s1424164 s1717759

ANLP Assignment 3: Exploring distributional similarity on Twitter Data

We set out to identify whether the scores emitted by the Jaccard, Jensen Shannon Divergence (JSD) and Cosine metrics, matched SIMLEX-999 for synonyms and matched our intuitions for unrelated words, with respect to the underlying context vector. In doing so, we attempt to explain some of the deeper reasons underlying the scores that are given. We conjecture that similarity ought to be close to one for synonyms, which do not significantly change the sentence meaning when substituted for one another. Unrelated words should have a similarities as close to zero as possible and antonyms should have a negative similarity, although in this report, antonyms are not studied. We then go on to change the context vector

Implementation details

Synonyms (Table 1) were sampled from the SIMLEX-999 (Levy et al. 2015) dataset, which had a similarity of 8.5 or over, out of 10. An identical number of unrelated words were chosen from a list, selected based on intuition. This was because SIMLEX-999 did not provide scores of unrelated word pairs, instead choosing to overload the meaning of zero scores as antonyms and scores mid-way between zero and ten as moderately similar. The context vector was constructed with PPMI. The similarities between these vectors were calculated with the Jaccard, Cosine and JSD metrics. The Jaccard metric (Formula 2) of two vectors calculates the sum of the element-wise intersection divided by the total union of both vectors. The Cosine metric of two vectors measures the cosine of the angle between two vectors. The JSD (Formula 1), a symmetrized bounded (between zero and one) version of Kullback-Leibler (KL) divergence, measures the expected log likelihood between the probability density function of each PPMI context vector and the probability density function of the mean of two PPMI context vectors, which is high when two context vectors have significantly different distributions. JSD measures dissimilarity, and has a value of zero when the vectors are identical. As such, it was adjusted into one minus the measure, to keep the semantics in line with the meaning of the Jaccard and Cosine measures, which have value one when identical. The JSD, from hereon refers to this adjusted measure.

Experiments

The scores for the Jaccard, cosine and JSD are shown in Fig. 1. Ideally, similarity measures should show high scores (close to 1) for all synonyms we chosen. However, Fig. 1, shows that scores for all three similarity metrics, are very low, i.e. upper-bounded by 0.5 for synonyms. However, the unrelated words are quite close to zero, imbuing the metrics with a 'one-sided' accuracy. However, the ranking performance of each scoring metric is good. In this assignment, we concentrate on dissecting why scores for synonyms are upper-bounded at only 0.5. With this weakness in mind, we attempt to improve the performances of all three similarity measures, by attempting to yield synonym scores close to one and scores from unrelated words, close to zero, such that the absolute differences are minimised. From hereon, the spread denotes the number of greater than zero elements in each PPMI vector and the spread ratio refers to the spread of one vector divided by the other. We noticed a trend that all similarity metrics exhibited, which was, as the frequency of the target word relative to the context word increased, the scores decreased. We hypothesised that synonyms with similar spread, will have higher scores than those with a dissimilar spread, due to the relationship between spread and intersection and "non-overlapping" elements of the vectors. A large spread ratio, could therefore be a factor contributing to a low score for synonyms.

We plotted spread ratio against similarity for all three measures, revealing an interesting relationship, as shown in Fig. 2. Fig. 2 shows, that as the spread ratio deviates from one, the similarities decrease for all measures, in an approximately log relationship (R^2 = 0.7, 0.3, 0.7). In order to understand why the scores for unrelated words were higher than zero for some words, we wanted to understand whether the measures were simply using the spread ratio as a *proxy* for similarity. We tested the same relationship on unrelated words, finding the same negative correlation as shown in Figure 3. The similarity values of unrelated words tended to be about a third than those of synonyms. However, as different words were tested, we can only infer that spread ratio likely plays at least part of the role in similarity. This is exemplified by two pairs of unrelated words, the stemmed words "committe" and "cent", achieving a moderately high JSD of 0.21, and a spread ratio of about 1.3, versus "broccoli" and "umbrella", which have a JSD of 0 and a spread ratio of 55.5. We then plotted the spread ratio versus word frequency for both synonyms and unrelated words, as shown in Figure 12. It can be noticed that regardless of whether the words were synonyms or not, that there is a strong positive correlation (R = 0.87) between the two. An order two polynomial, was generated as a line of best fit, with an R^2 values of 0.86 and 0.94, for synonyms and unrelated words. This suggests, differences in word frequencies are associated with changes in spread ratio which influence scores from the three similarity metrics for both synonyms and unrelated pairs and as such demonstrates some of the underlying reasons, for the low scores, given to some synonym word pairs.

However, even if the relative word frequencies are similar, the scores are still much lower than one for synonyms. To control for the effects of relative frequency ratio and spread ratio, we chose the synonym pair 'jet' and 'plane' which has a word frequency ratio of 1.4 and spread ratio at 1.06, yet only scores 0.17, 0.27 and 0.35 for Jaccard, cosine and JSD, respectively. The similarity metrics depend on the elements of the PPMI vector that intersect (intersection), i.e. those that are non-zero in both vectors, as well as the "non-overlapping" vector elements, i.e. those that are zero in one vector and non-zero in the other. The numerator of the Jaccard measure consists solely of the intersection, whereas the denominator depends on the intersection plus the "non-overlapping" elements. The cosine measure relies on the dot product which will be zero if they are "non-overlapping" and non-zero otherwise and finally the JSD texit{similarity} depends on the expected log likelihood which decreases if one vector is non-zero and the other is zero. As the number of "non-overlapping" regions increase, the scores will decrease and vice versa. No formal proof of this claim is given. Based on this intuition, and to give a deeper understanding of the low scores for this particular word pair, we set out to understand the demographic of the PPMI-based vectors, by plotting the "non-overlapping" PPMI values.

Fig 4 and 5, show PPMI values versus occurrence counts for each non-overlapping, non-zero, context word of the "jet" and "plane" context vectors, respectively. These demonstrate the existence of numerous low frequency words which also tend to have the highest PPMI values and that high frequency words tend to have lower PPMI values. To identify the importance of the frequency of the words, the proportion of the L1 vector norm of the non-overlapping regions, for both words, against the total L1 norms of both vectors added together, was plotted against the occurrence count, for words with occurrence count less than 100,000, as shown in Fig. 6. From hereon, we denote the quantity on the y-axis of Fig. 6 as the L1 proportion. Fig. 6 conveys the extent to which each occurrence band (per 2000) contributes to the L1 proportion, and hence, which occurrence values are more likely to be responsible for undermining the similarity measures tested. The plot shows that words that occur in total 6,000 times or less (864 points) contribute massively (38%) to the L1 proportion. In total, the non-overlapping regions (2935 points) contribute to 61% of the L1 proportion, which, following on from previous analysis on non-overlapping regions, explains the low scores for all metrics for synonym pair "jet" and "plane". We wanted to understand whether there were many words that were responsible for this difference or whether it was just a few. As such, we plotted two graphs Fig. 7, Fig. 8 which show the frequency of words of various occurrence count bands of size 1000, beneath 30,000 occurrence counts, for both the overlapping and non-overlapping regions. One can now draw the conclusion that the majority of words in the non-overlapping regions are low frequency (below 3000 occurrence count) and these words are responsible for almost 50% of the L1 proportion in the non-overlapping regions, suggesting an improvement may be to reduce the impact of these words. The overlapping regions have far fewer words relative to the total number of words in the 0-3000 occurrence count range than the non-overlapping regions. These low frequency words generally had the majority of their co-occurrences with the target word. An example of this is the word "@nhljets" which had 107 out of its 335 total occurrences, co-occurring with "jet", with a PPMI of 9.90. 37% of the words beneath 1,000 occurrence counts contained "#" or "@" symbols, reducing to 22% below 6,000 occurrences. This could be due to the data consisting of many re-tweets with spelling errors or other low usage words that only have one of the two words in them. Many words below 3000 occurrence counts are spelling error words, such as "greatbacaus" which had 112 out of its 118 total occurrences, co-occurring with "jet", with a PPMI of 11.51. These show that the similarity measures struggle with numerous low frequency words, which represent a large fraction of the total PPMI.

Another point that we now establish is that even if the vectors intersect, there may be variation within the vector entries, which encourage smaller similarity scores to be allocated. In this vein of thinking, we also plot a scatter plot of the words in the intersection. Fig. 9 shows details of the intersection, i.e. vector entries from both vectors, and plots the occurrence count versus PPMI value for synonym pair "jet" and "plane". The scatter distribution of "plane" is shifted relative to the words from "jet" in the bottom-left corner, which is because the word frequencies between "plane" and "jet" are not exactly the same. (The word frequency of "plane" is 1.3 times larger than that of "jet"). The minimum PPMI value that can be assumed of 'plane' is lower than that of 'jet', when the context words are the same. This phenomena, can also be seen in Fig. 10 for words 'odd' and 'weird', which have a relative word frequency ratio of 4. This suggests comparing words of differing frequency is not only difficult because of differing spread ratios, but also because the PPMI value itself has different lower bounds, for low frequency words, depending on the count of the target word. The L1 proportion is plotted for the intersection, as shown in Fig 11. The intersection takes up 39% of the L1 proportion. As such, the y-axis is upper-bounded by 4% rather than 25%. From Fig. 11, it can seen that the L1 proportion for the intersection is flatter than non-overlapping in Figure 7, which indicates that unlike non-overlapping, low occurrence words (less than 2000) do not significantly dominate the intersection L1 proportion.

Taking into account all of the available evidence, an idea to improve the performance of three similarity methods is deleting the words with occurrence counts less than 2000. For "jet" and "plane", we deleted the words with 2000 occurrences, yielding an average increase in score of 0.1. We also had to consider the potential increase of unrelated word scores. The plot, as shown in Figrue 10, show the results of pre-processing. From Figure 10, it can be seen that pre-processing operator exactly improves the absolute difference of the scores for synonyms for all three similarity measures and it keeps the absolute difference of the scores for unrelated words approximately constant. Hence, our pre-processing method can be used to improve the performance of Jaccard, Cosine and JSD similarity measures. Figure 13 shows that all three similarity metrics with pre-processing by deletion of low frequency words, for synonyms became closer to the scores given by SIMLEX-999. Figure 14 shows that the absolute difference in scores for un-related word pairs, after pre-processing, is smaller than that of the synonyms. Here, we re-scaled the SIMLEX-999 scores, by dividing them by 10, to keep the semantics in line with the meaning of Jaccard, Cosine and JSD. Hence, our pre-processing method can be used to improve the performance of Jaccard, Cosine and JSD similarity measures.

Conclusion

Differences in word frequencies are associated with changes in spread ratio, henceforth influencing the scores for Jaccard, Cosine and JSD measures for both synonyms and unrelated pairs. Low frequency words, often accompanied with the symbols “#” and "@" as well as words with spelling errors, influence the scores for all three similarity measures using PPMI context vectors. However, pre-processing, deleting words with occurrence less than 2000, enabled an improvement in the scores for synonyms and keeps scores for unrelated pairs approximately constant, thus improving the overall performance of all three similarity measures. It is unclear

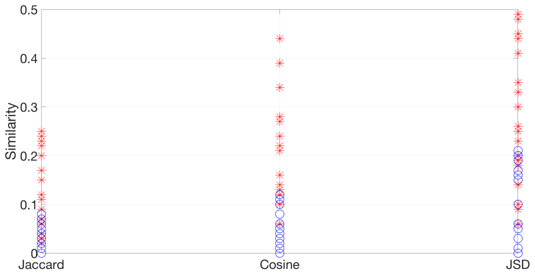

Fig 2. Jaccard, cosine and JSD scores for 20 synonym word pairs and 20 unrelated word pairs. Red points represent synonym word pairs, while blue points represent unrelated word pairs. Some points coincide due to rounding to two decimal places.

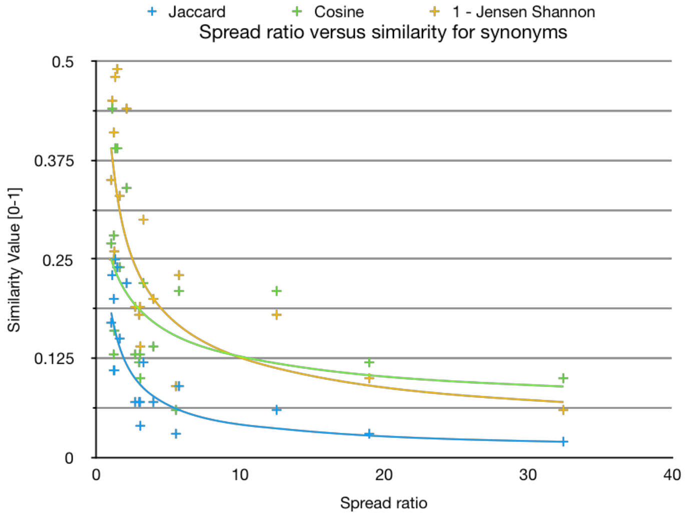

Fig 2. Spread ratio versus word frequency for synonyms.

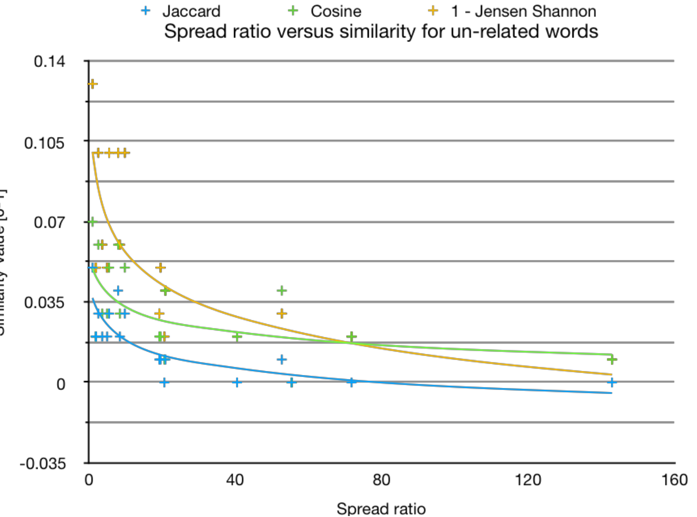

Fig 3. Spread ratio versus word frequency for unrelated words

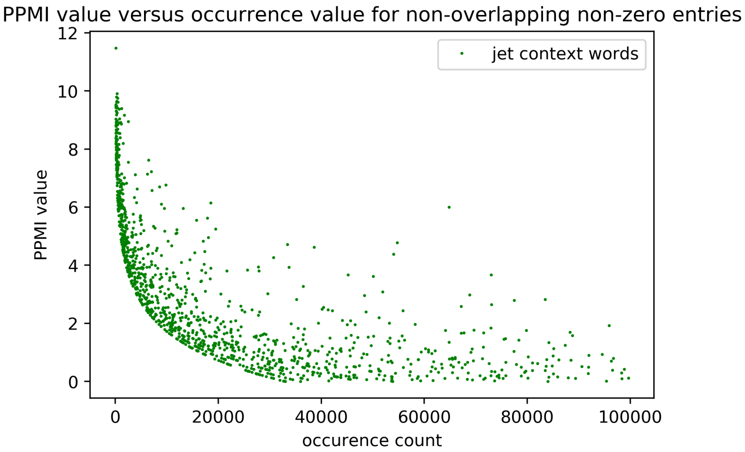

Fig 4. PPMI value versus occurrence count value for non-overlapping non-zero entries for word "jet"

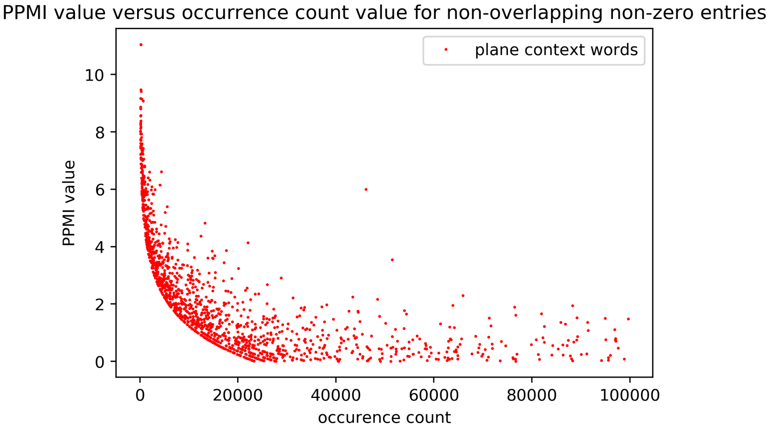

Fig 5.PPMI value versus occurrence count value for non-overlapping non-zero entries for word "plane"

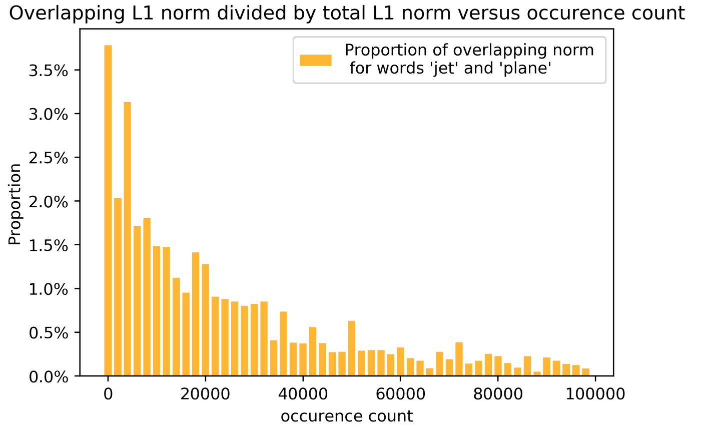

Fig 6. This shows the L1 proportion for words in the non-overlapping vector.

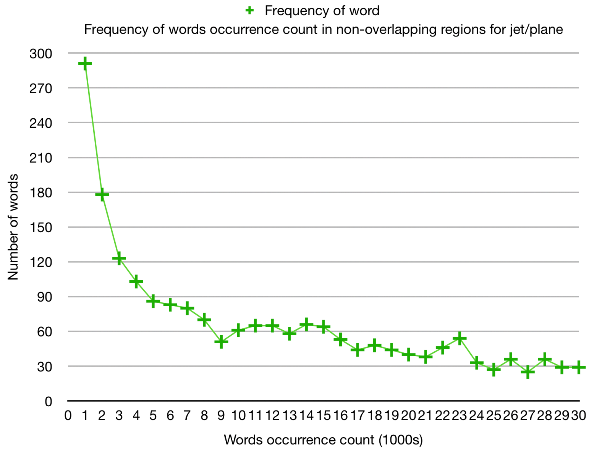

Fig 7. This depicts the occurrence counts of the words in the non-overlapping regions versus the number of them for 'jet' and 'plane'.

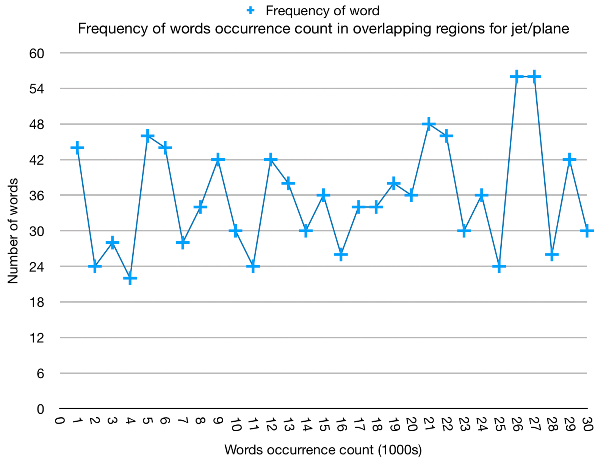

Fig 8. This depicts the occurrence counts of the words in the non-overlapping regions versus the number of them for 'jet' and 'plane'.

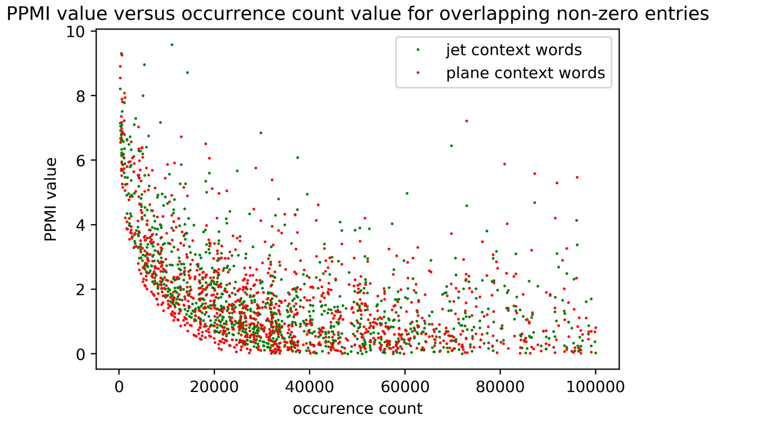

Fig 9. This shows the vector entries in jet and plane that are non-zero in both. For each entry, the occurrence count and PPMI value is plotted.

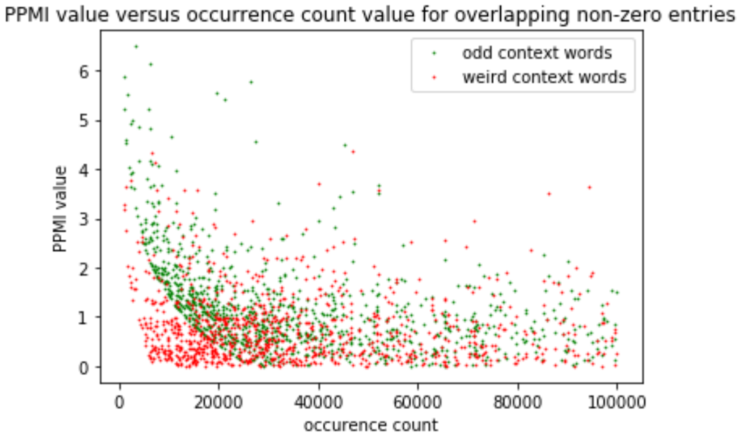

Fig 10. This shows the vector entries in odd and weird that are non-zero in both. For each entry, the occurrence count and PPMI value is plotted.

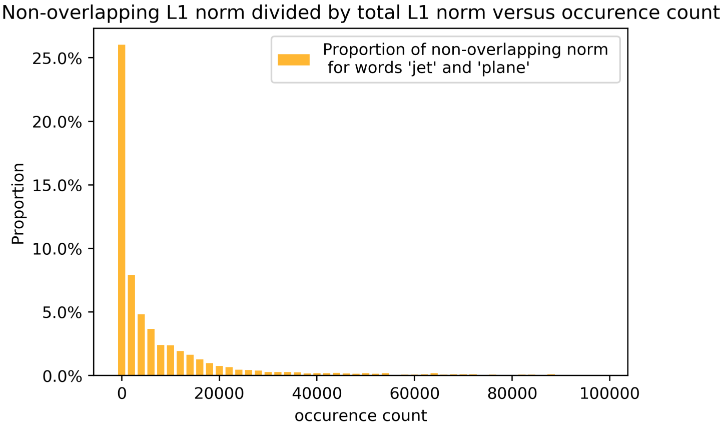

Fig 11.This shows the L1 proportion for words in the overlapping vector.

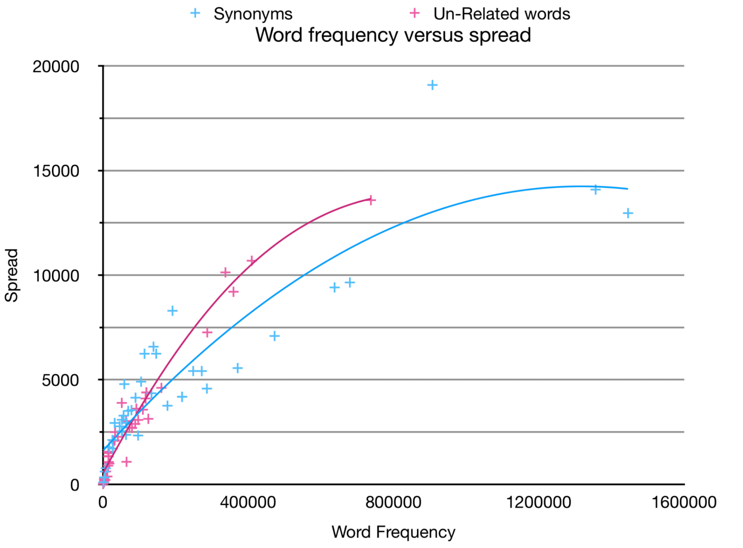

Fig 12. Spread ratio versus word frequency for both synonyms and unrelated words

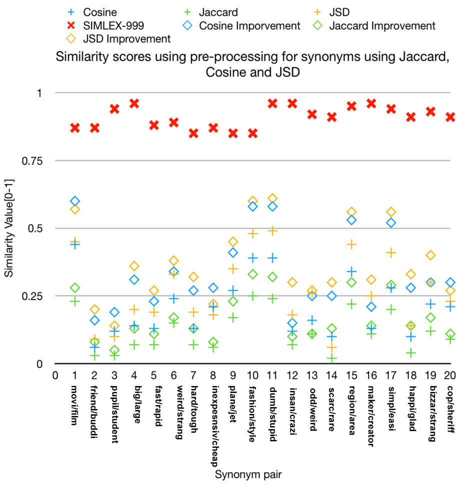

Fig 13. This depicts the score improvement for each of the synonyms when the preprocessing step is applied. The SIMLEX-999 Score has been scaled down to the range 0-1.

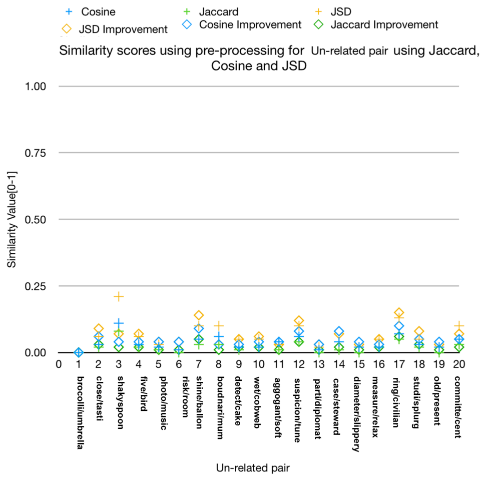

Fig 14. This depicts the score improvement for each of the unrelated words when the preprocessing step is applied. The SIMLEX-999 Score has been scaled down to the range 0-1.

Synonyms		Unrelated	
Movi/Film	dumb/stupid	brocolli/umbrella	wet/cobweb
Friend/Buddi	insan/crazi	close/tasti	arrogant/soft
Pupil/Student	odd/weird	committee/cent/no	suspicion/tune
Big/Larg	scarc/rare	shaky/spoon	parti/diplomat
fast/rapid	region/area	five/bird	case/steward
weird/strang	maker/creator	plato/music	diameter/slippery
hard/tough	simpl/easi	risk/room	measure/relax
inexpensiv/cheap	happi/glad	shine/balloon	ring/civilian
plane/jet	bizzar/strang	boundari/mum	studi/splurg
fashion/style	cop/sheriff	detect/cake	old/present

Table 1. The words used in this report.

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$$

Formula 1. Wikipedia: JSD

$$\frac{|A \bigcap B|}{|A \bigcup B|}$$

Formula 2. Wikipedia: Jaccard measure

Levy, Omer, Yoav Goldberg, and Ido Dagan. "Improving distributional similarity with lessons learned from word embeddings." *Transactions of the Association for Computational Linguistics* 3 (2015): 211-225.