0.06522('white', 'fuchsia') 383490 591 25
0.06448('yellow', 'indigo') 61507 1984 18
0.06403('green', 'crimson') 193219 2929 25
0.06115('green', 'amber') 193219 40048 103

SpearmanrResult(correlation=0.79732961360138555, pvalue=3.2404846351869032e-23)

The top and bottom five results of the cosine similarity computations (using t-test) are:

0.44439('blue', 'red') 199619 307667 18578
0.38443('white', 'yellow') 383490 61507 1825

0.33720('pink', 'blue')	98870	199619	3632
0.28721('black', 'white')	509536	383490	48705
0.27089('blue', 'yellow')	199619	61507	2432
...			
0.03780('black', 'crimson')	509536	2929	37
0.03654('black', 'amber')	509536	40048	388
0.03609('black', 'indigo')	509536	1984	36
0.03440('green', 'indigo')	193219	1984	30
0.03230('white', 'crimson')	383490	2929	171

SpearmanrResult(correlation=0.67081923185028158, pvalue=2.2453681293507913e-14)

The top and bottom five results of the Euclidean distance computations (using PPMI) are:

213.65763	('black', 'orang')	509536	71226	1615
202.71984	('black', 'green')	509536	193219	3642
197.64990	('orang', 'red')	71226	307667	1627
196.45780	('orang', 'white')	71226	383490	1231
195.49236	('black', 'amber')	509536	40048	388
...				
138.62604	('purpl', 'magenta')	54737	1851	18
137.96389	('yellow', 'magenta')	61507	1851	143
137.90376	('purpl', 'yellow')	54737	61507	1010
133.81594	('amber', 'gray')	40048	17520	69
133.79053	('yellow', 'beig')	61507	2615	17

SpearmanrResult(correlation=0.35023478439840194, pvalue=0.00035405665154563965)

4 . Example plots of word frequency vs computed similarity

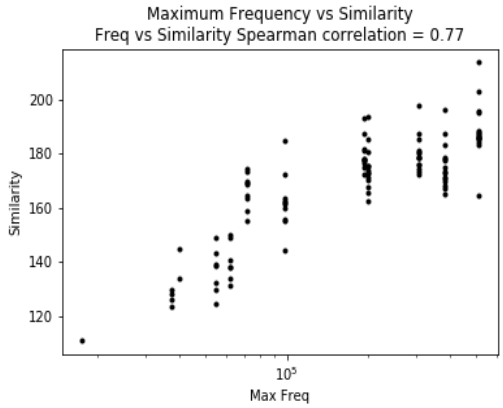

Figure 3: Euclidean distance using PPMI

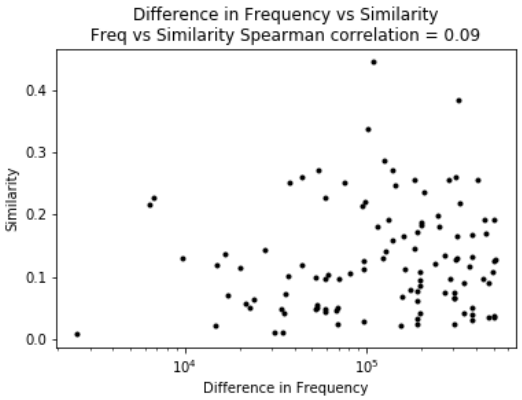

Figure 4: Cosine similarity using t-test

References:

Daniel Jurafsky and James H. Martin (2017). *Speech and Language Processing* (3rd Edition draft).

Curran, J. R. (2003). *From Distributional to Semantic Similarity*..Ph.D. *thesis*, University of Edinburgh.